How Good is My Prediction? Finding a Similarity Measure for Trajectory Prediction Evaluation.

Jannik Quehl, Haohao Hu, Ömer Şahin Taş, Eike Rehder, Martin Lauer

***Abstract*—The reliable prediction of traffic participants' trajectories is an important challenge for automated driving. Prediction methods that try to deal with this challenge need similarity measures for trajectories in order to evaluate the quality of their prediction. Currently, there exists no commonly accepted similarity measure suitable for this task. In this paper we review common trajectory similarity measures and analyze them with regard to prediction evaluation. Further we introduce a new approach for synthesizing a hybrid measure that combines a set of similarity measures and provide a heuristic to determine the parameters for this approach.**

I. Introduction

One of the fundamental problems in the area of automated driving and advanced driver assistance systems is the prediction of other traffic participants' behavior. Without a good estimate of a car's future surroundings it is very challenging to plan a suitable trajectory for the car. The most important parts of information about these surroundings are the position, orientation, velocity and acceleration of dynamic objects and traffic participants in particular. These information can be described by the trajectory of each object. This means that the essential part of behavior prediction for traffic participants in the context of automated driving consists mostly of trajectory prediction.

A major challenge for all trajectory prediction approaches is the evaluation and quantization of the performance of the approach. Since in practice an exact prediction of trajectories is not possible, metrics to quantify the difference – or similarity – between the predicted trajectory and the actual trajectory are needed in order to evaluate the performance of a prediction approach. However, currently there doesn't exist one commonly agreed upon metric that can be used to evaluate a trajectory prediction.

There have been several surveys on the topic of trajectory analysis and trajectory comparison. In 2006 Zhang et al. [1] compared different similarity measures like the Hausdorff distance, dynamic time warping (DTW) and longest common subsequence (LCSS) distance with respect to clustering properties in outdoor surveillance situations. The apparent conclusion was that a principle component analysis (PCA) with an Euclidean distance measure is sufficient for clustering in this context. In 2009 Morris et al. [2] examined a different set of metrics but also including the PCA approach, DTW, LCSS and a modified version of the Hausdorff metric (MODH). This examination came to the conclusion that LCSS and DTW perform better than the PCA-approach. More recent overviews can be found in [3] and [4]. However, none of these overviews focused on trajectory prediction evaluation and no overview provided a definite way to choose or create a suitable metric.

TABLE I
Trajectory Metrics

Similarity Measures	Publication
DTW	Keogh 2000 [5]
Hausdorff	Lou 2002 [6]
LCSS	Buzan 2004 [7]
MED	Hu 2007 [8]
PCA-Euclid	Bashir 2007 [9]
Traj. Hausdorff	Lee 2007 [10]
MODH	Atev 2010 [11]
CLEAR-MOTA	Bernadin 2008 [12]

An overview over the most relevant distance measures used in the context of trajectory analysis can be found in table I. In this paper we provide an analysis of the most common existing distance measures that are used in the context of trajectory analysis and their strengths and deficiencies. Further, we propose an approach to combining different metrics to a new metric that can be applied in the context of trajectory prediction analysis and a heuristic that can be used to determine its parameters.

II. Notation and Definitions

Since we compare and analyze several different similarity measures it is first necessary to define a common notation in order to prevent inconsistencies. In general a trajectory $T = ((a_{1t_1}, a_{2t_1}...), (a_{1t_2}, a_{2t_2}...)...)$ is a temporally ordered sequence of a set of attributes $(a_1...a_n)$ which describe the movement state of an object at given time stamps $(t_1, t_2...)$. We define the path of a trajectory as the unordered set of all positional information contained in the trajectory, i.e. a trajectory without time stamps. In most cases trajectories can be simplified by assuming that t starts with 1 and $t_n - t_{n-1} = 1$ for all n and by reducing the set of attributes to a list of points in two dimensional space (and in some cases the first and second order derivatives): $T = ((x_1, y_1, \dot{x}_1, \dot{y}_1, \ddot{x}_1, \ddot{y}_1)...)$ or $T = (\vec{p}_1, \vec{p}_2...)$ with $\vec{p}_t = (x_t, y_t)$. Next we define $T^{i,j} = (\vec{p}_i...\vec{p}_j)$ as the partial trajectory of $T = (\vec{p}_1...\vec{p}_i...\vec{p}_j...)$ with beginning index i and final index j.

