IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2024.0429000

Complementarity-Oriented Feature Fusion for Face-Phone Trajectory Matching

CHANGFENG CAO[1], WENCHUAN ZHANG[1], HUA YANG[1], DAN RUAN[2],
[1]Department of Big Data and Computer Science, Guizhou normal university, Guiyang, Guizhou 550000, China
[2]Department of Radiation Oncology, David Geffen School, Department of Bioengineering, Henry Samueli School of Engineering, University of California, Los Angeles, California, CA 90095, USA

Corresponding author: Hua Yang (e-mail: yanghuastory@whu.edu.cn).

C.C. and W.Z. contributed equally to this work. Writing-original draft, C.C. and W.Z.; Writing-review and editing, C.C. and W.Z.; Software, W.Z.; Conceptualization, methodology, supervision, H.Y.; Supervision, D.R.. All authors have read and agreed to the published version of the manuscript.
This work is supported by the Natural Science Foundation Project of China (Grant No. 61070243); Guizhou High-level Talent Research Project (Grant No. TZJF-2010-048);

ABSTRACT CCTVs and telecom base stations act as sensors, and collect massive face and phone related data. When used for person localization and trajectory characterization, they each present quite different spatiotemporal characteristics: CCTV is associated with slowly sampled face ID trajectories with spatial resolution of approximately 20 meters, while telecom readings provide fast sampled phone ID trajectories with spatial uncertainty of a few hundred meters. Face or phone trajectory can be seen as an observation of the real trajectory of a moving pedestrian. It is useful to identify correspondence between face and phone trajectories to reconstruct trajectory of moving persons. To this end, we propose a complementarity-oriented feature fusion mechanism (COFFM) to model and utilize the common embedding and complementarity of these two measurement modalities. Specifically, a Cycle Heterogeneous Trajectory Translation Network (CCTTN) is proposed to realize a TFE (Trajectory Feature Extractor) which captures the latent transforming relationships between the face and phone modalities. The latent features from both transforming directions are concatenated in the Feature Unifying (FU) module and fed into a binary face-phone trajectory matching discriminator (FPTPMD) to infer whether a face-phone trajectory pair corresponds to the same underlying motion trajectory. We evaluated our method on a large real-world face-phone trajectory data set, and showed promising results with the accuracy of 97.1% which exceeds the comparable similarity-based methods. The developed principle and framework generalize well to other multi-modality trajectory matching tasks.

INDEX TERMS Multi-modality trajectory matching, feature fusion, trajectory feature extraction, common domain embedding, pedestrian tracking, trajectory reconstruction.

I. INTRODUCTION

RECONSTRUCTION of pedestrian trajectory is an important task in tracking human subjects, for security or safety purposes. Security CCTV can capture a person's facial image when the subject is within the view range, typically within a radius of about 20 meters. Face ID is typically generated with face recognition technique, and coordinates of camera when such identifying frame is captured approximate the subject's location associated with the same time tag. Therefore, querying face ID from the CCTV footage can provide an approximate trajectory of the pedestrian. Similarly, as mobile phones move, their communication routing via the telecom base stations are also recorded and their geographic trajectory can be approximated by the location of their closest base station, tagged by the time of communication. If a mobile phone is associated with a pedestrian, then both the CCTV-derived face ID trajectory and the telecom-derived phone ID trajectory manifest from the same underlying continuous pedestrian movement and can be considered as two observation modalities, each with its own characteristics. The face trajectory has a high geographic resolution but a low temporal sampling rate, available only when the pedestrian is within 20 meters of a camera. On the other hand, the mobile phone trajectory has a high temporal sampling rate but at the cost of low geographic position resolution, in the order of a few hundred meters.

These two observation modalities provide complementary information about the pedestrian trajectory and it is desirable

to integrate them for the underlying reconstruction problem. One critical task for this Integration is to identify the corresponding face-phone trajectory pairs among the massive raw data.

Existing works on trajectory matching mostly focus on mono-modality trajectory data collected by same kind of sensors [1]–[11] typically relying on measurement of (dis)similarities between trajectories to establish correspondence. Unfortunately, such setup does not translate well to the face-phone matching problem where spatiotemporal resolution differ significantly across modalities. To fill in this critical gap between available methodology and application gap, we propose an innovative Complementarity-Oriented Feature Fusion Mechanism (COFFM) framework as shown in Fig 1. At the center of the COFFM is a Cycle Consistent Trajectory Translation Network (CCTTN) for bi-directional embedding of sequence-encoded features from both modalities. These features are concatenated and fed into the Face-Phone Trajectory Pair Matching Discriminator (FPTPMD, yellow in Fig 1) to distinguish true corresponding face-phone trajectories from other pairings. Experiments on large scale data have demonstrated that our complementarity-based approach achieves good performance for face-phone trajectory matching with the accuracy of 97.1%.

This framework aims to address the modality trajectory matching problem, particularly with different spatiotemporal characteristics. The central Cycle Consistent Trajectory Translation Network uses LSTM as basic building blocks for dynamic information encoding and constructs a mutual embedding to capture the complementarity of the modalities. Subsequent discriminator based on the concatenation of abstract features from CCTTN ensures that both the common and differential information are efficiently utilized. While this work is motivated by the face-phone matching problem, our development and design generalize to other multi-modality trajectory fusion problems quite naturally.

