Only Trajectory is Needed: Recognizing Emotion Behind Social Behavior

Anonymous submission

A. Pseudo-code for Related Algorithms

Stops Extraction via GPS Trajectories

The key to transforming GPS trajectories into event sequences lies in extracting stops, setting POIs for unattributed stops, and selecting the most appropriate POIs for enriched stops. We show pseudo-code for these three critical algorithms in particular.

Algorithm 1: Extract Stops

Input: T, T_{th}, D_{th}
Output: S

```
S ← ∅, stop ← ∅
for coordinate in T do
    while duration(stop) < D_th do
        stop.append(coordinate)
    end
    for coord in stop do
        if distance(coord,stop) > D_th then
            stop.pop_first()
            goto line2
        end
    end
    while distance(coordinate,stop) ≤ D_th do
        stop.append(coordinate)
    end
    S ← S∪ stop, stop ← ∅
end
```

Algorithm 1 employs specific strategies to reduce its time complexity. For instance, we reduce the time complexity from $O(n^2)$ to $O(n)$ by setting the $stop$ variable in row one to a queue structure. When determining whether the coordinates to be joined are reasonable, we follow a specific process. We first test if they match the last coordinate of the $stop$, which is a more efficient approach than recalculating whether the $stop$ satisfies Definition 2 each time.

POI Creation for UStops

The function of Algorithm 2 is to aggregate unattributed stops and then artificially set the POIs of these clusters. In extreme cases, each unknown stop is set to a new POI.

Algorithm 2: Creat POIs

Input: UStops
Output: POISet

```
POISet ← ∅, N ← 1
while N has been increased do
    centroid, clusters ← K-Means(N, UStops)
    for center, cluster in centroid, clusters do
        if distance(center, cluster) > D_th then
            N ← N + 1
            break
        end
    end
end
for center, cluster in centroid, clusters do
    POISet ← POISet ∪ POI(center, best type)
end
```

POI Selection for EStops

We implement Strategy 2 and Strategy 3 with two dictionaries, respectively. Strategy 2's dictionary keyword is the period; the value is the type of the recommended POI. Strategy 3's dictionary keyword is the previous POI; the value is the list. Specifically, iterating through all the event sequences, we take out the POIs of two neighboring events at a time, set the previous POI as the keyword of the dictionary, and the latter POI is inserted into the corresponding list, and the POIs in the list are sorted in descending order of counts.

B. Implementation Details of *Eventlet*

When transforming the raw data into input sequences, we optimized the input sequences, considering the differences and characteristics between the dimensions.

- The integer portions of longitude and latitude usually do not change much and need to be removed in order for the model to capture sequence changes better;
- If the dimension of the input sequence contains a start time and an end time, the start and end times need to be converted to their relative times during the sampling period (morning, afternoon, or evening);
- Keeping the values of all dimensions between 0 and 1 is necessary to close the absolute gap between dimensions.

Algorithm 3: Find Best POI

Input: EStop
Output: POI

$\rho \leftarrow$ EStop.get_candidate_poi()
if *Satisfies Strategy 1* **then**
  $\sigma \leftarrow \rho(most\ type)$
  POI $\leftarrow \arg\min_{poi \in \sigma}$ distance(poi, stop)
else if *Satisfies Strategy 2* **then**
  $\tau \leftarrow \rho(special\ type)$
  POI $\leftarrow \arg\min_{poi \in \tau}$ distance(poi, stop)
else if *Satisfies Strategy 3* **then**
  $\varphi \leftarrow \rho(previous\ type)$
  POI $\leftarrow \arg\min_{poi \in \varphi}$ distance(poi, stop)
else
  POI $\leftarrow \arg\min_{poi \in \rho}$ distance(poi, stop)
end

C. Dataset

Basic Information

The two datasets recorded GPS trajectories for 30 student volunteers for 10 to 30 consecutive days. The emotional patterns captured in these trajectories, influenced by the relative freedom of student life and the pressures of graduation or job searching, are of great value. Other differences between the two datasets are described below.

- Restricted movement. The XDU-R1510 was collected in December 2022 for 10-12 days, when the university was under control, so the student's range of events was near a university.
- Freedom of movement. The XDU-F1530 was collected in March-April 2023 for 29-30 days when the university was decontrolled, so the student movement was utterly free.

Collection Method

During the collection period, the APP recorded the volunteers' coordinates every 5 to 15 seconds, with each coordinate containing three features: longitude, latitude, and timestamp. Volunteers were asked to report their emotion index (negative, neutral, or positive) in the morning (6:00-12:00), afternoon (12:00-18:00), and evening (18:00-24:00) of each day so the GPS trajectories for each period were labeled with a corresponding emotion.