A similarity measure $m(T_p, T_g)$ in our context is a function that compares the predicted trajectory T_p to the ground truth trajectory T_g and returns a value as indicator for the difference in both trajectories. The higher the difference is the larger is the return value of m. Further, it should hold that $m(T_1, T_2) = 0 \Leftrightarrow T_1 = T_2$ and $m(T_p, T_g) \geq 0$. Since in this context one trajectory is always a prediction while

the other trajectory is the corresponding ground-truth, both trajectories have different semantic meaning. This means that a comparison of T_p to T_g does not necessarily carry the same meaning a comparison of T_g to T_p. Thus similarity measures do not have to be symmetric.

III. Similarity Measures

In this work, we rely on a set of readily available metrics and similarity measures which are presented in the following in detail. A fundamental distinction for the similarity measures in Table I is whether they evaluate based on the trajectory or based on its path. In the context of this work we define all similarity measures that perform their evaluation considering temporal features (eg. velocity or order of positions) as trajectory measures. If they on the other hand only regard spatial differences between trajectories we define them as path measures.

A. Euclidean Mean

One example for why the distinction between path and trajectory measures is important is the mean Euclidean distance (MED). MED can be used as a path or a trajectory measure depending on which points are compared with each other. If used as a trajectory similarity measure (MEDT), it is necessary to define a point in time for each trajectory from which onwards the comparison starts. Each point of the predicted trajectory is then compared with the point on the ground truth that has the same temporal distance from their respective starting points. For this to be possible both trajectories have to contain the same number of time steps $n_p = n_g = n$ with the same time between each time step. This means that for starting point t_{p1} of T_p and starting point t_{g1} of T_g MED is defined as

$$m_{\mathrm{MEDT}}(T_p, T_g) = \frac{1}{n} \sum_{k=0}^{n-1} \|p_{p,t_{p1}+k} - p_{g,t_{p1}+k}\|_2. \quad (1)$$

The obvious problem with this similarity measure is that a temporal misalignment leads to values larger than 0 even if both trajectories are identical. Further, two trajectories that follow the exact same path can get an arbitrary large distance value if the velocities are different enough. Temporal misalignment can in the context of trajectory prediction be prevented by dictating the timestamps for which the prediction should occur. For some applications, however, it is undesirable to get large distances for different velocities. If the prediction for example is used in order to check if a pedestrian will cross the road, the velocity with which he does that is not necessarily relevant while the path he takes is.

The alternate form of MED measures the difference between the predicted paths (MEDP). For each point p_p of a predicted trajectory T_p with time steps $1...n_p$ MEDP searches for the nearest point in the ground truth trajectory and evaluates the Euclidean distance to that point

$$m_{\mathrm{MEDP}}(T_p, T_g) = \frac{1}{n_p} \sum_{k=1}^{n_p} \min_x \|p_{p,k}, p_{g,x}\|_2. \quad (2)$$

This version of the MED has the property that it is robust towards velocity errors and temporal misalignments. Additionally, MEDP is able to compare trajectories of different length. If two trajectories follow the same path but with different velocities they will be evaluated as very similar by MEDP while MEDT finds them dissimilar. MEDP is additionally robust towards periods of stopped motion. The path of a pedestrian waiting for a traffic light to change can easily be predicted while the actual duration how long the pedestrian will wait is significantly harder to predict. However, by using MEDP the temporal order of points is completely lost. This means that a trajectory $T = (\vec{p}_1, \vec{p}_2 ... \vec{p}_n)$ can not be distinguished from trajectories with the same points, e.g. from its own reverse version $T' = (\vec{p}_n, \vec{p}_{n-1} ... \vec{p}_1)$. This is a major problem inherent to path similarity measures.