II. RELATED WORKS

A. SIMILARITY-BASED METHOD FOR MONO-MODALITY TRAJECTORY MATCHING

Most existing work on trajectory matching focus on trajectories collected from the same type of sensors. For example, reference [12] worked on travel recommendations by mining multiple pedestrians' GPS traces. The common theme is to investigate similarity metric to define difference between two trajectories [1]–[11], with the possibly semantic considerations. These methods typically require domain knowledge to design similarity metrics on pairing trajectories with compatible lengths and sampling rates. In contrast, our work automatically extracts features for multi-modality sampled trajectories originated from the same underlying continuous movement by leveraging deep network, and performs trajectory matching without the need for ad-hoc sample alignment across the different trajectories.

B. MULTI-MODALITY TRAJECTORY MATCHING

Multi-modality trajectory matching has been a focal point of research, given the potential of combining diverse data sources to enhance tracking precision. Several studies, including [13], [14], and [15], have explored methods for associating heterogeneous trajectories. Among these, the work in [13] is particularly relevant to ours, as it presents a feature engineering-based approach for matching face and phone trajectories, utilizing spatiotemporal features such as Multi-Granularity SpatioTemporal Window Searching (MGSTWS) and Big GeoDis and Small TimeDiff (BGST) to identify potential matches. While effective in controlled settings, this method is constrained by its reliance on manually crafted features, which limits adaptability in scenarios with varying spatial resolutions and temporal frequencies—a challenge commonly encountered in face-phone trajectory data. In contrast, our approach employs a data-driven feature fusion mechanism, designed to flexibly integrate cross-modality information and enhance robustness in complex environments.

In addition, [15] investigates vehicle-phone trajectory matching, employing multiple similarity metrics to increase robustness. However, this approach does not address key challenges specific to face-phone matching, such as differing statistical properties between data sources and the need for modality complementarity. Likewise, [14] introduces Vi-Fi, a framework that associates visual and wireless data through a deep learning-based similarity matrix, effectively linking camera-detected individuals with their corresponding wireless signals. Although Vi-Fi mitigates some limitations of vision-based matching (e.g., reliance on color features) and demonstrates adaptability in dynamic conditions, it primarily targets visual-wireless alignment and does not tackle the unique requirements of face-phone data.

Our work advances beyond these methods by addressing the distinct challenges of face-phone trajectory matching through an approach that unifies and maximizes complementary information across modalities, enabling more effective and adaptive trajectory alignment.

C. TRAJECTORY EMBEDDINGS

Trajectory embedding is necessary to provide a homogenized inputs from different modalities by transforming raw trajectories into a common linear vector space. The design of this trajectory embedding should account for the data characteristics – in the current context, the face ID/phone ID raw trajectories are presented as variable length sequences and needs to be converted to a homogeneous form for further processing.

There are few works concerning this trajectory embedding process. Reference [16] uses Recurrent Neural Network to characterize trajectory. In contrast, we use LSTM to learn the trajectory embedding since LSTM excels in extracting formation in long sequences. References [17]–[19] learn a mapping function to transfer trajectories into low-dimensional vector representations based on only geographical correlativeness in trajectories. In contrast, our framework considers not only geographical correlation but also temporal correlation in tra-

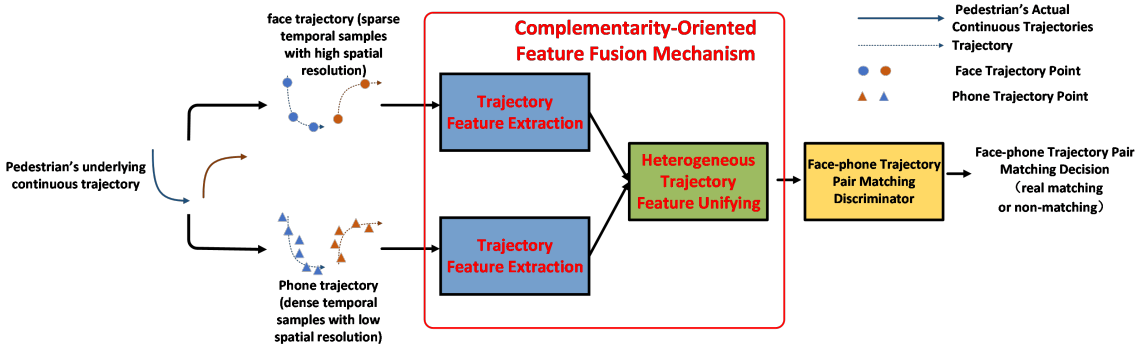


FIGURE 1. **Framework for face-phone trajectory matching based on Complementarity-Oriented Feature Fusion Mechanism.**

jectories. Ref. [20] attempts to capture the semantic similarities in trajectories by directly learning an embedding from predicting the context aspect tokens whose generation relies on clustering. In sharp contrast, our method automatically extracts trajectory semantic with versatility. Moreover, we obtain trajectory embedding with a two-phase encoding framework. Phase-1 is analogous to transforming autoencoders in previous synthesis works, phase-2 further enhances the embedding by unifying the abstract representation for different trajectory modalities. In doing so, our two-phase trajectory embedding addressed the desire to capture complementarity for true matched face-phone pairs.

D. CYCLE-CONSISTENT GENERATIVE NETWORK

Motivated by Cycle-GAN [21] we design a cycle consistent trajectory translation network (CCTTN) to model the complementarity in real matching face-phone trajectory pairs and to enhance the trajectory embedding's distinctiveness orienting towards differentiating true matching face-phone trajectory pairs from the artificial ones.

