Hidden Follower Detection: How Is the Gaze-Spacing Pattern Embodied in Frequency Domain?

Shu Li[1], **Ruimin Hu**[1*], **Suhui Li**[1], **Liang Liao**[2]

[1]School of Cyber Engineering, Xidian University
[2]School of Computer Science and Engineering, Nanyang Technological University
{shuli, rmhu, suhuili}@xidian.edu.cn, liang.liao@ntu.edu.sg

Abstract

Spatiotemporal social behavior analysis is a technique that studies the social behavior patterns of objects and estimates their risks based on their trajectories. In social public scenarios such as train stations, hidden following behavior has become one of the most challenging issues due to its probability of evolving into violent events, which is more than 25%. In recent years, research on hidden following detection (HFD) has focused on differences in time series between hidden followers and normal pedestrians under two temporal characteristics: gaze and spatial distance. However, the time-domain representation for time series is irreversible and usually causes the loss of critical information. In this paper, we deeply study the expression efficiency of time/frequency domain features of time series, by exploring the recovery mechanism of features to source time series, we establish a fidelity estimation method for feature expression and a selection model for frequency-domain features based on the signal-to-distortion ratio (SDR). Experimental results demonstrate the feature fidelity of time series and HFD performance are positively correlated, and the fidelity of frequency-domain features and HFD performance are significantly better than the time-domain features. On both real and simulated datasets, the accuracy of the proposed method is increased by 3%, and the gaze-only module is improved by 10%. Related research has explored new methods for optimal feature selection based on fidelity, new patterns for efficient feature expression of hidden following behavior, and the mechanism of multimodal collaborative identification.

Introduction

Many achievements have been made in modeling abnormal behaviors with obvious visual features (Liu et al. 2022; Ionescu et al. 2019). As a prelude to many criminal crimes, hidden following is a behavior of secretly tracking and monitoring behind the target. Hidden follower detection (HFD) seeks to identify the hidden follower in all pedestrians from the surveillance video. However, methods based on computer vision (Li, Zhang, and Diao 2020; Jiang et al. 2020; Duan et al. 2020), end-to-end learning (Zhou et al. 2019), or sensors (Wang et al. 2017) are helpless for this task because there are no obvious posture characteristics; Then mainly studies the relative position of two trajectories (Andersson et al. 2008; Siqueira et al. 2011) and temporal-spatial trajectory (Kjærgaard et al. 2013; Li et al. 2013; Xie, Ren, and Liu 2020; Jiang et al. 2018), nevertheless, the following behavior is ubiquitous, making it difficult to distinguish whether the pedestrians behind with similar trajectories are due to coincidence or hidden following intention.

The study of "hidden" behavior has seen new light in recent works (Xu et al. 2022, 2021). They consider behavior patterns can reflect human intentions: first found there are significant pattern differences in gaze-spatial behavior between hidden followers and normal pedestrians. In the time domain, they extract the gaze state series and distance series from the surveillance video and generate gaze-spacing-flow features to represent the gaze-spatial low of walking. Then, a hidden follower detection framework embedded with gaze-spacing-flow (HFDF-GS) is proposed to improve the accuracy of HFD. However, the time domain representation of time series often leads to information loss, which is mainly attributed to 1) *Irreversibility*: time-domain features are irreversible, for example, we cannot reconstruct from the gaze frequency to the gazing state series. This one-direction feature extraction will inevitably lose some original information; 2) *One-sidedness*: the subjective time-domain features such as gaze frequency only describe the hidden following behavior in limited aspects.

Nevertheless, early research has proved the vulnerability of time-domain parameters (S.B. and Rao 2016). Signal changes not only with time but also with frequency and phase, etc. Any movement signal in frequency domain can be decomposed into different sine waves, which makes it reflect the essence of things or phenomena from an objective view (Yadav and Rai 2020): 1) *Better representation*: many studies have demonstrated the frequency-domain features can characterize more accurate and comprehensive original information (Van Segbroeck, Tsiartas, and Narayanan 2013), such as Mel-frequency cepstrum coefficient (MFCC) (Lai et al. 2022), wavelet feature (Lee et al. 2022), spectral entropy (Yu et al. 2022), and Constant-Q Cepstral Coefficients (CQCC) (Bhattacharjee et al. 2020), etc; 2) *Reversibility*: reversible frequency-domain features can be lossily reconstructed, which maximizes the preservation of original information.

