What Does a Good Prediction Look Like?

Florian Wirth[1], Stephan Krane, Melanie Loos, Eike Rehder[2] and Carlos Fernandez[1]

***Abstract*— In the developement of safe intelligent vehicles, intentions of other road users need to be estimated in order to plan a safe and convenient trajectory. The present work tries to approach the answer to the question of how a prediction could be assessed holistically using the example of pedestrian prediction. Recently used metrics are compared on a state of the art (SOTA) prediction method. A good prediction does not only have to verify itself on a limited data set, so machine versus human pedestrian prediction are compared. A study with test persons points out the similarity of human predictions compared to SOTA prediction. However, further research is required since we cannot propose a holistic metric that answers the question in the title, yet.**

***Index Terms* — pedestrian prediction, human prediction, prediction metrics, human machine interaction.**

I. Introduction

Superordinate to saving energy, accelerating traffic, improving comfort and enabling efficient usage of travel time, increasing traffic safety is the primary demand on automated vehicles. If traffic safety decreases due to automated vehicles, it is likely that automated driving will not prevail compared to manual driving. Minding the safety of third party traffic participants is inseparably included in the demand of driving safely. In the future, intelligent vehicles will be connected and share information about desired future trajectories. So-called vulnerable road users (VRU) that participate in traffic will not be able to share detailed information of future trajectories. A significant proportion (46% [1]) of road deaths represent VRUs which mainly include pedestrians, cyclists and motorcyclists. The focus of attention currently is on pedestrians because they represent the weakest type of VRU.

If the motion of traffic participants can be predicted successfully, the behaviour of the ego vehicle can adopt to the changing traffic situation in advance. This will affect all demands on automated driving metioned above. Like all measurement tasks, prediction is only useful if assumptions about its uncertainty are provided.

We would like to tackle two questions with the present work:

1) *How may prediction quality be quantified?*
   In contrast to other tasks in perception, prediction measures a quantity which is not defined at the moment of measuring. Even if all existing knowledge was available, events in the future could not be predictied with absolute certainty.
   The knowledge necessary to estimate the scale of an object - e.g. a vehicle's dimensions - exist as a physical quantity, they are written in the car's manual and some people might know those value by heart [1]. From this, we derive the necessity to evaluate prediction differently from classical estimation tasks.
2) *How accurate does a prediction need to be for automated driving?*
   We assume that the way humans predict is good enough for driving. This is a rather strong assumption since there obviously are traffic fatalities due to bad intention interpretation. Still, in the overwhelming majority of road conflicts human intention interpretation seems to be sufficient. The remaining question is:
   How accurate is human prediction compared to SOTA prediction and recorded Ground Truth (GT)?

II. Related Work

A summary of current pedestrian detection, prediction and interpretation SOTA can be found in recent surveys. Herein, Riedel *et al.*[2] published a literature survey of progress in pedestrian intention estimation. Earlier, Brouwer *et al.* [3] and Schneider *et al.* [4] quantitatively compared state of the art motion prediction models.

In general, research work in pedestrian behaviour prediction can be divided in two kinds: Prediction as binary intention interpretation and forecasting of destinations.

Contributions of the first kind are means to an end, they constitute a service providing module for automated driving [5]–[11]. In these works, pedestrians' intentions are derived corresponding to potential conflicts with the ego vehicle in the near future. Here, predictions provides scalar states $0 \leq s \leq 1,\ s \in \mathbb{R}$ with i.e. 0 being a risk free state and 1 being an assured conflict.

Contributions of the second kind work on estimating pedestrians' future positions as a stand-alone field of research. Future positions can be predicted independently of interaction with road traffic [12]–[15], by taking road topology into account [16], [17] and by modeling the interaction between road traffic and pedestrians [18].

Schmidt *et al.* [19] conducted similar experiments with test persons in order to figure out which visual features are used by observers to predict whether a pedestrian crosses a road. They masked constituent visual features and analyzed how the prediction quality changes if the masked feature is not observable for the test person.

*This work was supported by Intel Corporation.