The period of the sample and the design of the emotion labels are related. First, the period is set considering that:

- Human emotions usually go through multiple changes within a day, so the period cannot be too long;
- The event sequence should contain an appropriate number of events so the period cannot be too short;
- 00:00-6:00 is usually a resting time, and emotions during unconsciousness are out of the scope of our study.

Second, the emotion labels were set considering that:

- People may have multiple emotions at the same time at the exact moment, so specific emotions are difficult to describe precisely;
- The duration of specific emotions may be short;
- Abstract emotion labels can better summarize the emotions over a more extended period.

Quality Assessment

To assess the quality of the datasets, we compared their critical metrics to those of some widely used emotion recognition datasets, and the results are shown in Table 1.

Table 1: Comparison of critical metrics between the collected dataset and other emotion recognition datasets

Dataset	Users	Samples	Labels
MAHNOB-HCI(Soleymani et al. 2011)	30	532	10
DEAP(Koelstra et al. 2011)	32	1,280	5
eNTERFACE'05(Martin et al. 2006)	42	1,166	5
SEMAINE(McKeown et al. 2011)	150	959	5
XDU-R1510 (Ours)	30	490	3
XDU-F1530 (Ours)	30	1,286	3

According to Table 1, XDU-R1510 and XDU-F1530 are comparable to the other emotion recognition datasets in terms of critical metrics, such as the number of participants and samples, so the collected datasets satisfy the quality requirements of emotion recognition datasets. In addition, according to the distribution of labels (Figure 1), the dataset does not show a disparity in the proportion of labels.

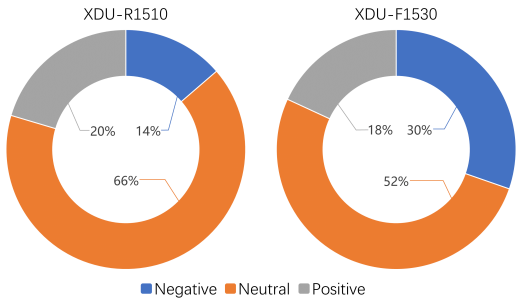

Figure 1: The distribution of emotion indices for both datasets

D. Experimental Setup

PyTorch implements the proposed method. The experimental runtime environment is Ubuntu 20.04. The CPU is an AMD EPYC 7543 32-core Processor, and the GPU is NVIDIA A100-PCIE-40GB. For the proposed method, the granularity size of the Eventlet is 10 minutes, the time threshold of the event T_{th} is 15 minutes, and the distance threshold D_{th} is 200 meters. In addition, all hyperparameters of the validation method are the settings recommended in the related literature.

E. Selection of *Eventlet* Granularity

According to the role of *Eventlet*, the candidate values of *Eventlet* granularity that satisfy the conditions are $5min$, $10min$, and $15min$, and the corresponding *Eventlet* sequence lengths are 72, 36, and 24, respectively. The experimental results corresponding to different *Eventlet* granularities are shown in Table 2.

Table 2: Accuracy comparison with *Eventlet* granularity as independent variable

G(min) / Method	XDU-R1510			XDU-F1530		
	5	10	15	5	10	15
KNN	**66.40**	65.60	62.40	54.13	**57.75**	56.68
K-Means	**41.60**	39.20	34.40	37.60	**39.84**	35.56
MLP	61.60	**68.80**	63.20	49.60	**54.55**	50.53
RNN	58.40	**64.00**	53.60	50.53	53.48	**54.01**
GRU	62.40	**71.20**	66.40	55.20	**61.23**	56.68
LSTM	64.00	**68.00**	62.40	54.13	61.23	**62.30**
Transformer	64.80	**68.00**	63.20	49.87	**57.49**	52.67
K-Shape	**45.60**	24.80	21.60	33.07	33.96	**37.70**
Tapnet	68.80	**71.20**	66.40	55.35	**59.36**	57.49
Time2Graph	65.60	**68.80**	64.00	54.81	54.81	54.01
DA-Net	56.80	**68.00**	67.20	49.47	**54.81**	49.20
MultiRocket	67.20	**75.20**	64.80	56.68	**58.29**	57.75
Best	3	9	0	0	8	3

The experimental results show that the optimal *Eventlet* granularity is $10min$. Why is there no positive or negative correlation between model performance and *Eventlet* granularity? Because we use "rounding" to set up *Eventlets* that are not fully covered by events, coarse-grained *Eventlets* may cause more errors. At the same time, fine-grained *Eventlets* may classify events too accurately (However, human behavior is biased; e.g., the same person cannot eat at a fixed period every day), thus affecting the model's generalization ability.