In this paper, in order to prove the information loss of

*Corresponding author

time-domain features, we define fidelity using signal-to-distortion ratio (SDR) (Boeddeker et al. 2021) to estimate the expression efficiency of time/frequency features and provide a universal selection method for frequency-domain features. Based on frequency-domain analysis, we then propose a decision-fusion hidden follower detection framework based on reversible time-frequency transform (HFDF-TF). HFDF-TF detects the gaze direction and trajectory for all pedestrians in the surveillance to generate the gaze state series and distance series (Xu et al. 2022), and then our framework acts on it to mine the frequency-domain features that distinguish hidden followers from location followers. The contributions of this paper are as follows:

- We demonstrate there is information loss in time-domain features by exploring the recovery of features to the original time series. We establish a fidelity estimation method for feature expression and a selection model for frequency-domain features based on SDR.
- We explore the expression mechanism of human behavior in frequency domain, there are also significant differences in gaze-spacing patterns between the hidden follower and normal pedestrians.
- In HFDF-TF, we first redefine the gaze state in gaze state series; then introduce the decision-fusion training to fully integrate the characteristics of both gaze features and spacing features. Compared to the baselines, HFDF-TF achieves a considerable improvement.

Definitions and Preliminaries

Xu et al. (2022) defines two kinds of following pedestrians: *Position follower*: pedestrians with following characteristics in temporal-spatial position and *Hidden follower*: position followers with real hidden following intention.

Definition 1 (Position follower) *Given two moving pedestrians $\mathcal{F} = \{\boldsymbol{f}_1, \cdots \boldsymbol{f}_k, \cdots\}$ and $\mathcal{T} = \{\boldsymbol{t}_1, \cdots \boldsymbol{t}_k, \cdots\}$, where $\boldsymbol{f}_k = (u_{fk}, v_{fk})$ and $\boldsymbol{t}_k = (u_{tk}, v_{tk})$ are 2D locations at the timestamp k. Given a distance threshold τ, if 1): $\Delta(k) = \|\boldsymbol{f}_k - \boldsymbol{t}_k\| < \tau$; 2): $(\boldsymbol{t}_k - \boldsymbol{f}_k) \cdot (\boldsymbol{f}_{k+1} - \boldsymbol{f}_k) > 0$ (the angle of $\mathcal{F}$ and $\mathcal{T}$ in moving direction is less than $90°$); and 3): $(\boldsymbol{t}_{k+1} - \boldsymbol{t}_k) \cdot (\boldsymbol{f}_{k+1} - \boldsymbol{f}_k) > 0$ ($\mathcal{T}$ is in front of $\mathcal{F}$ in $\mathcal{F}'s$ moving direction). In a time interval, if $\mathcal{F}$ follows $\mathcal{T}$ more than ϵ (default=50%) frames and the distance is always less than τ, we say $\mathcal{F}$ follows $\mathcal{T}$ in this time interval, and $\mathcal{F}$ is defined as a position follower.*

Definition 2 (Hidden follower) *If a pedestrian $\mathcal{F}$ is walking with the intention to "know the real-time position of $\mathcal{T}$" and to "not be found by $\mathcal{T}$". We say $\mathcal{F}$ hidden follows $\mathcal{T}$, and $\mathcal{F}$ is defined as a hidden follower $\mathcal{H}$.*

We call who are not hidden followers $\mathcal{H}$ as normal pedestrians $\overline{\mathcal{H}}$, there are generally two types: acquaintances ($\overline{\mathcal{H}}_1$) and strangers ($\overline{\mathcal{H}}_2$). $\overline{\mathcal{H}}$ may be the position followers or not.