[1]are with Intitute of Measurement and Control Systems, Karslruhe Institute of Technology, Karlsruhe, Germany {firstname.surname}@kit.edu

[2] Eike Rehder is with Daimler R&D AG, Sindelfingen, Germany eike.rehder@daimler.com

[1]This deliberation is independent of the observability of the object's scale.

The present work does not focus on the importance of visual features and how they affect prediction quality. Instead, it focuses on prediction quality obtained by test persons and compares their prediction results with prediction generated by a SOTA approach [17]. The overall goal of our ongoing work is to find a holistic prediction metric. Hub *et al.* [15] also identified the problem of not having good prediction metrics.

III. Towards New Prediction Metrics

A. Motivation

This work was inspired by a specific finding mentioned in Rehder *et al.* [17]: we found the superiority of learning from the maximum error of a batch

$$l_{max} = -\log(\min(p_i)) \tag{1}$$

when a batch is the path of a tracked pedestrian for which corresponding GT is known. p_i is the reward corresponding to timestep i. The default option in machine learning is to train with the mean error

$$l_{avrg} = -\log(\text{avrg}(p_i)). \tag{2}$$

This superiority of learning with l_{max}, however, was only observed when looking at the output PDF manually. The output PDFs look smother and more intuitive. The mean loss when evaluating was roughly in the same scale for both loss functions, so numerically, one option is not preferable over the other.

For training, two slightly different cost functions to calculate the rewards p_i are used. The first reward function is given by the Gaussian term of eq. 3 in Rehder *et al.* [17] and is the 2D extension of Bishop's Mixture Density Network (MDN) [20]. Here, GT is modeled as a point $\vec{x}_{GT}$:

$$p_{Dot} = \sum_{j=1}^{\text{Mix}} \Pi_j \mathcal{N}(\vec{x}_{GT}; \vec{\mu_j}, \Sigma_j). \tag{3}$$

The second cost function led to visually smoother and more intuitive results. It models a pedestrian as a Gaussian with fixed variance $\sigma_{Ped}^2 = 0.25^2$. This complies roughly with the physical expansion of a person and was emperically verified. The predicted PDF is multiplied with the pedestrian's Gaussian, which results in

$$\begin{aligned} & p_{Vol} \\ &= \sum_{j=1}^{\text{Mix}} \Pi_j \iint_{-\infty}^{+\infty} \mathcal{N}_{Pred}(\vec{x}|\vec{\mu_j}, \Sigma_j)\mathcal{N}_{Ped}(\vec{x}|\vec{x}_{GT}, \Sigma_{Ped})d\vec{x} \\ &= \sum_{j=1}^{\text{Mix}} \Pi_j \mathcal{N}_{Pred}(\vec{\mu_j}|\vec{x}_{GT}, \Sigma_j^c) \end{aligned} \tag{4}$$

with

$$\Sigma_j^{\mathbf{c}} = \Sigma_j + \Sigma_{Ped}, \Sigma_{Ped} = \begin{pmatrix} \sigma_{Ped}^2 & 0 \\ 0 & \sigma_{Ped}^2 \end{pmatrix} \tag{5}$$

being the combined covariance matrix [21]. The cost is calculated as in Bishop [20] by using the negative logarithm of the reward p

$$l = -\log(p), \tag{6}$$

with either $p = \max(p_i)$ or $p = \text{avrg}(p_i)$. To summarize, we do now have a set of 2×2 cost functions[2]. Each of them can be used for training while the others are used for evaluation purpose only.

Visually, PDFs generated by a network trained with the combination $l_{max} \leftrightarrow p_{Vol}$ look best but the superiority of this combination could not be proven numerically. In conclusion, we are looking for a metric or a set of metrics which can assess the holistic quality of a prediction.

B. Recent Prediction Metrics

Metrics for prediction are used in all kind of data driven research fields: In economics stock prices get predicted, in meteorology weather and climate forecasts are a major field of interest, and critical machine states shall be predicted in engineering.

In order to assess the quality of a prediction, scoring rules can be used. They provide a value that is interpreted as a measure for sharpness and concordance between predicted and „real“ distribution [22]. By doing so, the goodness of a prediction model can be assessed which allows a comparison between different models [23].