F. Selection of T_{th} and D_{th}

In the proposed method, D_{th} is used as a distance threshold for stop extraction and POI selection to compensate for the error of the GPS acquisition device. To ensure the robustness of our methodology, we thoroughly tested the maximum error of the GPS acquisition device, which was found to be 200 meters. Therefore, D_{th} should be set to 200.

To analyze the effect of T_{th} on classification performance, we set T_{th} to $10min$, $15min$, $20min$, $25min$, and $30min$ and conducted the experiments, respectively. T_{th} serves as a time threshold for stop extraction, which filters the events with a shorter duration. It means that T_{th} negatively correlates with the number of events found. The corresponding experimental results are shown in Table 3.

The experimental results show that the value of T_{th} is not significantly correlated with model accuracy. Although a smaller T_{th} can dig out more events, these small events may not be the actual events of the object (e.g., waiting for a train at the station), and they may have little effect on the emotion. In comparison, a more extensive T_{th} filters small events, but some behavioral information of the object will be lost. Therefore, the best T_{th} must be found according to the situation. For the dataset used in this experiment, we take 15 minutes as the default value of T_{th}.

G. Information on the four sequences

The four sequences used for the ablation experiments and the dimensions they contain are as follows:

- The original stop sequence: user ID, latitude, longitude, start time, end time, segment ID.
- The standard stop sequence: user ID, latitude, longitude, segment ID.
- The original event sequence: user ID, event type, latitude, longitude, start time, end time, segment ID.
- The standard event sequence: i.e., *Eventlet* sequence.

H. Limitations

The key to the proposed method is to establish a mapping between event sequences and emotions. However, a prerequisite for the method to work well is that the subject has a certain degree of behavioral freedom. Unfortunately, for some special groups, their events are restricted most of the time due to objective factors such as occupation and physical functioning (e.g., military personnel, disabled people). Nonetheless, the proposed method still shows good prospects for application due to its all-weather, non-sensory characteristics.

I. Related Work

Our research mainly concerns emotion recognition, event mining, and Time Series Classification (TSC), so we investigated related work in these three areas.

Related Work on Emotion Recognition

Researchers have made significant progress in the problem of emotion recognition in recent years. For example, (Lee et al. 2019) proposed a network for the emotion recognition of facial expressions that considers the context information in the scene; using the characteristics of contextual contexts in dialogs, (Shen et al. 2021) proposed a network that can store longer historical contexts and has dialog-aware capabilities; (Tao et al. 2020) proposed an attention-based convolutional recurrent neural network that can extract more discriminative features from EEG signals; (Zhang, Wang, and Yang 2023) proposed a cross-modal spatio-temporal cancellation network that locates contextual and audio-related information in a weakly supervised manner. (Yang et al. 2024) proposed a framework based on generalized causal graphs to mitigate contextual bias interference and enhance model robustness. Researchers have proposed multimodal models to deal with different input types (Wang, Li, and Cui 2024), e.g., a fusion of facial expression and speech (Tellamekala et al. 2023), a fusion of image and text (Zhu et al. 2022), and a fusion of image, speech, and text (Ren et al. 2023; Xu et al. 2024). There has also been good progress in further work exploring the generation of faces with emotions (Xu et al. 2023) and sarcasm context detection (Wen, Jia, and Yang 2023).

Table 3: Accuracy comparison of *Eventlet* sequences with T_{th} as independent variable

$T_{th}(min)$ / Method	XDU-R1510					XDU-F1530				
	10	15	20	25	30	10	15	20	25	30
KNN	64.80	**65.60**	60.66	64.75	65.57	54.67	57.75	**58.71**	57.91	56.57
K-Means	44.00	39.20	32.79	32.79	**44.26**	30.67	39.84	**43.97**	37.27	43.43
MLP	64.80	**68.80**	62.30	64.75	56.56	45.07	54.55	57.37	**58.18**	56.03
RNN	60.80	58.40	**65.04**	55.28	55.28	**56.80**	45.99	46.92	48.53	55.23
GRU	62.40	**71.20**	68.03	69.67	62.30	55.73	**61.23**	60.32	60.86	58.98
LSTM	57.60	68.00	61.48	**69.67**	65.57	53.87	61.23	58.18	**61.66**	60.59
Transformer	64.80	**68.00**	67.21	61.48	60.66	54.67	**57.49**	53.62	54.96	53.62
K-Shape	25.60	24.80	22.13	**40.98**	33.61	31.47	**33.96**	24.66	22.25	27.08
Tapnet	69.60	71.20	68.85	67.21	**71.31**	58.40	**59.36**	58.71	56.03	58.98
Time2Graph	65.60	**68.80**	68.03	68.03	63.11	54.40	54.81	55.23	54.16	55.23
DA-Net	63.20	68.00	**68.03**	59.02	65.57	46.13	54.81	**54.96**	49.60	52.28
MultiRocket	68.80	**75.20**	71.31	68.03	73.77	58.93	58.29	59.52	57.37	**59.79**
Best	0	6	2	2	2	1	4	3	2	1