Gaze-Spacing Flow. Xu et al. (2022) define the gaze-spacing pattern to represent behavioral differences between the hidden follower $\mathcal{F}$ and the target $\mathcal{T}$:

- *Gaze pattern*: $\mathcal{H}$ need to gaze at $\mathcal{T}$ frequency to prevent being lost, nor too frequently to avoid being found.
- *Spacing pattern*: $\mathcal{H}$ should not be too far away from $\mathcal{T}$ to prevent being lost, nor too close to avoid being found.

In the time domain, Xu et al. (2022) represents the spacing pattern of the distance series using the spacing flow features: distance range and average distance; and represents the gaze pattern of the gaze state series using the gaze flow features: gaze frequency and gaze density.

Gaze State Series and Distance Series. In HFDF-GS (Xu et al. 2022), the gaze state series only uses a coarse-grained representation (1 or 0) of threshold dichotomy to determine gaze or not, which is not enough to describe complex gaze behavior. In this paper, we introduce a dynamic score to represent the gaze degree at timestep i.

For the "short-time analysis" in the frequency domain, given a gaze state series and a distance series with L frames for a video, the frame rate is fps, we first divide the series into multiple overlapping segments through a sliding window (the overlap between two frames is to maintain smooth transition): we set the window length as w(s), each movement of the window (frameshift) is inc(s), so the overlap is $overlap = w - inc$. Then, the number of segments N is:

$$N = \frac{L - w}{inc} + 1. \tag{1}$$

Each segment contains $w \times fps$ frames (timestamps) of the gaze state or distance information. For each segment, the gaze state series is described by: $\mathcal{G} = \{g_1, g_2 \cdots g_{w \times fps}\}$, where $-1 \leq g_i \leq 1$ that means the degree of $\mathcal{F}$ gaze at $\mathcal{T}$ who in front in timestamp $\boldsymbol{i}$. Similarly, the distance series is recorded as $\mathcal{D} = \{d_1, d_2 \cdots d_{w \times fps}\}$, where d_i denotes the following distance in timestamp i. Obviously, the segmentation of time series can also preserve temporal variability.

Frequency-Domain Features. To analyze the universality of frequency-domain features for hidden following behavior, we choose a classical feature and a complex feature:

Mel-Frequency Cepstral Coefficient (MFCC) (Murty and Yegnanarayana 2006; Brown et al. 2020): MFCC is the most common and representative feature (Lee et al. 2019). The time-frequency transform is based on the Short-Time Fourier Transform (STFT) (Lu et al. 2009).

Constant-Q Cepstral Coefficients (CQCC) (Todisco et al. 2016; Bhattacharjee et al. 2020): CQCC has a better time-frequency resolution, but has high time and computational complexity. The time-frequency transform is based on constant-Q transform (CQT) (Shah et al. 2023).

Considering the CQCC is not conducive to the timeliness of HFD, the subsequent frequency-domain analysis is mainly based on the most representative MFCC feature.

Expression Efficiency of Features

In this section, we will evaluate the expression efficiency of time/frequency domain features from two aspects: fidelity and mode differentiation, while proving there is greater information loss in time-domain features.

Fidelity. The signal-to-distortion ratio (SDR) (Boeddeker et al. 2021) when recovering the initial time series can imply the information loss rate, the smaller the SDR, the more

Feature	Time domain		Frequency domain	
	Gaze	Spacing	Gaze	Spacing
Fidelity	52%	60%	73%	77%

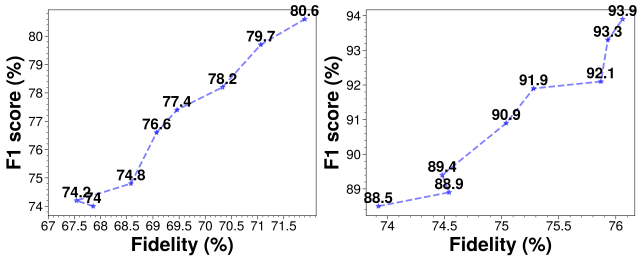
Table 1: The fidelity of time/frequency features.