A third variation of MED proposed in [9] is a trajectory similarity measure that computes the Euclidean distance in lower dimensional parameter space (PCA-Euclid). The x and y coordinates of each point are transformed in one-dimensional space by principle component analysis (PCA) decomposition. The PCA-Euclid measure is computed as the mean Euclidean distance between PCA coefficients of both trajectories. This version of MED requires reduced computational effort and offers some more robustness compared to MEDT via the PCA shape decomposition [2].

B. Dynamic Time Warp

The dynamic time warping distance (DTW) as proposed in [5] is a trajectory measure that can be used on general time series. Unlike MEDT, DTW does not require both trajectories to have the same length. Instead DTW measures the temporal changes that are necessary in order to warp one trajectory into another. In order to do that, DTW first finds for every point on T_p a matching point on T_g. Then dynamic programming is used in order to find a time warping that minimizes the total distance between these pairs of points. DTW then evaluates the length of the warping path that is needed in order to warp T_p into T_g.

$$m_{\mathrm{DTW}}(T_p, T_g) = \min \sqrt{\sum_{k=1}^{K} \frac{w_k}{K}} \quad (3)$$

With $w_k \in W$ as possible warping path with length K that warps T_p into T_g. DTW is especially useful if trajectories with different lengths are compared. In the context of prediction evaluation, however, we can ensure that the predicted trajectory is as long as the ground truth trajectory. However, DTW is very susceptible to outliers and is not very robust [3].

C. Hausdorff and modified Hausdorff

The simple Hausdorff distance (HAU) is a similarity measure used to compare sets and is commonly used in the geometric calculation and image processing domain [6]. In the context of trajectory similarity measures it describes a path similarity measure since it does not account for an ordering of either sets it compares. The simple Hausdorff measure

describes the maximal minimal distance between the points in T_p and T_g.

$$\begin{aligned} m_{\mathrm{HAU}}(T_p, T_g) &= \max\{d_h(T_p, T_g),\ d_h(T_g, T_p)\} \\ d_h(T_p, T_g) &= \max_{p_p \in T_p} \min_{p_g \in T_g} \|p_p - p_g\|_2 \end{aligned} \tag{4}$$

It can be used for trajectories with differing lengths but suffers from the same drawbacks as all other path measures: Trajectories that have a high spatial proximity will be scored as highly similar even if the movement performed is different e.g. in the opposite direction. Moreover HAU is very sensitive to outliers because of the prominent use of $\min/\max$ functions.

The modified Hausdorff measure (MODH) aims to reduce these drawbacks by adding a correspondence function C which controls the trajectory alignment by forcing matching points to occur at a similar fraction of total trajectory length and a neighborhood window which allows for some slack and more robustness [11]. This modified version is neither a true path measure since it accounts for a loose ordering of position nor a true trajectory measure since time related aspects only indirectly influence the result of the measure.

D. Longest Common Subsequence

The longest common subsequence (LCSS) measure is a trajectory measure that evaluates for how many time steps two trajectories are *matching* each other [7]. In order to compute LCSS, an alignment is searched that maximizes the length of common subsequence. A LCSS based similarity measure that evaluates what fraction of all points belonging to the shorter trajectory can be matched to the longer trajectory can be defined as

$$m_{\mathrm{LCSS}}(T_p, T_g) = 1 - \frac{\mathrm{LCSS}(T_p, T_g)}{\min(n_p, n_g)} \tag{5}$$

with $\mathrm{LCSS}(T_1, T_2)$ denoting the number of matching points between both trajectories. In order to define what constitutes a matching between points, LCSS uses two thresholds that determine how large the maximum distances in time and space are allowed to be in order for two points to match. Finding suitable thresholds is not easy and largely defines how well LCSS performs in practice. If the thresholds are chosen too large, a lot of trajectories will be deemed completely identical while too small thresholds may lead to very similar trajectories getting a large distance.