There are numerous types of scoring rules. Well known examples are the quadratic scoring rule and the Kullback-Leibler divergence.

The quadratic scoring rule is given by

$$S_{\text{disc}}(\vec{x}_{GT}, p_{Pred}) = 2 \cdot p_{Pred}(\vec{x}_{GT}) - \sum_{k}^{\text{Bins}} p_{Pred}(\vec{x}_k)^2 \tag{7}$$

in discrete formulation with $S_{disc} \in [-1, +1] \cap \mathbb{R}$ and

$$S_{\text{conti}}(\vec{x}_{GT}, p_{Pred}) = 2 \cdot p_{Pred}(\vec{x}_{GT}) - \iint_{-\infty}^{+\infty} p_{Pred}(x, y)^2 dxdy \tag{8}$$

in continuous form. Higher scores correspond to higher prediction quality. For the discrete scoring rule, the Gaussian mixture was discretized in equidistant bins and at least $99,9\%$ of probability mass was used. $p_{Pred}(\vec{x}_{GT})$ is the probability of the bin containing the GT position. A baseline was defined by deploying the „pedestrian's distribution“ with covariance Σ_{Ped} defined in eq. 5. The Kullback-Leibler scoring rule is applicable by using the predicted and the GT PDF. If $\frac{1}{2\pi\sqrt{\det \Sigma_{Ped}}} = p_{Pred}(\vec{x}_{GT})$, the score is 0.

The Kullback-Leibler divergence given by

$$KL(p_{Ped}||p_{Pred}) = \iint p_{Ped} \cdot \log\left(\frac{p_{Ped}}{p_{Pred}}\right) d\vec{x} \tag{9}$$

can be used to measure the loss of information between a given distribution and its approximation [24], [25]. To

[2]The notation $l_{\text{reduction method}} \leftrightarrow p_{\text{prob. model}}$ states which cost funtion was used.

assess a prediction, the information loss can be regarded as a measure of unsuitability to approximate the original distribution.

Furthermore, the well-known χ^2-test which tests a statistical model for its validity on a given set of data shall be applied [26]. It is not commonly used to test prediction methods because approaches based on classical statistical methods like (Extended or Unscented) Kalman filters are optimal estimators and fulfill a χ^2-test adequately.

In order to perform Pearson's χ^2-test with respect to the goodness of fit on a set of samples, the expected distribution needs to be discretized in n bins for which the expected probability mass p_n is known. The numbers of samples in those bins are then compared to a model which is supposed to represent this set of samples.

For the present use case, each input of the prediction network produces a different prediction. Each prediction consists of a Gaussian mixture. In order to test the goodness of fit, the PDF needs to be discretized in a constant number of bins. These bins are supposed to contain the same probability mass for each prediction. For Gaussian mixtures, this cannot be achieved in a useful way[3]. However, it can be achieved for a single 2D-Gaussian as described in Moore *et al.* [27]. A 2D-Gaussian can be analytically separated into a number of elliptical probability rings with a desired probability mass.

Assuming a Gaussian

$$f(x) = \exp\left(-\frac{1}{2}(\mu - x)^T \Sigma^{-1} (\mu - x)\right) \tag{10}$$

that only differs from a valid PDF $g(x)$ in the scaling term $\frac{1}{\sqrt{2\pi\sigma}}$. The maximum of $f(x)$ is located at $f(\mu) = 1$. By assigning a desired value (e.g. 10% for the first bin) to the integral from $\mu - x_{\text{bin}}$ to $\mu + x_{\text{bin}}$ of $g(x)$, x_{bin} can be determined and a constant threshold $c_{\text{bin}} = f(\mu + x_{\text{bin}})$ is calculated. The region $\{x : f(x) > c_{\text{bin}}\}$ now contains the same probability mass for every PDF $g(x)$ of same dimension. This procedure can be conducted until all of the Gaussian's probability mass is split to one of n (e.g. 10) rings that now all contain 10% of probability mass. The split to 10 elliptical bins is shown in Fig. 1. With knowing these thresholds, the χ^2-test can be applied.