information the feature saves. We further define fidelity to represent the degree to which features retain information:

$$fidelity = 1 - SDR = 1 - \frac{\left\|\mathcal{L} - \widehat{\mathcal{L}}\right\|^2}{\|\mathcal{L}\|^2} \quad (2)$$

Fidelity for Time/Frequency Domain Features. For the time-domain features, we try to reconstruct using an auto-regressive model: Transformer, we keep the network structure, set $\widehat{\mathcal{L}}$ as the outputting reconstructed series, and introduce the SDR as the loss function. 70% of the real dataset for training and 30% for testing, finally obtaining the average fidelity of the test set. For frequency-domain features, we obtain the optimal 4D gaze/spacing MFCCs, and the initial time series is rebuilt by its inverse operation. As depicted in Table 1, the fidelity of time-domain features is much lower than frequency-domain features, which proves time-domain features may lose more critical information, while frequency-domain features have better expression efficiency for time series. However, the significant fidelity difference in the time-frequency domain in the table may also be due to the technical limitations of time-domain recovery.

Fidelity for Feature Selection in Frequency Domain. To explore the impact of the fidelity of frequency-domain features on the recognition performance of hidden followers, we obtain the fidelity of MFCCs under different $dimension$, w(s), and $overlap$(s). The line graph is shown in Fig. 1, both in gaze-spacing pattern, we find HFD performance is positively correlated with feature fidelity, although there will be a slight disturbance in F_1 score when the fidelity of the two features is very similar. We can say when the feature fidelity increases by more than 0.5%, the HFD performance is also likely to show positive growth. So the optimal frequency-domain feature can be chosen based on fidelity without tedious training. This method is widely applicable to other reversible frequency-domain features or tasks.



(a) Gaze MFCCs (b) Spacing MFCCs

Figure 1: Relationships between the fidelity of various gaze/spacing MFCCs and HFD performance (F_1 score).

Control group	Real-HFD		Sim-HFD	
	Gaze	Spacing	Gaze	Spacing
$\mathcal{L}\overline{\mathcal{L}}$	8.42E-03	1.51E-10	6.30E-04	2.74E-10
$\mathcal{L}\mathcal{H}_T$	0.076	1.25E-04	0.032	1.76E-06
$\mathcal{L}\mathcal{H}_F$	1.21E-39	8.52E-23	5.65E-50	2.85E-38
$\mathcal{H}_F\overline{\mathcal{H}}_F$	8.82E-03	5.50E-08	2.64E-10	4.53E-15
$\mathcal{H}_F\mathcal{H}_F$	0.250	0.497	0.311	0.428
$\overline{\mathcal{H}}_F\overline{\mathcal{H}}_F$	0.437	0.372	0.496	0.410
$\mathcal{H}_F\mathcal{H}_T$	1.15E-21	5.44E-29	3.44E-47	7.31E-55

Table 2: Mode differentiation in control groups: p-value.

Mode Differentiation of Source Information and Features. On both real and simulated datasets, by performing K-S tests (Bickel 1969) on the k-means clustering modes, we set up some control groups to compare the pattern differences of time-frequency features or source series (Table 2). In this section, we only analyze the mode differences between initial time series $\mathcal{L}$ and time/frequency domain features: gaze state series and gaze flow of $\mathcal{H}_T$ ($\mathcal{H}$ in time domain), distance series and spacing flow of $\mathcal{H}_T$, gaze state series and gaze MFCCs of $\mathcal{H}_F$ ($\mathcal{H}$ in frequency domain), distance series and spacing MFCCs of $\mathcal{H}_F$. The pattern differences between $\mathcal{L}$ and the MFCCs ($\mathcal{L}\mathcal{H}_F$) are more significant from the gaze-spacing flow ($\mathcal{L}\mathcal{H}_T$), this indicates frequency-domain features have strong expressive power for time series. Moreover, our gaze-only module (HFDF-GF, see Table 3) is greatly enhanced due to the significant difference in gaze pattern (0.076 in $\mathcal{L}\mathcal{H}_T$ vs 1.21E-39 in $\mathcal{L}\mathcal{H}_F$).