E. CLEAR-Multi Object Tracking Acurracy

The CLEAR multiple object tracking accuracy (CLEAR-MOTA) is as the name implies a measure that is usually used in order to evaluate how well a multi object tracking algorithm performs [12]. However, in the context of prediction evaluation it is possible to define matches for each predicted point similar to the LCSS measure. Based on these matchings CLEAR-MOT Accuracy can be defined as

$$m_{\mathrm{MOTA}}(T_p, T_g) = 1 - \frac{\sum_t (m_t + fp_t + mme_t)}{n_p} \tag{6}$$

with m_t, fp_t and mme_t as number of misses, false positives and mismatches when matching T_p to T_g. A miss in this context occurs for every point in T_g that did not get matched. A false positive is defined as a match to a point that is too far away (defined by a threshold) to constitute a match. A mismatch happens if more than one point in T_p is matched to the same point in T_g.

Similar to LCSS this measure describes which fraction of predicted points can be matched to the ground truth without an error. Unlike LCSS, however the total fraction of matches is regarded instead of the longest sequence of matches. Further the definitions for match/mismatch are complexer compared to LCSS. CLEAR-MOTA is a trajectory measure since velocity differences influence the result of the evaluation. The larger the velocity difference the more misses and mismatches are found. CLEAR-MOTA requires a spatial threshold similar to LCSS to define what constitutes a point matching making it also susceptible to similar problems as LCSS.

F. Trajectory Hausdorff

The trajectory Hausdorff similarity measure (THAU) describes a path measure approach that consists of a weighted sum of several path distances [10] $d_\perp, d_{||}, d_\theta$.

$$\begin{aligned} m_{\mathrm{THAU}}(T_p, T_g) =& w_\perp d_\perp(T_p, T_g) + w_{||} d_{||}(T_p, T_g) \\ &+ w_\theta d_\theta(T_p, T_g) \end{aligned} \tag{7}$$

Each of these path distances focuses on a different aspect of the path. The orthogonal distance $d_\perp$ measures the separation between both trajectories, $d_{||}$ describes the difference in length of both trajectories and d_θ captures the difference in orientation [3]. How the weights $w_\perp$, $w_{||}$ and w_θ are chosen depends on the application and data sets. In many cases $w_\perp = w_{||} = w_\theta = 1$ can be used in order to get acceptable results. As this similarity measure is a pure path measure it suffers from the same limitations as m_{MEDP} or m_{HAU}.

IV. A Novel Trajectory Similarity Measure

Each of the previously introduced similarity measures has its own advantages and disadvantages. The problem in choosing or defining a suitable similarity measure for prediction evaluation is that each of these measures will usually produce quite different results when applied to two trajectories. If the prediction approach that is to be evaluated for example has a high accuracy for the velocity of the predicted object, DTW will evaluate this prediction as very good. However, a path measure like the Hausdorff may yield significantly worse results depending on the accuracy of the spatial prediction.

In order to avoid a (positive or negative) bias caused by choosing a similarity measure based on the prediction method, it is necessary to make the choice of similarity measure independent of the prediction approach. Instead it is reasonable to make a choice based on the data set that is used as ground truth for the evaluation. If the data set for example includes many car trajectories that take a turn at crossroads, it would be reasonable to use a similarity measure that incorporates the change in direction into its result. Similarly, a data set

which includes a lot of different velocities should not solely be evaluated based on path similarity measures.

Since there can always be aspects relevant to the prediction approach, data set or trajectory format that are not incorporated in a single similarity measure a general approach to trajectory prediction evaluation should be configurable to the specific situation. In order to enable this, a combination of different measures similar to the THAU approach can be used to define a new metric that combines the considered features

$$m_{GPE}(T_p, T_g) = w_1 m_1(T_p, T_g) + ... + w_n m_n(T_p, T_g) \quad (8)$$

with $m_1 ... m_n$ different path and trajectory similarity measures and $w_1 ... w_n$ their respective weights. This equation defines the generalized prediction evaluation (GPE) trajectory measure. Such an approach requires an appropriate selection of similarity measures together with their weights. One problem in this context are the different value ranges that the similarity measures produce. The LCSS and MOTA approaches, for example, produce results between 0 and 1 while both MED approaches can produce arbitrary large numbers for sufficiently distant trajectories. This implies that the weights for these similarity measures have to compensate for different ranges, scales or units used for each trajectory. For this to be possible, some quantitative analysis of the proportional differences and general performance for these measures is necessary. Further, the weights of the similarity measures should as previously mentioned incorporate the specific properties of the evaluation dataset with which they should be used.