III. PROPOSED APPROACH

Let $O^A = \{o_1^A, \ldots, o_M^A\}$ denote a set of M face ID trajectories. In O^A, each face trajectory $o_i^A = [o_i^A(1), \ldots, o_i^A(x)]$ consists of x trajectory points for the $i-th$ face ID. The αth-trajectory point where $\alpha \in 1, \ldots, x$ is formally represented as a triplet $o_i^A(\alpha) = [lon, lat, ts]$, where *lon* and *lat* denotes the longitude and latitude of the camera respectively which approximates a pedestrian's geographic position, and ts denotes the timestamp for the pedestrian's appearance on the CCTV camera. Similarly, $O^B = o_1^B, \ldots, o_N^B$ denotes a set of N phone ID trajectories. In O^B each phone trajectory $o_j^B = [o_j^B(1), \ldots, o_j^B(y)]$consists of y trajectory points for the $j-th$ phone ID. The βth-trajectory point where $\beta \in 1, \ldots, y$ is formally represented as a triplet $o_i^B(\beta) = [lon, lat, ts]$, where *lon* and *lat* denotes the longitude and latitude of the communication base station respectively that the phone communicates with at timestamp *ts* , approximating the pedestrian's location. The aim of this work is to perform trajectory pairing of the face trajectory $o_i^A \in O^A$ and the mobile phone trajectory $o_i^B \in O^B$, originated from the same one pedestrian.

The entire network is shown in Fig 1. First, to fully exploit this latent complementarity, we build a two-phase-embedding process called Complementarity-Oriented Feature Fusion Mechanism (COFFM), containing a Trajectory Feature Extraction (TFE) module and a Multi-modality Trajectory Feature Unifying (MTFU) module. Specifically, the TFE module encodes a face/phone trajectory into an initial latent feature vector such that it is amicable to decoding into the other modality. The subsequent MTFU module uses a cycle-consistent Trajectory Translation Network (CCTTN) to obtain a unified abstract trajectory representation. Second, we formulate the trajectory matching problem as a binary classification task, and train a Face-Phone Trajectory Pair Matching Discriminator (FPTPMD) module to accomplish it – it takes the output of MTFU module to infer a decision of whether a face-phone trajectory is a true match.

A. TRAJECTORY FEATURE EXTRACTION (TFE)

Phase one of the COFFM is the Trajectory Feature Extraction (TFE) module that is designed to extract preliminary features to capture modality transfer behaviors. To this end, we propose to use a general auto-encoder-decoder framework with long-short-term memory (LSTM) blocks [22]–[26]. The modality transfer aspect is achieved by having different modalities as the encoder input and decoder output. The LSTM blocks are amicable to variable and inconsistent sequence lengths, alleviating the stringent requirement of existing work for matched lengths in the sequences [26]–[30]. The contrast between MLP and LSTM respectively used for trajectory feature extraction is shown in Fig 2.

When TFE is used to perform face-to-phone trajectory transformation, the encoder takes the face trajectory o_i^A as input and encodes o_i^A into a 64-dimensional feature vector v_i^{A2B}, and the decoder uses this feature vector v_i^{A2B} to reconstruct corresponding phone trajectory o_i^B. The latent feature vector v_i^{A2B} contains directional information of face trajectory o_i^A and the phone trajectory o_i^B. Upon completion of training, the encoder can be applied to any face trajectory to obtain a representation v_i^{A2B} that is phone-compatible. Similarly, a face-compatible phone representation v_j^{B2A} can be obtained when the roles of face and phone sequence are reverted.

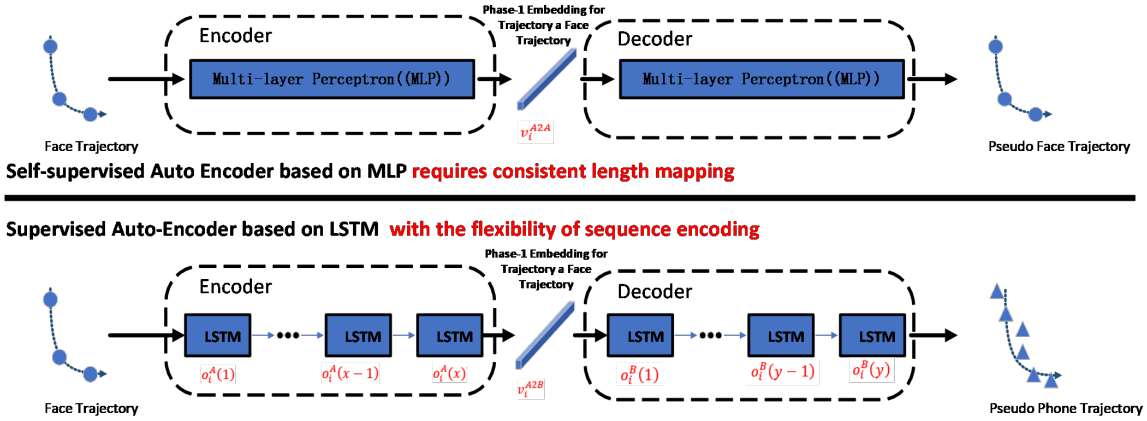


FIGURE 2. Trajectory Feature Extraction (TFE) module.