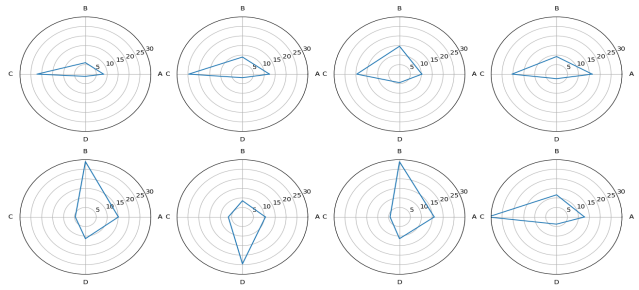
Gaze-Spacing Pattern in Frequency Domain

Based on the above analysis, we make and verify the following assumption:

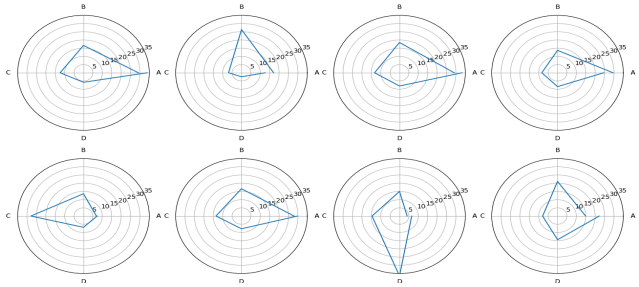
Assumption 1 (Gaze-Spacing pattern) *In frequency domain, the gaze-spacing pattern between the hidden followers and the normal pedestrians is significantly different.*

MFCC Clustering Modes. For all hidden following pairs and normal walking pairs in real dataset, we extracted the d-dimensional MFCCs and then conducted the k-means clustering analysis, they are clustered into four modes: A, B, C, D. Fig. 2 shows the comparison of MFCC mode distribution with the radar chart. It indicates the gaze pattern of $\mathcal{H}$ is more likely concentrated in mode C, and the spacing pattern of $\mathcal{H}$ is more likely distributed in mode A and B; yet the clustering modes of $\overline{\mathcal{H}}$ is disorderly and different to $\mathcal{H}$.

Mode Differentiation. From the three control groups in Table 2: $\mathcal{H}_F\overline{\mathcal{H}}_F$, $\mathcal{H}_F\mathcal{H}_F$ and $\overline{\mathcal{H}}_F\overline{\mathcal{H}}_F$ ($\overline{\mathcal{H}}$ in frequency domain). As for $\mathcal{H}_F$ and $\overline{\mathcal{H}}_F$, there are significant differences in spacing pattern ($p \ll 0.05$), but relatively not significant in gaze pattern on Real-HFD (p=8.82E-03), which also leads to our spacing-only module (HFDF-SF) being better than the gaze-only module (HFDF-GF) (see Table 3). Besides, there are significant differences in gaze-spacing pattern between



(a) Gaze modes



(b) Spacing modes

Figure 2: The MFCC clustering modes of gaze-spacing pattern (radar chart). (a) and (b) showed the mode distribution of gaze and spacing MFCCs respectively. Each radar chart displays the mode distribution of one video, the first and second line is the hidden following and normal walking pairs.

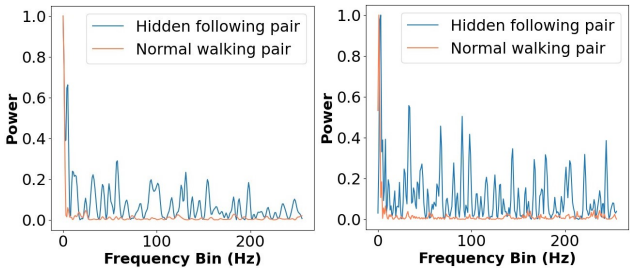
$\mathcal{H}_F$ and $\mathcal{H}_T$ (p$\ll$ 0.05). It proves that frequency-domain features are completely different from the time domain.

Power Spectrum and Spectrogram. At the micro level, we acquired the power spectrum (Kugler 2022) of each segment in $\mathcal{G}$ and $\mathcal{D}$. When the hidden following occurs, compared to normal walking, it shows a significant increase in high-frequency components (see Fig. 3(a)(b)).