In order to be able to on one hand find weights based on the data, and on the other hand compare their result ranges we propose the following procedure:

1) choose a prediction horizon h and a prediction basis b and create an evaluation set $\{T\}$ of all trajectories T_i with $|T_i| \geq b + h$
2) choose a set of similarity measures $M = m_1 ... m_n$ that use as different features as possible for their evaluation
3) For each trajectory $T_i \in \{T\}$ and for each similarity measure m_j:
   a) split the trajectory T_i in two parts $T_i^{1,b}, T_i^{b+1,n_1}$
   b) find the trajectory $T_2 = \text{argmin}_{T_k} m_i(T_i^{1,b}, T_k^{1,b})$, $i \neq k$ that has the most similar partial trajectory $T_2^{1,b}$ with respect to m_i
   c) Evaluate the similarity $m_j(T_i^{b+1,b+h}, T_2^{b+1,b+h})$ for each other similarity measure $m_l \in M, j \neq l$
4) Calculate the average similarity for each combination of similarity measures.

This procedure is in essence an evaluation of a nearest neighbor prediction approach. For each similarity measure m_i a nearest neighbor predictor is initialized and the resulting predictions ($T_2^{b+1,b+h}$) are compared to the ground truth using each of the similarity measures. This produces a table in which the row i describes the similarity measure $m_i \in M$ used for prediction. Each column j, on the other hand, describes the average of how the prediction was assessed by a similarity measure $m_j \in M$. This table can then be used to estimate how suitable the features used by each evaluated similarity measure are for trajectory prediction and evaluation.

In order to be able to make quantitative statements about similarity measures, first the table has to be normalized. By dividing each value of the table by the minimum value in each column $\min_j$ a normalized table A^N is generated for which all entries a_{ij} in row i and column j are larger or equal to 1. Each entry is now a factor that describes how much worse than the best prediction the measure j evaluates the prediction of measure i.

The first step to determine weights from the table is to analyze the primary diagonal. If there are entries that are significantly larger than 1 this means that the features used by the metric are not stable enough to provide a good prediction and by extension a good evaluation of predictions. The weight of the measures corresponding to such entries is set to 0. Second, by comparing each of the entries of a column similarity measures that do not contribute in a meaningful way can be found. If for a column all entries are close to 1 this means that the metric in average does not yield different results depending on the selected prediction method. These measures can also be ignored and set to 0. All measures which now have their weight set to 0 are deleted from the table resulting in a new table A'^N

The remaining measures in A'^N can still be considered for the weighted sum. A mainly negative evaluation of a prediction made by measure m_i is an indicator that the features that m_i analyses are not very useful in the context of prediction and should be weighted less. If a measure m_j on the other hand evaluates other measures as negative it implies that m_j evaluates aspects that are ignored by the other measures and m_j should be rated higher. By evaluating the average of each column i and each row j two values are produced: The average of how this measure was scored by all remaining measures $s_{m_i} = \frac{\sum_{j=1}^{n'} a_{ij}}{n'}$ and the average how this measure evaluated by other measures $e_{m_i} = \frac{\sum_{j=1}^{n'} a_{ji}}{n'}$. The ratio of these two numbers combined with the normalization factor $\min_i$ that was used for generating the normalized table yields the final weight for each measure:

$$w_i = \frac{s_{m_i}}{e_{m_i}} \cdot \frac{1}{\min_i} \quad (9)$$

A. Discussion of Approach

The presented approach to combining trajectory similarity measures has both advantages and disadvantages. One of the main advantages is that it provides a method to choose an evaluation measure for trajectory prediction that is not biased towards the chosen prediction approach. However, the resulting measure still exhibits a bias based on the data set. A measure created with the proposed approach may not yield the same performance if it is used on a different data set. Because of this, our approach needs re-calibration whenever the underlying data set is modified. However this is not a problem e.g. in the context of benchmark data sets in which the data won't change.