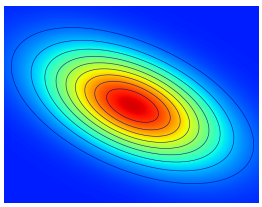


Fig. 1: Shares of 2D Gaussian with equal amount of probability mass (10%) in each ellipse ring

[3]Discretize with a equidistant grid in a limited region does not work, because the cells would always contain different shares of probability mass. Integrating each grid cell so that it contains a fixed share of mass is possible but the shape of cells might not be steady and therefore, cells of two PDF cannot be associated. Integrating independently in each dimension is an option, but we found a different and more useful method.

Still, it is necessary to form a single Gaussian from the present mixture. When observing the scales of mixture coefficients Π_i, it is noticeable that a significant share of this mass (roughly around 80%) is often united to only a few of the mixture components. Furthermore, all mixture components with significant share of probability mass are often close to each other compared to the corresponding standard deviations. Combining Gaussian mixture components is a separate field of research (e.g. Henning [28]) we do not want to enter in-depth. A simple estimate based on a weighted sum was used according to the following Algorithm 1.

Data: $var_i : \mu_{x,i}, \mu_{y,i}, \rho_i, \sigma_{x,i}, \sigma_{y,i}, \Pi_i, i \in [1, \text{Mix}] \cap \mathbb{N}$
Result: Approximate Gaussian mixture components as a single Gaussian that incorporates at least 80% of prob. mass
initialize $var^c : \mu_x^c, \nu_y^c, \rho^c, \sigma_x^c, \sigma_y, \Pi^c = 0$
while $\Pi^c < 80\%$ **do**
  $m = argmax(\Pi_i)$
  $var^c = var_m$
  $\sigma_x^c = \frac{\sigma_{x,m} \cdot \Pi_m + \sigma_x^c \cdot \Pi^c}{\Pi^c + \Pi_m} + \frac{\mu_{x,m} \cdot \Pi_m + \mu_x^c \cdot \Pi^c}{\pi_m + \pi^c}$
  $\sigma_y^c = \frac{\sigma_{y,m} \cdot \Pi_m + \sigma_y^c \cdot \Pi^c}{\Pi^c + \Pi_m} + \frac{\mu_{y,m} \cdot \Pi_m + \mu_y^c \cdot \Pi^c}{\Pi_m + \Pi^c}$
  $\rho^c = \frac{\rho^c \cdot \Pi^c + \rho_m \cdot \Pi_m}{\Pi_m + \Pi^c}$
  $\mu_x^c = \frac{\mu_{x,m} \cdot \Pi_m + \mu_x^c \cdot \Pi^c}{\Pi_m + \Pi^c}$
  $\mu_y^c = \frac{\mu_{y,m} \cdot \Pi_m + \mu_y^c \cdot \Pi^c}{\Pi_m + \Pi^c}$
  $\Pi^c = \Pi^c + \Pi_m$
  $\Pi_m = 0$
return $\mu_x^c, \nu_y^c, \rho^c, \sigma_x^c, \sigma_y, \pi^c$

Algorithm 1: Procedure to simplify Gaussian mixture based on weighted sum.

With combining Gaussians mixtures to a single Gaussian and dividing it in 10 bins it is now possible to apply χ^2-test in order to verify the model.

C. Human Prediction

Fig. 2: GUI used for experiments. The pedestrian marked by the red dot shall be predicted. The uncertainty σ_{hu} defines the radius of the covariance ellipsis.