At the global level, the spectrogram (Decorsière et al. 2015) represents the time variation of the frequency structure. As shown in Fig. 4, for both gaze and spacing pattern, the whole moving process of $\overline{\mathcal{H}}$ is relatively concentrated in low frequency (Fig. 4(b)(d)). Fig. 4(d) shows disordered power changes in spacing pattern, it implies the random walking speed of $\overline{\mathcal{H}}$ because there is no special intention. Fig. 4(b) have high-frequency aggregation that may be due to the $\overline{\mathcal{H}}$ always habitually staring at a certain place (no need to hide). On the contrary, the higher power of $\mathcal{H}$ changes frequently but not intensively (Fig. 4(a)(c)), it unveils the law of hidden following behavior: 1) *Can't gaze frequently, whereas can't ignore*; 2) *Not too close nor too far*.

HFDF Based on Time-Frequency Transform

The decision-fusion hidden follower detection framework based on time-frequency transform (HFDF-TF) is designed to detect the hidden follower in the surveillance: specify a target pedestrian $\mathcal{T}$, HFDF-TF will predict the probability P that each pedestrian hidden following $\mathcal{T}$. If P>0.6, the pedestrian is recognized as a hidden follower. Fig. 5 shows the overall architecture of HFDF-TF.



(a) Gaze structure (b) Spacing structure

Figure 3: Power spectrum of the gaze-spacing pattern when hidden following occurs. (a) and (b) depicted the frequency (power) structure of the gaze and spacing patterns of the hidden following pair and the normal walking pair respectively.

Trajectory Tracking

We first use the TraDeS (Wu et al. 2021) model (online multi-target tracker) to track the pixel coordinates of all pedestrians in the surveillance video and also convert the pixel coordinates into real ground coordinates by perspective transformation. The distance series of each pedestrian pair is calculated by their relative real ground distance.

Gaze State Series Extraction

The gaze state of the follower to the target needs to be determined by gaze direction.

Gaze Direction Detection First, the HFDF-FT combines DensePose model (Güler, Neverova, and Kokkinos 2018) with the TraDeS tracking results for more accurate head-box tracking. Then, the Gaze360 model (Kellnhofer et al. 2019) outputs the 2D gaze direction for each pedestrian.

Gaze State Series The gaze state series $\mathcal{G}$ consists of the gaze state of each frame. Note that we introduce "gaze degree (score)" to indicate the gaze state instead of using a simple 1 or 0 to denote gaze or not (1 when the gaze angle is less than 60, otherwise it is 0) (Xu et al. 2022). For i-th frame in the video, the gaze state g_i ($g_i \in \mathcal{G}$) from $\mathcal{F}$ to $\mathcal{T}$ is calculated by the ground coordinates of $\mathcal{F}$: (u_{fi}, v_{fi}) and $\mathcal{T}$: (u_{ti}, v_{ti}) and the $2D$ gaze directions of $\mathcal{F}$: (gd_x, gd_y).

$$\Delta u_i = u_{fi} - u_{ti},\ \Delta v_i = v_{fi} - v_{ti}, \tag{3}$$

$$\alpha = arccos \frac{(gd_x, gd_y) \cdot (\Delta u_i, \Delta v_i)}{|(gd_x, gd_y)| \cdot |(\Delta u_i, \Delta v_i)|}, \tag{4}$$

$$g_i = 1 - 2 * angle\,(\alpha)\,/180, \tag{5}$$

where $angle\,(\alpha) \in [0, 180]$, and $-1 \leq g_i \leq 1$ represents the degree of $\mathcal{F}$ gaze at $\mathcal{T}$ in timestamp i. The closer the gaze angle is to 0, the $\mathcal{F}$ is more likely to gaze at $\mathcal{T}$.