LSTM network modules are used as basic building blocks in the TFE to capture the semantics of the sequential data.

The v_i^{A2B} or v_j^{B2A} preliminarily capture the complementarity within corresponding face-phone trajectory pairs. In subsequent HTFU, the complementarity will be enhanced. To express the phased process of capturing the complementarity, we would call the v_i^{A2B} or v_j^{B2A} as Phase-1 embedding for a trajectory.

B. MULTI-MODALITY TRAJECTORY FEATURE UNIFYING (MTFU)

The second phase of the COFFM is the MTFU module to obtain abstract mapping embedding. Specifically, it is designed to enhance the complementarity of the vector representations from TFE alone. While the face-to-phone and phone-to-face paths are modality-translating in their own rights, there is an intrinsic directionality that does not quite offer the symmetry we desire for an ideal complementary common embedding.

Whether a 1-phase trajectory embedding is originated from a face trajectory or phone trajectory, HTFU deals with it in the same way. Specifically, HTFU transforms a 1-phase face trajectory embedding to a 2-phase face trajectory embedding and HTFU deals with a 1-phase phone trajectory embedding in a quite similar way. For simplicity, in this section we temporarily do not differentiate the specific type of original trajectory source, but generally call one source as A-type and another source as B-type. For example, if A-type means that the source is face trajectory and then B-type means phone trajectory source, or vice versa.

Motivated by the idea of cycle consistency [21], [31] in generative adversarial networks, we design a Cycle-consistent Trajectory Translation Network (CCTTN) to achieve unified abstract representation, as shown in Fig 3.

The preliminary features $\{v_i^{A2B}\}$ and $\{v_j^{B2A}\}$ from TFE are used as input to a cycle consistent auto-encoder pair $AE_{A2B\to B2A}$ and $AE_{B2A\to A2B}$. The intermediate feature vectors $w^{A2B\to B2A}$ and $w^{B2A\to A2B}$ are concatenated to form a single representation. The training of CCTTN is driven by a loss $L(AE_{A2B\to B2A}, AE_{B2A\to A2B})$ (as shown in (4)) consisting weighted sum of mean squared error losses for directional translators $AE_{A2B\to B2A}$ and $AE_{B2A\to A2B}$, and a cycle consistency criterion $L_{cyc}(G,F)$ (as shown in (3), (4).) The weight λ is set to one in our experiments.

$$L_{AE_{A2E\to B2A}}(\{v^{A2B}, v^{B2A}\}; AE_{A2B\to B2A}) = E_{v^{A2B}}\left[\left\|AE_{A2B\to B2A}\left(v^{A2B}\right) - v^{B2A}\right\|_2\right]. \tag{1}$$

$$L_{AE_{B2A}}(F,B,A) = E_{o_i^{B,1-phase}\sim P_{data}(o_i^{B,1-phase})}\left[\left\|G(o_i^{B,1-phase}) - o_i^{A,1-phase}\right\|_2\right]. \tag{2}$$

$$\begin{aligned} L_{cyc}(G,F) = & \\ & E_{o_i^{A,1-phase}\sim P_{data}(o_i^{A,1-phase})}\left[\left\|F(G(o_i^{A,1-phase})) - o_i^{A,1-phase}\right\|_2\right] \\ & + E_{o_i^{B,1-phase}\sim P_{data}(o_i^{B,1-phase})}\left[\left\|G(F(o_i^{B,1-phase})) - o_i^{B,1-phase}\right\|_2\right]. \end{aligned} \tag{3}$$

$$\begin{aligned} & L(AE_{A2B\to B2A}, AE_{B2A\to A2B}) = \\ & L_{AE_{A2B}} + L_{AE_{B2A}} + \lambda L_{cyc}(AE_{A2B\to B2A}, AE_{B2A\to A2B}). \end{aligned} \tag{4}$$

Concatenating the intermediate vectors w_i^{A2B} and w_i^{B2A} provides symmetric two-way translation, preserving both the inherent similarity and the directional translation difference.

C. FACE-PHONE TRAJECTORY PAIR MATCHING DISCRIMINATOR

The face-phone trajectory matching problem can be formulated as a binary classification problem to indicate a true match. We call this classifier as Face-Phone Trajectory Pair Matching Determinator, abbreviated as FPTPMD.

For a face-phone trajectory pair, COFFM respectively generates an abstract representation vector for the face trajectory and the phone trajectory. We concatenate these two vectors into one vector as the abstract representation of the face-phone trajectory pair (abbreviated as AFPPV for simplicity). The classification network generates an output of a two-dimensional category vector. The two elements in this category vector respectively indicate the probability of matching