In order to obtain a better prediction metric, experiments ware conducted. The test persons watched 22 videos of traffic scenes recorded by an experimental vehicle. Each video was about 10 sec. One pedestrian per video was marked. At the end of each video, test persons were asked to predict this pedestrian by annotating the presumed destination in the

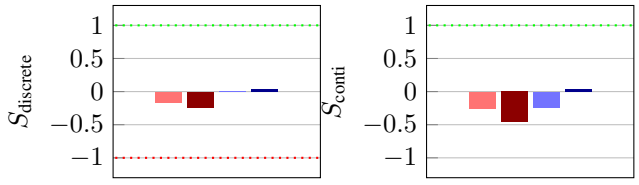
image. Three types of prediction were tested: predicting a destination with uncertainty, predicting only the direction of a destination with uncertainty, and predicting a path. All methods were evaluated with a prediction horizon of $2.5sec$. To make the concept of uncertainty more intuitive, test persons had to choose a **h**uman **u**ncertainty $\sigma_{hu} \in [0, 100] \cap \mathcal{N}$ (Fig. 2). This value is then mapped to a Gaussian or a von-Mises uncertainty (σ^2 or $\frac{1}{\kappa}$, respectively) by a nonlinear function whose parameters were determined empirically. They are categorized subjectively as „easy“, „moderate“ and „hard“.

The goal is to discover how good a prediction needs to be in order to drive with it and to derive rules for the prediction quality required for automated driving. For the experiments only people who participate in traffic on a regular basis were chosen. In total, the set of **t**est **p**ersons is $\Omega_{tp} = 32$.

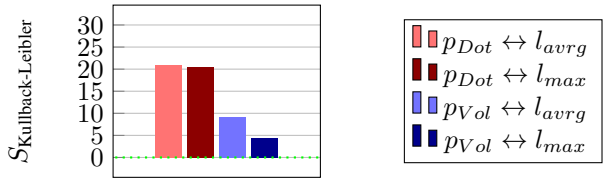
IV. Results

A. Prediction Metrics Applied to Single Pedestrian Network

According to Fig. 3 all prediction metrics clearly rate $l_{max} \leftrightarrow p_{Vol}$ higher than the competing approaches. However, according to these metrics, learning from the largest error seems to have a smaller impact on the score than modeling a pedestrian as a PDF (eq. 4) instead of a point (eq. 3). When looking at a set of PDFs manually, the difference of $l_{max} \bowtie l_{mean}$ is more remarkable than $p_{Dot} \bowtie p_{Vol}$, so none of those metrics used in other research fields fully confirm our observations.



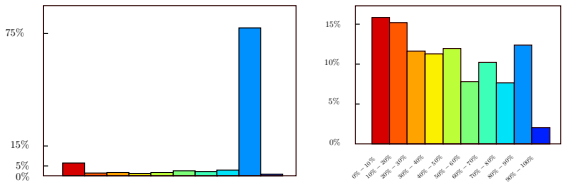
(a) Discrete quadratic scores (b) Continuous quadratic scores



(c) Kullback-Leibler scores

Fig. 3: Results of 2×2 different loss functions, evaluated with three different scoring rules. Green dashed line: Best score, red dashed line: Worst score. p_{Vol} outperforms p_{Dot} but none of the scores reflect the observed superiority of l_{max} over l_{avrg}.

Last, results of the χ^2-test are shown in Fig. 4. For all tests a statistical significance of 5% is used. According to Pearson's goodness of fit test, $p_{Vol} \leftrightarrow l_{avrg}$ does not represent the dataset ($300 < \chi^2 < 600$), since its histogram is far from the uniform distibution expected. Compared to l_{avrg}, a model trained with l_{max} yields to a valid distribution on the given test dataset ($5 < \chi^2 < 15$). The χ^2-test is the only metric that is able to substantiate our observations.



(a) $p_{Vol} \leftrightarrow l_{avrg}$ (b) $p_{Vol} \leftrightarrow l_{max}$

Fig. 4: Histogramm of GT destinations in 10 ellipse bins according to color scheme of Fig. 1.

B. Prediction Results of Humans

The images shown in Fig. 6 and 7 represent the overall results that are statistically shown in Fig. 5.

In Fig. 6, point-shaped predictions are shown. Predictions of test persons (blue points with uncertainties as circles) and the automatically generated predictions (green ellipses) have a large uncertainty. In Fig. 7 directions of the future destination are shown. With both prediction methods direction in which a future position is located is met quite precisely. However, both the human and the machine prediction often misestimate the correct distance to the pedestrians' destinations. The third kind of prediction, a pedestrian's path, led to similar results. Note that test persons determined the future position in the image which was then mapped to the ground plane by raycasting. This means the result is not biased by the lack of humans to numerically estimating distances[4].