Frequency-Domain Features Extraction

After segmentation by Eq. 1, the distance series and gaze state series are divided into N segments. We extracted the d-dimensional MFCC of each gaze segment and spacing segment, and finally obtained the $(N \times d)$ MFCC feature matrix as the gaze feature and the spacing feature, and input them

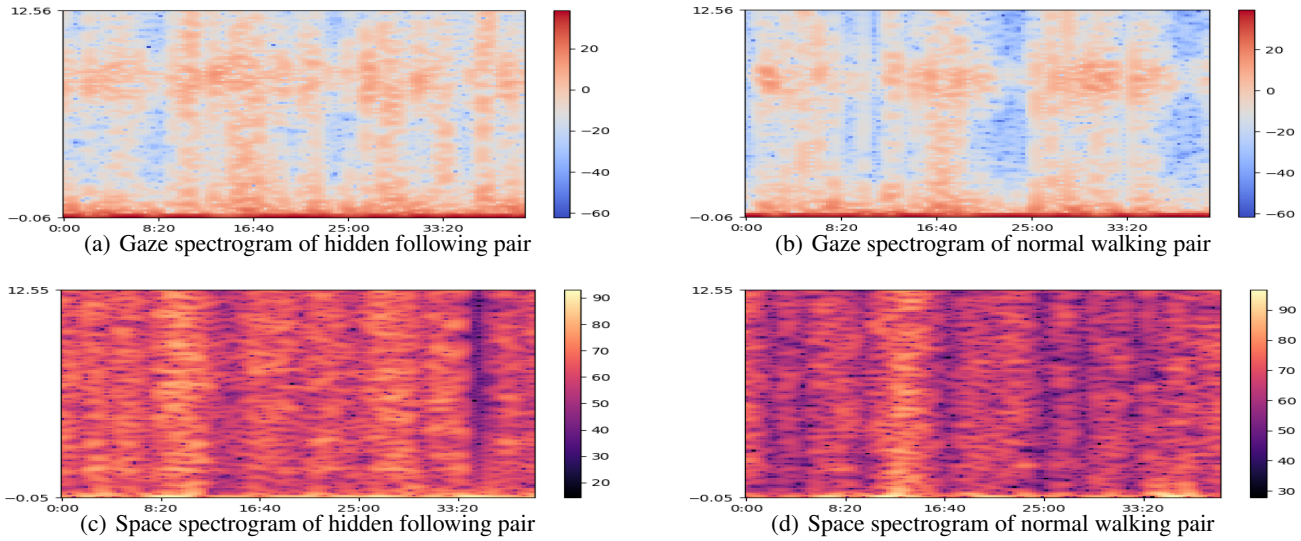


Figure 4: Spectrograms of the gaze-spacing pattern in hidden following and normal walking pair (x label: Time, y label: Hz).