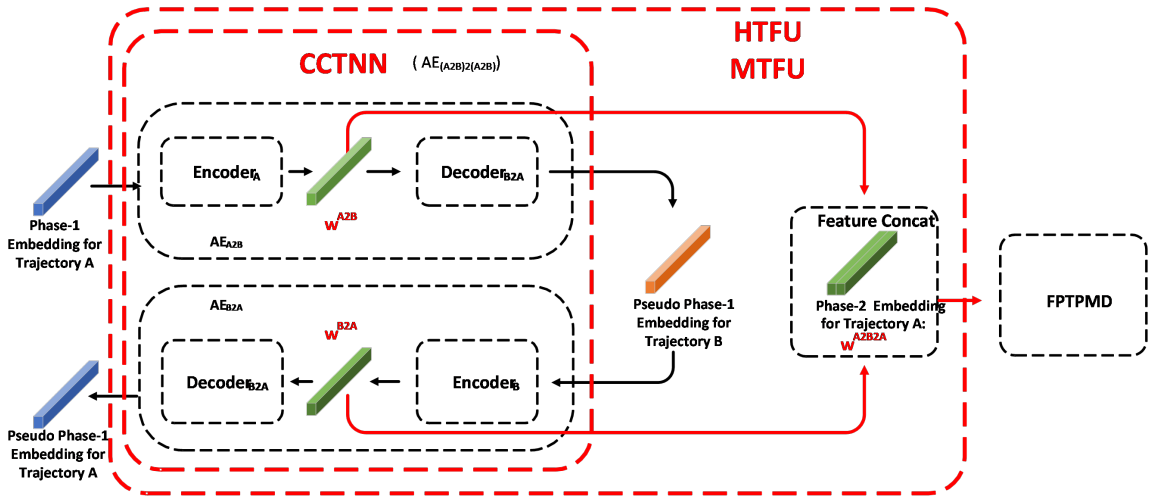


FIGURE 3. **Architecture of MTFU module based on cycle consistent trajectory translation network (CCTTN).**

between the face and phone trajectory at input. The higher possibility is taken to predict the matching decision as true or false, which respectively represents that the original face-phone pair is real matching or non-matching.

Since the raw trajectory dataset contains only true matching face-phone trajectory pairs, it is necessary to generate negative samples to prevent the discriminator from overfitting the matching pairs. Since one pedestrian's face trajectory does not match another pedestrian's phone trajectory, we randomly select face trajectories and phone trajectories to generate non-matching face-phone trajectory pair to generate AFPPV as negative samples.

IV. EXPERIMENTS AND ANALYSIS

A. RAW DATASET DESCRIPTION AND PROPERTIES

Face and phone trajectory datasets were used in our experiments. The face trajectory dataset D_{ft} consisted of 43423 face trajectories. A specific example of a face trajectory point is a triplet <1598316054, $1 * 4.72184214004113$, $2 * .344966375887772$> where the '$*$' hides the true values for legal considerations. All face trajectories have been transformed to CCTV coordinates face recognition performed by the data provider, ridding all privacy information. The mobile-phone trajectory dataset D_{pt} is consisted of 43423 phone trajectories and each of these phone trajectories is composed of a series of trajectory points in the form of <time, longitude, latitude>. A specific example of a face trajectory point is a triplet like <1598314054, $1*4.72184214004113$, $2*.344966375887772$>. The third dataset D_{fp_match} appends an indicator at face-phone pairs to indicate true match.

A trajectory's length is defined as the number of trajectory points in the trajectory. I presents the statistics of the lengths of D_{ft} and D_{pt}. It can be observed that the face trajectories are typically very sparse, with a median length of 3. This indicates that only partial information about pedestrian movement can be derived from face trajectory.

Significant length difference can also be observed between the face and phone trajectories. Let $M_{inconsistency}$ be the ratio of a phone trajectory length to a face trajectory length and describe the severity of length inconsistency. II and Fig 4 show the distribution of $M_{inconsistency}$ of the 43423 matching face-phone pairs in D_{fp_match}, with the average and median being 24 and 8.6 respectively, illustrating significant length inconsistency between the two modalities, making classic similarity-based methods inapplicable.

B. EXPERIMENTAL SETUP

The CCTTN consists of two pairs of cascaded encoder-decoder, and each encoder/decoder consists of a 64-dimension input layer and two 64-dimension hidden layers. FPTPMD is a 2-layer Multi-Layer Perceptron (MLP) with 128-dimensional layers, Relu as activation functions between hidden layers, and Softmax following the output layer. We train CCTTN and FPTPMD separately, each with RMSprop [32] for 40 epochs at the learning rate of 0.1. For CCTTN, batch size is set as 10 and dropout rate as 0.5. For FPTPMD, the batch size is set as 100.

For training our frame in Fig 1, we reserve 20% of face-phone trajectory pairs indicated in D_{fp_match} as testing sets (D_{test} in III). The rest of the real matching pairs are used as positive pair samples, together with randomly constructed non-matching trajectory pairs as negative pair samples, for training. Only the positive pairs are used in training CCTTN.

C. EVALUATION

Since the face-phone trajectory matching is solved by using binary classifier, we use four metrics: Accuracy (A), Precision (P), Recall (R), and F1-score (F1) to comprehensively evaluate the performance of different methods. A, P, R, F1 are respectively defined in (5). In these definitions, TP is the number of true positive face-phone matching pairs; FP is the number of false positive pairs, TN is the number of true negative pairs, and FN is the number of false negative pairs.

TABLE I. Statistical property of the number of trajectory points.

Trajectory Dataset	Number of Trajectories	Average	Minimum	Quarter	Median	Third-Quarters	Maximum
Face Trajectory	43423	8	1	1	3	7	7744
Phone Trajectory	43423	92	1	8	37	123	5640

TABLE II. Statistics of $M_{inconsistency}$ of real matching face-phone trajectory pairs.