The average Euklidean distances $\mathbb{E}_{\Omega_{tp}}(|\vec{x}_{hp} - \vec{x}_{GT}|)$ of **h**uman **p**redictions $\vec{x}_{hp}$ is larger than the networks error (Fig. 5 (a)). Test persons overestimate their ability to predict pedestrians. We derive that from the smaller uncertainties in Fig. 5 (c) and the fact that the ANN is a valid model according to Fig. 4 and has smaller errors according to 5 (a). The classification of sequences seems to be reasonable, since the humans' errors and uncertainties increase with the difficulty class. When comparing the results for the direction prediction, humans predict slightly better than the ANN.

C. Discussion

A new holistic prediction metric cannot be stated at the moment. A superiority of p_{Vol} over p_{Dot} is observed and numerically shown by all tested metrics. Still, it is necessary to evaluate destinations and uncertainties separately, because the observed advantage of l_{max} over l_{avrg} could only be shown with a goodness of fit test.

When comparing the combined human versus machine prediction, both are surprisingly close to each other. The mean error of ANN and humans have roughly the same

[4]Such a bias might be implicitely introduced by the numerically fixed prediction horizon of 2.5 sec, though.

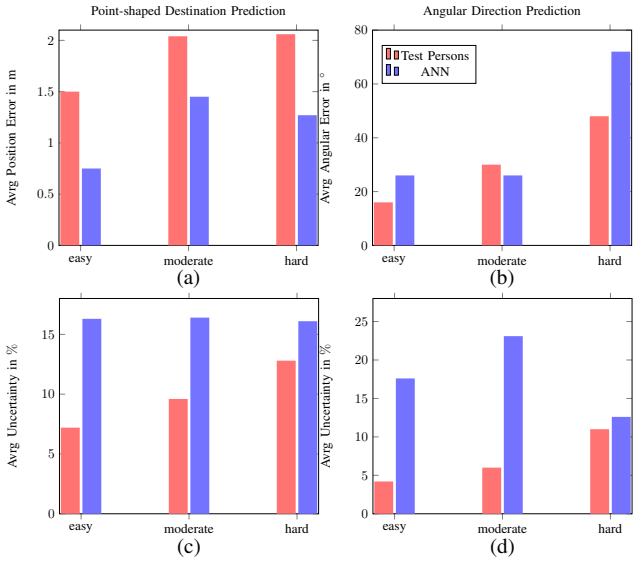


Fig. 5: Statistics of experiments with test persons. Left column: Position prediction, right column: Angle prediction. Upper row: Error, lower row: Uncertainties. For easier interpretability, the Euklidean distance to maximum mode of PDF is shown.

scale. The network, however, does not fully represent the human predictions. This is interpreted as a risk introduced by the artificial prediction. Even though neither the human prediction nor the uncertainty stated by the test persons seems to obviously correlate with GT, the human prediction could take information into account that is only implicitly available by observing the traffic situation as a whole. However, we neither sought nor casually found any indications of this conjecture.

Futhermore, better predictions of point-shaped destinations might stem from the numerical superiority of the ANN. The numerical differences of the results of ANN and humans are not considered large enough to derive a holistic superiority of the machine made prediction.

V. Conclusion

The present work summarizes new findings with regards to a better prediction metric, especially related to pedestrian prediction in public traffic. A collection of metrics recently used is presented and tested on a SOTA prediction method. A new loss function for Mixture Density Networks is presented (eq. 4). Furthermore, machine prediction was compared versus human prediction in a experiment with test persons. For our future work, we conclude with the following findings:

- Position *and* uncertainty need to be evaluated separately in order to assess a prediction holistically. The χ^2-test might be a suitable and well-established option.
- Learning from minimum or maximum (or a specific quartile) error might be an easy empirical regularization method in order to prevent uncertainty misestimation.
- Human and machine made predictions are surprisingly close to each other. Either current prediction approaches are already sufficiently good to drive with it or bad prediction skills are a major cause of man-made accidents.