One disadvantage of this approach is that it is partly based on the assumption that features that are suitable to use for prediction are also features that are predicted well. This is the case for most features that are evaluated in the presented similarity measures. However, if similarity measures are combined for which this assumption does not hold, the resulting measure may not provide a good evaluation. Further, the quality of the resulting measure mainly depends on the measures that are combined. If several measures are chosen that are not suitable for prediction evaluation, the resulting measure will probably exhibit the same problems. Another drawback of such a hybrid measure is that unlike e.g. MED its value has no semantic meaning other than a score that can be compared to other trajectory pairs evaluated the same way.

Lastly it should be noted that this approach should be seen as a heuristic. Since there is no objective quality measure for similarity measures, it is not possible to find an optimal evaluation metric. However, the proposed approach and the analysis of similarity measures should prove useful for future evaluations of trajectory prediction approaches.

V. Evaluation

The evaluation of our similarity measure is done in form of a case study on a trajectory data set recorded at a busy intersection in Karlsruhe, Germany. The data set is created with an experimental vehicle equipped with a Velodyne LiDAR HDL-64 [13] scanner and parked next to the intersection. The pointclouds received from the laser scanner are segmented by removing the ground plane and then executing a euclidean cluster extraction provided by the point cloud library (PCL) [14]. The resulting objects are tracked for as long as they are observable with a simple tracking algorithm based on a constant velocity motion model and a nearest neighbor association approach. Based on the motion we further extract an orientation and velocity for each observed time frame. All tracked objects that moved during the observation are manually labeled and classified either as pedestrian, bike, car or neither.

As prediction basis and horizon we chose 5 time frames since that number for this data set is enough to evaluate whether changes to the trajectory are correctly predicted or not. Further, a relatively short time frame allows for a splitting of longer trajectories in order to get more data. The result of this observation is a data set of about 16500 trajectories with a length of at least 10 frames: 5660 car trajectories, 4647 bike trajectories and 6212 pedestrian trajectories. A visualization of all tracked trajectories are plotted in Figure 1. Since the movement patterns of each class of traffic participants greatly differ the data set was split into one data set for each class of traffic participants.

The set of metrics that are to be combined in our evaluation is chosen based on their inherent properties and features they compare: To test the scaling properties of the proposed heuristic a measure with bounded output is useful. This leaves m_{MOTA} and m_{LCSS}. Since both are trajectory measures and compare similar features and since m_{MOTA} only needs one threshold parameter it was preferred over m_{LCSS}. The second and third measures were selected as the euclidean mean distances m_{MEDT} and m_{MEDP}. This choice is made since both consider spatial similarity in a comparable way while being different enough based on the fact that one is a trajectory measure and one a path measure. Further, the euclidean distance can be considered as the most intuitive measure which warrants a comparison. Since the euclidean distances mostly cover spatial similarity we decided to exclude m_{HAU} and m_{MODH}. The trajectory Hausdorff metric, on the other hand, additionally uses orientation similarity d_θ. Because this is not covered in the selection we include this in our weighted sum. Since the prediction horizon is the same throughout the experiment and m_{DTW} is mainly useful in situations where the compared trajectories' length differ, m_{DTW} was not included in the set. In order to test these established similarity measures a further metric, the average velocity difference m_{AVD}, is added. This measure only regards the velocity difference as feature ignoring even spatial similarity.