	Average	Minimum	Quarter	Median	Third-Quarters	Maximum
$M_{inconsistency}$	24	0.01	3.3	8.6	23	3110

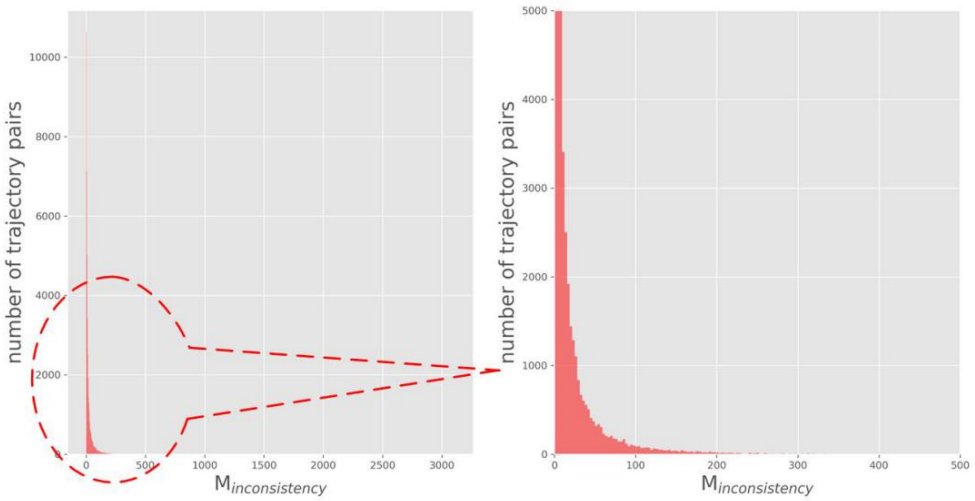


FIGURE 4. Distribution of $M_{inconsistency}$ of matching face-phone pairs. The left histogram shows the distribution of the $M_{inconsistency}$ of all matching face-phone trajectory pairs, the right histogram specially shows the $M_{inconsistency}$ of matching face-phone trajectory pairs with $M_{inconsistency} \leq 500$.

$$
\begin{aligned}
Accuracy &= \frac{TP+TN}{TP+FP+TN+FN}, \\
Precision &= \frac{TP}{TP+FP}, \\
Recall &= \frac{TP}{TP+FN}, \\
F1 &= 2 * \frac{Precision * Recall}{Precision + Recall}.
\end{aligned} \tag{5}
$$

Our complementarity-based method is compared with traditional similarity-based trajectory matching methods. Two earlier methods, Dynamic Time Warping (DTW) [33] and Discrete Frechet (DF) [34], [35], may also be used to calculate similarity between a pair of heterogeneous trajectories such as face-phone trajectory pair. DTW captures flexible similarities under time distortions, and DF measures the similarity of two polygonal curves – both compatible with sequences of different length and can be used as baselines.

In the benchmark similarity-based methods, mismatch is expressed with a real number, in contrast to our binary decision. To fairly and meaningfully compare the effectiveness of our method with DTW/DF concerning the face-phone trajectory matching task, since the testing sets are with ratio of 1:1 (number of positive pairs to number of negative pairs), we use the median of the trajectory similarities computed with DTW/DF based on the training set as the threshold for determining a match.

To comprehensively compare various methods as for face-phone matching task, we set aside 3 testing sets for evaluation. The 3 testing sets are: (1) D_{test} to evaluate a method's general performance. (2) D_{sparse} consists of face trajectory and phone trajectory with very short face trajectories to evaluate a method's robustness towards sparse face trajectory. (3)$D_{inconsistency}$ with severe length inconsistency to evaluate a method's performance on inconstant face-phone trajectory pairs. The description of the 3 testing datasets is summarized in III.

We applied the benchmark DTW, DF, and the proposed COFFM on the above 3 testing datasets. IV reports the performance, among which accuracy is the evaluation metric most concerned by our project-launching part as for the real application. It can be observed that our approach generally outperforms the baseline on all testing datasets.

Our approach achieves an Accuracy of 97.1% on D_{test}, exceeding DTW by 48.1% and DF by 38.6%. We mainly attribute the advantage of our method over DTW/DF to that our

TABLE III. Building of testing datasets.

Testing Dataset	Number of Trajectory Pairs	Testing Dataset Building Description
D_{test}	8685×2	Half of the face-phone trajectory pairs in D_{test} is sampled in this way: we randomly select real matching trajectory pairs' faceIDs and phoneIDs from D_{fp_match}, and take corresponding trajectory form D_{ft} and D_{pt} to form real matching trajectory pairs, i.e., positive pairs. Another half of face-phone trajectory pairs in D_{test} is obtained by randomly selecting face trajectory in D_{pt} and phone trajectory in D_{ft} excluding the pairs which are labeled as real matching in D_{fp_match}.
D_{sparse}	4894×2	The construction of D_{sparse} is the same with that of D_{test}, but the trajectory pairs in Dsparse are those with very sparse face trajectory (face trajectory $length \leq 3$).
$D_{inconsistency}$	4010×2	The construction of $D_{inconsistency}$ is the same with that of D_{test}, but the trajectory pairs in $D_{inconsistency}$ are those with very high inconsistency ($M_{inconsistency} \geq 10$).

complementarity-oriented method obtains discriminative feature representation of face-phone trajectory pair, and avoids confusion between positive pairs and negative pairs, while DTW and DF ignore the complementarity between positive face-phone trajectory pairs.