Acknowledgements

We would like to thank Intel Corporation for supporting our work.

References

[1] European Commission, *2017 road safety statistics: What is behind the figures?*, 2018. [Online]. Available: `http://europa.eu/rapid/press-release_MEMO-18-2762_en.pdf`.

[2] D. Riedel, E. Rehder, M. Lauer, C. Stiller, and D. Wolf, "A literature review on the prediction of pedestrian behavior in urban scenarios", *IEEE Trans. Intell. Transp. Syst.*, 2018.

[3] N. Brouwer, H. Kloeden, and C. Stiller, "Comparison and evaluation of pedestrian motion models for vehicle safety systems", *IEEE 19th International Conference on Intelligent Transportation Systems*, 2016.

[4] N. Schneider and D. M. Gavrila, "Pedestrian path prediction with recursive bayesian filters: A comparative study", *Pattern Recognition: 35th German Conference*, 2013.

[5] S. Bonnin, T. H. Weisswange, F. Kummert, and Schmuedderich, "Pedestrian crossing prediction using multiple context-based models", *17th International IEEE Conference on In-telligent Transportation Systems (ITSC)*, 2014.

[6] S. Köhler, M. Goldhammer, S. Bauer, K. Doll, U. Brunsmann, and K. Dietmayer, "Early detection of the pedestrian's intention to cross the street", *15th International IEEE Conferenceon Intelligent Transportation Systems*, 2012.

[7] C. G. Keller and D. M. Gavrila, "Will the pedestrian cross? a study on pedestrian path prediction", *15th IEEE Transactions on Intelligent Transportation Systems*, 2014.

[8] R. Furuhashi and K. Yamada, "Estimation of street crossing intention from a pedestrian's posture on a sidewalk using multiple image frames", *Asian Conference on Pattern Recognition*, 2011.

[9] S. Köhler, M. Goldhammer, K. Zindler, K. Doll, and K. Dietmeyer, "Stereo-vision-based pedestrian's intention detection in a moving vehicle", *IEEE 18th International Conferenceon Intelligent Transportation Systems*, 2015.

[10] C. G. Keller, C. Hermes, and D. M. Gavrila, "Will the pedes-trian cross? probabilistic path prediction based on learned motion features", *33rd DAGM Symposium*, 2011.

[11] B. Völz, K. Behrendt, H. Mielenz, I. Hilitschenski, R. Siegwart, and J. Nieto, "A data-driven approach for pedestrian intention estimation", *IEEE 19th International Conference on Intelligent Transportation Systems*, 2016.

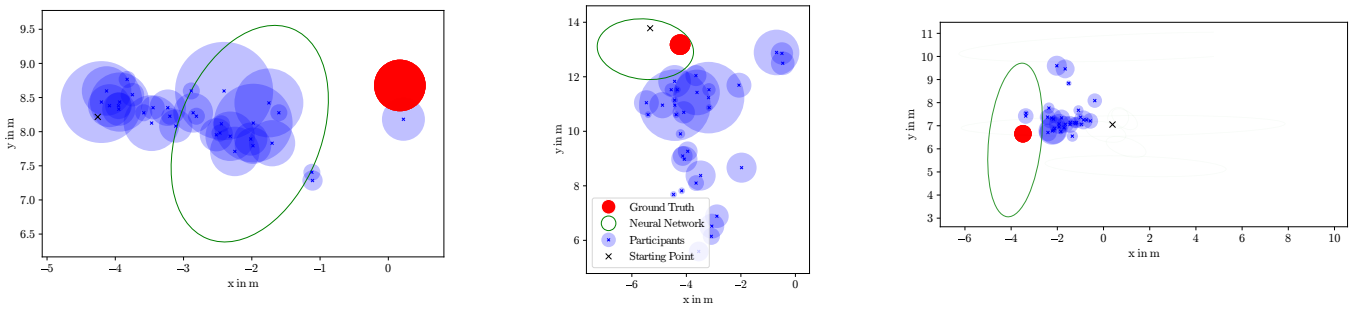


Fig. 6: Human point prediction of three examplenary pedestrians with different difficulties: easy (left), moderate (middle) and hard (right). Direction matches in most cases, actual distance is often misestimated by a factor of 2-3 both from the machine and the test persons.