TABLE II
Normalized results for bike trajectories

	m_{MOTA}	m_{MEDP}	m_{MEDT}	d_θ	m_{AVD}
m_{MOTA}	1.000	2.600	2.152	6.195	2.074
m_{MEDP}	1.014	1.000	1.000	5.859	2.052
m_{MEDT}	1.017	1.190	1.108	6.064	1.992
d_θ	1.104	9.754	7.661	1.000	2.793
m_{AVD}	1.111	11.619	9.059	8.382	1.000

TABLE III
Final normalized weights for each class of traffic participants

	m_{MOTA}	m_{MEDP}	m_{MEDT}	d_θ	m_{AVD}
car	0.000	8.396	5.272	2.678	1.500
bike	0.000	2.377	1.818	1.005	0.261
pedestrian	0.000	3.117	2.241	0.759	0.229

A. Discussion of Results

The normalized results of the evaluation of bike trajectories are presented in Table II. As can be seen on the primary diagonal of the table, all similarity measures score a prediction made with the same metric as very good (~ 1). Interestingly predictions made by m_{MEDP} are evaluated slightly better by m_{MEDT} than m_{MEDT}'s prediction itself. This leads to the conclusion that in the context of this dataset the exact location of each point is more relevant than e.g. the velocity for trajectory prediction. Since all values on the primary diagonal are close to 1 no measure can be discarded in the first step of analysis. However, the analysis of the columns reveal that m_{MOTA} yields very small differences for each prediction approach. This result was also observed for the evaluation of the car and pedestrian data sets. This means that m_{MOTA} may not be suitable in the context of this data set. The cause for this might either be a unsuitable threshold – in this instance 0.5m maximal distance for a match was chosen – or a general problem with this measure in the context of prediction.

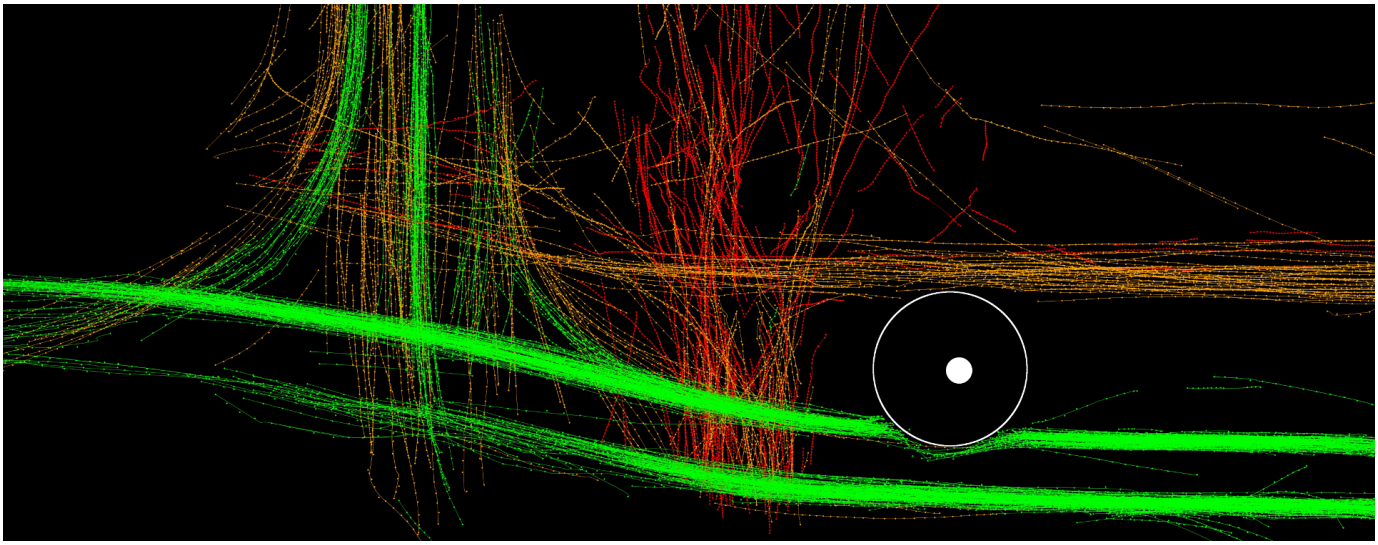

Fig. 1. Evaluation Data Set. Green, red, orange respectively indicate cars, pedestrians and bikes. White indicates the position of the sensor and nearest possible observations.