Our approach achieves an Accuracy of 94.8% on D_{sparse}, demonstrating our model's robustness against sample sparseness. This is partly because the dense (long) phone trajectory compensates the lack of useful information in the sparse face trajectories. In addition, we can observe that DTW shows a 15.3% Accuracy improvement on D_{sparse} than on D_{test}, which may be attributed to the point-reuse of sparse face trajectory by DTW.

Finally, our approach achieves an Accuracy of 95.6% on $D_{inconsistency}$, compared to 18.4% from DTW and 51.6% from DF, demonstrating our method's ability to capture the complementarity.

V. DISCUSSIONS AND CONCLUSION

Pedestrians typically move with high randomness in various direction at relatively low speeds, resulting in trajectories with relatively poor regularity and sparsity in face trajectory, compared to fast moving vehicles conforming to roads with relative regular and dense trajectories. The phase-phone matching problem presents significant challenges in coping with sample sparsity in face trajectories and low spatial resolution in phone trajectories.

The proposed COFFM has a novel CCTTN as its central component, capturing the implicit and hard-to-describe complementarity. The network structure and phased encoding alleviates the need to set ad-hoc parameters in our method, compared to other existing similarity-based and unisource modality oriented method. The phased scheme is simple and offers interpretability. One may consider a joint training approach for CCTTN and FPTPMD with more complex tasks.

There may be a need to emphasize recall compared to precision. Since false positivity may be corrected with a secondary module or even benefit for certain tasks such as companion or traffic behaviors among pedestrians [7], [12]. The TPTPMD classifier can be easily adjusted to reflect such preference.

Our study shows that two-way translation fusion mechanism is beneficial in solving trajectory matching problems, especially under challenging sparseness and length inconsistency conditions. While our work a framework for multi-modality matching, it is expected that application specific characteristics would necessitate network retraining depending on the setting.

REFERENCES

[1] N. Magdy, M. A. Sakr, T. Mostafa, and K. El-Bahnasy, "Review on trajectory similarity measures," in *2015 IEEE seventh international conference on Intelligent Computing and Information Systems (ICICIS)*. IEEE, 2015, pp. 613–619.

[2] C. Faloutsos, M. Ranganathan, and Y. Manolopoulos, "Fast subsequence matching in time-series databases," *ACM Sigmod Record*, vol. 23, no. 2, pp. 419–429, 1994.

[3] D. Papadias, J. Zhang, N. Mamoulis, and Y. Tao, "Query processing in spatial network databases," in *Proceedings 2003 VLDB conference*. Elsevier, 2003, pp. 802–813.

[4] B. Lin and J. Su, "Shapes based trajectory queries for moving objects," in *Proceedings of the 13th annual ACM international workshop on Geographic information systems*, 2005, pp. 21–30.

[5] H. Alt, "The computational geometry of comparing shapes," *Efficient Algorithms: Essays Dedicated to Kurt Mehlhorn on the Occasion of His 60th Birthday*, pp. 235–248, 2009.

[6] T. Nakamura, K. Taki, H. Nomiya, K. Seki, and K. Uehara, "A shape-based similarity measure for time series data with ensemble learning," *Pattern Analysis and Applications*, vol. 16, pp. 535–548, 2013.

[7] I. Assent, M. Wichterich, R. Krieger, H. Kremer, and T. Seidl, "Anticipatory dtw for efficient similarity search in time series databases," *Proceedings of the VLDB Endowment*, vol. 2, no. 1, pp. 826–837, 2009.

[8] E. Keogh and C. A. Ratanamahatana, "Exact indexing of dynamic time warping," *Knowledge and information systems*, vol. 7, pp. 358–386, 2005.

[9] D. Pudasaini and A. Abhari, "Scalable pattern recognition and real time tracking of moving objects," in *2019 Spring Simulation Conference (SpringSim)*. IEEE, 2019, pp. 1–11.

[10] L. Chen, M. T. Özsu, and V. Oria, "Robust and fast similarity search for moving object trajectories," in *Proceedings of the 2005 ACM SIGMOD international conference on Management of data*, 2005, pp. 491–502.

[11] J. W. Hunt and T. G. Szymanski, "A fast algorithm for computing longest common subsequences," *Communications of the ACM*, vol. 20, no. 5, pp. 350–353, 1977.

[12] Y. Zheng and X. Xie, "Learning travel recommendations from user-generated gps traces," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 1, pp. 1–29, 2011.

[13] Z. Dong, F. Tian, H. Yang, T. Sun, W. Zhang, and D. Ruan, "A framework with elaborate feature engineering for matching face trajectory and mobile phone trajectory," *Electronics*, vol. 12, no. 6, p. 1372, 2023.

[14] H. Liu, A. Alali, M. Ibrahim, B. B. Cao, N. Meegan, H. Li, M. Gruteser, S. Jain, K. Dana, A. Ashok, B. Cheng, and H. Lu, "Vi-fi: Associating moving subjects across vision and wireless sensors," in *2022 21st ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*, 2022, pp. 208–219.

TABLE IV. Evaluation of face-phone trajectories matching on testing datasets.