Related Work on Event Mining

One of the focuses of our work lies in finding POIs near stops. Regarding discovering POIs, (Zhou et al. 2007) extend existing clustering algorithms to discover locations and define a set of basic evaluation metrics and an interactive evaluation framework. (Zheng et al. 2009) mine POIs and classify travel sequences within a specific geospatial region based on GPS trajectories of multiple users. Regarding finding POIs, specifying one or more locations to be visited, and describing the events at each location, (Wang et al. 2017) provide an algorithm recommending POI sequences that fulfill the requirements. (Cao et al. 2019) proposed mining human behavioral semantics from GPS trajectories, extracting behavioral semantics' feature representations, and finally classifying human life patterns through K-Means.

Related Work on Time Series Classification

In order to establish a mapping of event sequences to emotion indices, we investigated a large number of TSC methods. First, for the problem of univariate TSC, (Lin et al. 2003) proposed a symbolic representation of time series, which can reduce the data's dimensionality and improve the classification algorithms' performance. (Ye and Keogh 2009) proposed the well-known Shapelet, a method to classify time series based on their local features. Subsequently, (Li et al. 2021) improved this work by applying Shapelet to the Multivariate Time Series Classification (MTSC) problem. Among the state-of-the-art methods we mentioned, (Cheng et al. 2020, 2021) extracted time-aware Shapelets via a two-level timing factor, defined and constructed an evolutionary graph of the Shapelet; (Zhang et al. 2020) proposed a novel MTSC model with an attention prototype network, which combines multilayer convolutional networks to learn low-dimensional features from multivariate time series data; (Chen et al. 2022) came up with a novel dual-attention based network for mining local-global features of multivariate time series, and (Tan et al. 2022) proposed a structurally simple TSC model that requires only a very short training time to achieve state-of-the-art accuracy. In addition, (Liu et al. 2024) and (Wang et al. 2024) proposed diffusion model-based and graph contrastive learning-based networks to alleviate the problem of missing labels in datasets, respectively.

References

Cao, H.; Xu, F.; Sankaranarayanan, J.; Li, Y.; and Samet, H. 2019. Habit2vec: Trajectory semantic embedding for living pattern recognition in population. *IEEE Transactions on Mobile Computing*, 19(5): 1096–1108.

Chen, R.; Yan, X.; Wang, S.; and Xiao, G. 2022. DA-Net: Dual-attention network for multivariate time series classification. *Information Sciences*, 610: 472–487.

Cheng, Z.; Yang, Y.; Jiang, S.; Hu, W.; Ying, Z.; Chai, Z.; and Wang, C. 2021. Time2Graph+: Bridging time series and graph representation learning via multiple attentions. *IEEE Transactions on Knowledge and Data Engineering*.

Cheng, Z.; Yang, Y.; Wang, W.; Hu, W.; Zhuang, Y.; and Song, G. 2020. Time2graph: Revisiting time series modeling with dynamic shapelets. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 3617–3624.

Koelstra, S.; Muhl, C.; Soleymani, M.; Lee, J.-S.; Yazdani, A.; Ebrahimi, T.; Pun, T.; Nijholt, A.; and Patras, I. 2011. Deap: A database for emotion analysis; using physiological signals. *IEEE transactions on affective computing*, 3(1): 18–31.

Lee, J.; Kim, S.; Kim, S.; Park, J.; and Sohn, K. 2019. Context-aware emotion recognition networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10143–10152.

Li, G.; Choi, B.; Xu, J.; Bhowmick, S. S.; Chun, K.-P.; and Wong, G. L.-H. 2021. Shapenet: A shapelet-neural network approach for multivariate time series classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 8375–8383.

Lin, J.; Keogh, E.; Lonardi, S.; and Chiu, B. 2003. A symbolic representation of time series, with implications for streaming algorithms. In *Proceedings of the 8th ACM*

SIGMOD workshop on Research issues in data mining and knowledge discovery, 2–11.

Liu, Z.; Pei, W.; Lan, D.; and Ma, Q. 2024. Diffusion language-shapelets for semi-supervised time-series classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 14079–14087.