The final normalized weights are depicted in Table III. The results show that for all three classes the orientation evaluated with m_θ and the velocity evaluated with m_{AVD} receive comparatively low weights compared to both euclidean distances. This leads to the conclusion that for this data set the prediction of all traffic participants should mainly be performed on basis of euclidean differences between predicted and actual trajectory. This seems plausible since an evaluation mainly evaluating velocity or differences in direction are improbable to produce meaningful results.

VI. Conclusion

In this paper we analyzed different approaches to trajectory comparison in the context of trajectory prediction evaluation. We reviewed eight established similarity measures by highlighting their similarities and differences, and subsequently inspected their applicability to our context. Based on this, we proposed a new similarity measure as a weighted sum of existing measures and a heuristic that can be used to determine suitable values for the weights. After we discussed the advantages and disadvantages of our approach, we evaluated the proposed heuristic with a set of five similarity measures on a data set we recorded. We concluded the paper by discussing the plausibility of our results.

Acknowledgment

The research leading to these results has received funding from the German collaborative research center “SPP 1835 - Cooperative Interacting Automobiles” (CoInCar) granted by the German Research Foundation (DFG).

References

[1] Z. Zhang, K. Huang, and T. Tan, “Comparison of similarity measures for trajectory clustering in outdoor surveillance scenes,” in *Proceedings of the 18th International Conference on Pattern Recognition. ICPR 2006*, vol. 3. IEEE, 2006, pp. 1135–1138.

[2] B. Morris and M. Trivedi, “Learning trajectory patterns by clustering: Experimental studies and comparative evaluation,” in *IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2009.* IEEE, 2009, pp. 312–319.

[3] Y. Zheng, “Trajectory data mining: an overview,” *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 6, no. 3, p. 29, 2015.

[4] X. Pan, Y. He, H. Wang, W. Xiong, and X. Peng, “Mining regular behaviors based on multidimensional trajectories,” *Expert Systems with Applications*, vol. 66, pp. 106–113, 2016.

[5] E. J. Keogh and M. J. Pazzani, “Scaling up dynamic time warping for datamining applications,” in *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2000, pp. 285–289.

[6] J. Lou, Q. Liu, T. Tan, and W. Hu, “Semantic interpretation of object activities in a surveillance system,” in *Proceedings of the 16 th International Conference on Pattern Recognition. ICPR 2002*, vol. 3. IEEE, 2002, pp. 777–780.

[7] D. Buzan, S. Sclaroff, and G. Kollios, “Extraction and clustering of motion trajectories in video,” in *Proceedings of the 17th International Conference on Pattern Recognition. ICPR 2004*, vol. 2. IEEE, 2004, pp. 521–524.

[8] W. Hu, D. Xie, Z. Fu, W. Zeng, and S. Maybank, “Semantic-based surveillance video retrieval,” *IEEE Transactions on Image Processing*, vol. 16, no. 4, pp. 1168–1181, 2007.

[9] F. I. Bashir, A. A. Khokhar, and D. Schonfeld, “Object trajectory-based activity classification and recognition using hidden markov models,” *IEEE Transactions on Image Processing*, vol. 16, no. 7, pp. 1912–1919, 2007.

[10] J.-G. Lee, J. Han, and K.-Y. Whang, “Trajectory clustering: A partition-and-group framework,” in *Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data*. ACM, 2007, pp. 593–604.

[11] S. Atev, G. Miller, and N. P. Papanikolopoulos, “Clustering of vehicle trajectories,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 11, no. 3, pp. 647–657, 2010.

[12] K. Bernardin and R. Stiefelhagen, “Evaluating multiple object tracking performance: The clear mot metrics,” *Journal on Image and Video Processing*, no. 1, pp. 1–10, 2008.

[13] Velodyne LiDAR, “Velodyne HDL-64E Data Sheet,” [last accessed 20-April-2017]. [Online]. Available: ”http://velodynelidar.com/docs/datasheet/63-9194_Rev-F_HDL-64E_S3_Data%20Sheet_Web.pdf”

[14] R. B. Rusu and S. Cousins, “3D is here: Point Cloud Library (PCL),” in *IEEE International Conference on Robotics and Automation*, May 9-13 2011.