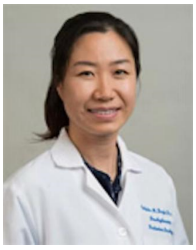
Testing Datasets	DTW				DF				Ours			
	A	F1	P	R	A	F1	P	R	A	F1	P	R
D_{test}	49.0%	49.0%	52.1%	46.3%	58.5%	58.5%	57.9%	59.1%	97.1%	97.1%	97.6%	96.3%
D_{sparse}	64.3%	64.3%	71.6%	57.8%	49.4%	49.4%	46.9%	52.2%	94.8%	94.8%	95.2%	94.4%
$D_{inconsistency}$	18.4%	18.4%	19.1%	17.7%	51.6%	51.6%	49.7%	53.6%	95.6%	95.6%	96.3%	94.9%

[15] W. Wan and M. Cai, "Phone-vehicle trajectory matching framework based on alpr and cellular signalling data," *IET Intelligent Transport Systems*, vol. 15, no. 1, pp. 107–118, 2021.
[16] D. Xiao, L. Song, R. Wang, X. Han, Y. Cai, and C. Shi, "Embedding geographic information for anomalous trajectory detection," *World Wide Web*, vol. 23, no. 5, pp. 2789–2809, 2020.
[17] X. Li, K. Zhao, G. Cong, C. S. Jensen, and W. Wei, "Deep representation learning for trajectory similarity computation," in *2018 IEEE 34th international conference on data engineering (ICDE)*. IEEE, 2018, pp. 617–628.
[18] W. Yang, Y. Zhao, B. Zheng, G. Liu, and K. Zheng, "Modeling travel behavior similarity with trajectory embedding," in *Database Systems for Advanced Applications: 23rd International Conference, DASFAA 2018, Gold Coast, QLD, Australia, May 21-24, 2018, Proceedings, Part I 23*. Springer, 2018, pp. 630–646.
[19] N. Zhou, W. X. Zhao, X. Zhang, J.-R. Wen, and S. Wang, "A general multi-context embedding model for mining human trajectory data," *IEEE transactions on knowledge and data engineering*, vol. 28, no. 8, pp. 1945–1958, 2016.
[20] T. Boonchoo, X. Ao, and Q. He, "Multi-aspect embedding for attribute-aware trajectories," *Symmetry*, vol. 11, no. 9, p. 1149, 2019.
[21] D. Xia, H. Liu, L. Xu, and L. Wang, "Visible-infrared person re-identification with data augmentation via cycle-consistent adversarial network," *Neurocomputing*, vol. 443, pp. 35–46, 2021.
[22] H. Bourlard and Y. Kamp, "Auto-association by multilayer perceptrons and singular value decomposition," *Biological cybernetics*, vol. 59, no. 4, pp. 291–294, 1988.
[23] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *science*, vol. 313, no. 5786, pp. 504–507, 2006.
[24] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proceedings of the 25th international conference on Machine learning*, 2008, pp. 1096–1103.
[25] S. Rifai, P. Vincent, X. Muller, X. Glorot, and Y. Bengio, "Contractive auto-encoders: Explicit invariance during feature extraction," in *Proceedings of the 28th international conference on international conference on machine learning*, 2011, pp. 833–840.
[26] A. Graves and A. Graves, "Long short-term memory," *Supervised sequence labelling with recurrent neural networks*, pp. 37–45, 2012.
[27] S. Hochreiter, "Untersuchungen zu dynamischen neuronalen netzen," *Diploma, Technische Universität München*, vol. 91, no. 1, p. 31, 1991.
[28] K. Kawakami, "Supervised sequence labelling with recurrent neural networks," Ph.D. dissertation, Ph. D. thesis, 2008.
[29] F. Gers, "Long short-term memory in recurrent neural networks," Ph.D. dissertation, Verlag nicht ermittelbar, 2001.
[30] K. Cho, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.
[31] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.
[32] T. Tieleman, "Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude," *COURSERA: Neural networks for machine learning*, vol. 4, no. 2, p. 26, 2012.
[33] Z. Zhang, R. Tavenard, A. Bailly, X. Tang, P. Tang, and T. Corpetti, "Dynamic time warping under limited warping path length," *Information Sciences*, vol. 393, pp. 91–107, 2017.
[34] P. K. Agarwal, R. B. Avraham, H. Kaplan, and M. Sharir, "Computing the discrete fréchet distance in subquadratic time," *SIAM Journal on Computing*, vol. 43, no. 2, pp. 429–449, 2014.
[35] K. Buchin, M. Buchin, W. Meulemans, and W. Mulzer, "Four soviets walk the dog: Improved bounds for computing the fréchet distance," *Discrete & Computational Geometry*, vol. 58, pp. 180–216, 2017.

CHANGFENG CAO received the B.E. degree from Henan University of Economics and Law. Now, he is a graduate student in Guizhou Normal University. His research interests include data mining, machine learning, natural language processing and application.

WENCHUAN ZHANG received the B.E. degree from Wuhan Institute of Technology and the M.E. degree from Guizhou Normal University. His research interests include data mining, machine learning, natural language processing and application.

HUA YANG received his Ph.D degree in computer software theory in 2009 from Wuhan University, Wuhan, China. He used to work in Wuhan University until 2015 and he is currently a professor with GuiZhou Normal University. He has been an academic visitor in the University of Nottingham, UK. from 2015 to 2016, His research interests include data mining, machine learning, and natural language processing.



DAN RUAN recieved Ph.D in Electrical Engineering and M.S. in Mathematics, both from the Unviersity of Michigan, Ann Arbor in 2008. She is currently a professor of Bioengineering, and Radiation Oncology, at University of Califorina, Los Aneles. Her research interest includes signal and image processing, optimization, and statistical learning.