Martin, O.; Kotsia, I.; Macq, B.; and Pitas, I. 2006. The eNTERFACE'05 audio-visual emotion database. In *22nd international conference on data engineering workshops (ICDEW'06)*, 8–8. IEEE.

McKeown, G.; Valstar, M.; Cowie, R.; Pantic, M.; and Schroder, M. 2011. The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE transactions on affective computing*, 3(1): 5–17.

Ren, M.; Huang, X.; Liu, J.; Liu, M.; Li, X.; and Liu, A.-A. 2023. MALN: Multimodal Adversarial Learning Network for Conversational Emotion Recognition. *IEEE Transactions on Circuits and Systems for Video Technology*.

Shen, W.; Chen, J.; Quan, X.; and Xie, Z. 2021. Dialogxl: All-in-one xlnet for multi-party conversation emotion recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 13789–13797.

Soleymani, M.; Lichtenauer, J.; Pun, T.; and Pantic, M. 2011. A multimodal database for affect recognition and implicit tagging. *IEEE transactions on affective computing*, 3(1): 42–55.

Tan, C. W.; Dempster, A.; Bergmeir, C.; and Webb, G. I. 2022. MultiRocket: multiple pooling operators and transformations for fast and effective time series classification. *Data Mining and Knowledge Discovery*, 36(5): 1623–1646.

Tao, W.; Li, C.; Song, R.; Cheng, J.; Liu, Y.; Wan, F.; and Chen, X. 2020. EEG-based emotion recognition via channel-wise attention and self attention. *IEEE Transactions on Affective Computing*.

Tellamekala, M. K.; Amiriparian, S.; Schuller, B. W.; André, E.; Giesbrecht, T.; and Valstar, M. 2023. COLD fusion: Calibrated and ordinal latent distribution fusion for uncertainty-aware multimodal emotion recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Wang, S.; Bao, Z.; Culpepper, J. S.; Sellis, T.; Sanderson, M.; and Qin, X. 2017. Answering top-k exemplar trajectory queries. In *2017 IEEE 33rd International Conference on Data Engineering (ICDE)*, 597–608. IEEE.

Wang, Y.; Li, Y.; and Cui, Z. 2024. Incomplete multimodality-diffused emotion recognition. *Advances in Neural Information Processing Systems*, 36.

Wang, Y.; Xu, Y.; Yang, J.; Wu, M.; Li, X.; Xie, L.; and Chen, Z. 2024. Graph-Aware Contrasting for Multivariate Time-Series Classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 15725–15734.

Wen, C.; Jia, G.; and Yang, J. 2023. DIP: Dual Incongruity Perceiving Network for Sarcasm Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2540–2550.

Xu, C.; Zhu, J.; Zhang, J.; Han, Y.; Chu, W.; Tai, Y.; Wang, C.; Xie, Z.; and Liu, Y. 2023. High-fidelity Generalized Emotional Talking Face Generation with Multi-modal Emotion Space Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6609–6619.

Xu, Y.; Chen, H.; Yu, J.; Huang, Q.; Wu, Z.; Zhang, S.-X.; Li, G.; Luo, Y.; and Gu, R. 2024. Secap: Speech emotion captioning with large language model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 19323–19331.

Yang, D.; Yang, K.; Li, M.; Wang, S.; Wang, S.; and Zhang, L. 2024. Robust emotion recognition in context debiasing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12447–12457.

Ye, L.; and Keogh, E. 2009. Time series shapelets: a new primitive for data mining. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, 947–956.

Zhang, X.; Gao, Y.; Lin, J.; and Lu, C.-T. 2020. Tapnet: Multivariate time series classification with attentional prototypical network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 6845–6852.

Zhang, Z.; Wang, L.; and Yang, J. 2023. Weakly Supervised Video Emotion Detection and Prediction via Cross-Modal Temporal Erasing Network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18888–18897.

Zheng, Y.; Zhang, L.; Xie, X.; and Ma, W.-Y. 2009. Mining interesting locations and travel sequences from GPS trajectories. In *Proceedings of the 18th international conference on World wide web*, 791–800.

Zhou, C.; Frankowski, D.; Ludford, P.; Shekhar, S.; and Terveen, L. 2007. Discovering personally meaningful places: An interactive clustering approach. *ACM Transactions on Information Systems (TOIS)*, 25(3): 12–es.

Zhu, T.; Li, L.; Yang, J.; Zhao, S.; and Xiao, X. 2022. Multimodal Emotion Classification with Multi-level Semantic Reasoning Network. *IEEE Transactions on Multimedia*.