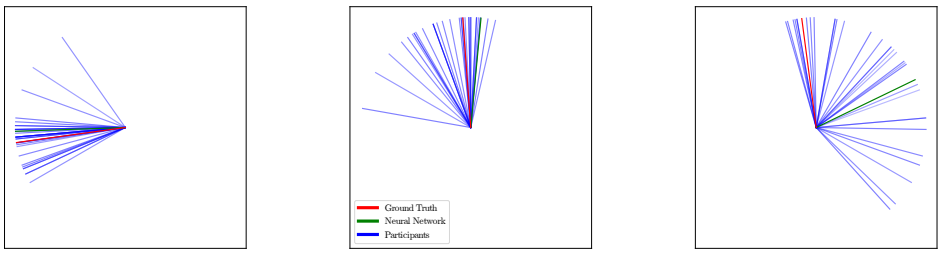


Fig. 7: Human direction prediction of three examplenary pedestrians with different difficulties: easy (left), moderate (middle) and hard (right). Direction of machine and human prediction matches roughly.

[12] J. Hariyono, A. Shahbaz, and K. H. Jo, "Estimation of walking direction for pedestrian path prediction from moving vehicle", *IEEE/SICE International Symposium on System Integration*, 2015.

[13] M. Goldhammer, S. Köhler, K. Doll, and B. Sick, "Camera based pedestrian path prediction by means of polynomial least-squares approximation and multilayer perceptron neural networks", *Intelligent Systems Conference*, 2015.

[14] A. T. Schulz and R. Stiefelhagen, "A controlled interactive multiple model filter for combined pedestrian intention recognition and path prediction", *IEEE 18th InternationalConference on Intelligent Transportation Systems*, 2015.

[15] R. Hug, S. Becker, W. Hübner, and M. Arens, "On the reliability of lstm-mdl models for pedestrian trajectory prediction", Dec. 2017.

[16] E. Rehder and H. Kloeden, "Goal-directed pedestrian prediction", *2015 IEEE International Conference on Computer Vision Workshop*, 2015.

[17] E. Rehder, F. Wirth, M. Lauer, and C. Stiller, "Pedestrian prediction by planning using deep neural networks", *IEEE International Conference on Robotics and Automation*, 2018.

[18] J. F. P. Kooij, N. Schneider, F. Flohr, and D. M. Gavrila, "Context-based pedestrian path prediction", *13th European Conference on Computer Vision*, 2014.

[19] S. Schmidt and B. Färber, "Pedestrians at the kerb – recognising the action intentions of humans", *Transportation Research Part F: Traffic Psychology and Behaviour*, 2009.

[20] C. M. Bishop, "Mixture density networks", *Neural Computing Research Group Report*, 1994.

[21] K. B. Petersen and M. S. Pedersen, "The matrix cookbook", Nov. 2012, Version 20121115. [Online]. Available: `http://www2.imm.dtu.dk/pubdb/p.php?3274`.

[22] D. Friedman, "Effective scoring rules for probabilistic forecasts", *Management Science No. 29*, 1983.

[23] G. Gneiting and A. E. Rafety, "Strictly proper scoring rules, prediction, and estimation", *Journal of the American Statistical Association 102*, 2007.

[24] R. Winkler and A. H. Murphy, ""good" probability assessors.", *Journal of Applied Meteorology*, 1968.

[25] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Wiley-Interscience, 1997, ISBN: 0-471-06259-6.

[26] K. Pearson, *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 50, no. 302, pp. 157–175, 1900. [Online]. Available: `https://doi.org/10.1080/14786440009463897`.

[27] D. S. Moore and J. B. Stubblebine, "Chi-square tests for multivariate normality with application to common", *Communications in Statistics - Theory and Methods*, vol. 10, no. 8, pp. 713–738, 1981.

[28] C. Henning, "Methods for merging gaussian mixture components", *Advances in Data Analysis and Classification, Vol. 4*, 2010.