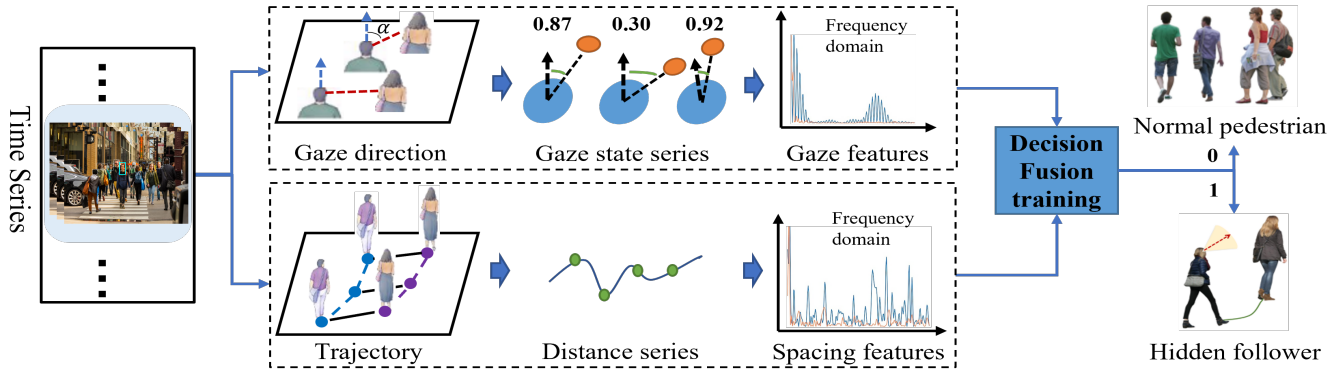


Figure 5: Architecture of the proposed HFDF-TF. HFDF-TF first detects the trajectory and gaze direction of each pedestrian in the surveillance video to acquire the distance series and gaze state series, then transforms them to frequency domain to obtain the spacing MFCCs and gaze MFCCs, they were each sent to a TSC model and outputs the probability of "being a hidden follower", respectively. Finally, the final output of HFDF-TF comes from the decision fusion of the two TSC models.

into two independent TWIESN (Tanisaro and Heidemann 2016) models for decision fusion training, respectively. Similarly, obtain the $(N \times d^{'})$ CQCC feature matrix.

Decision Fusion Training

We apply DS evidence theory (DS inference) (Martin, Zhang, and Liu 2010) as the decision fusion strategy (see Fig. 5). Regard the two TWIESN classifiers as m_1 and m_2, we defined the following identification domain:

$$\Psi = \{\theta_1, \theta_2\}, \tag{6}$$

where θ_1 and θ_2 represents the proposition of: $\mathcal{F}$ is a hidden follower and $\mathcal{F}$ is not a hidden follower, respectively. θ_1 and θ_2 output the prediction probability of m_1 and m_2, respectively, which is recorded as: $m_1 : \{m_1(\theta_1), m_1(\theta_2)\}$ and $m_2 : \{m_2(\theta_1), m_2(\theta_2)\}$. The final decision of proposition θ_1 is calculated as follows.

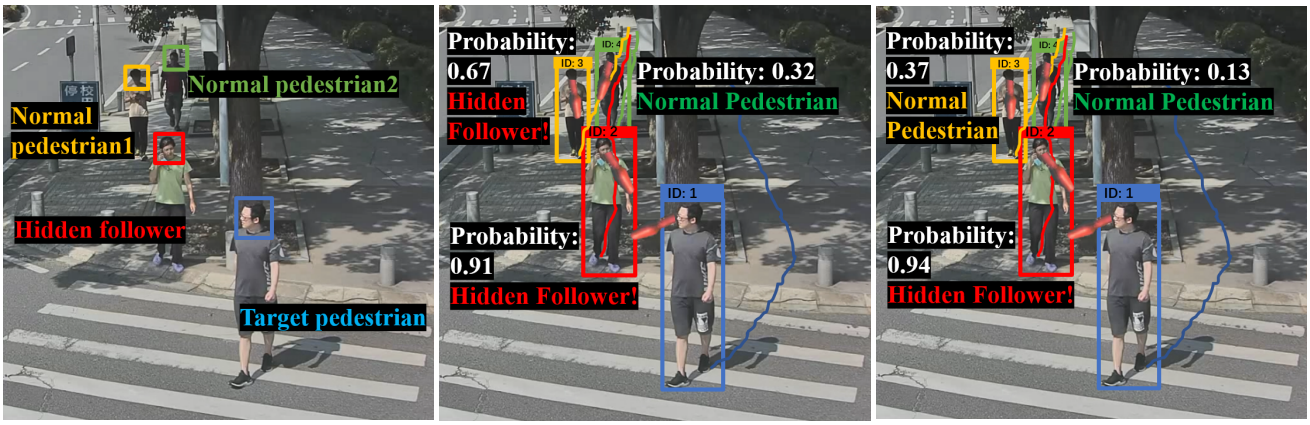
$$P(\theta_1) = \frac{m_1(\theta_1)\, m_2(\theta_1)}{K}, \tag{7}$$

where $K = \sum_{i=1}^{2} m_1(\theta_i)\, m_2(\theta_i)$ is the conflict coefficient.

Experiments and Results

Datasets

Real-HFD. The Real-HFD includes 20 pedestrians with a total video length of 160 minutes. Each video lasts one minute, in which the number of pedestrians varies from 5 to 12. In the hidden following videos, each contains 1 or 2 hidden following pairs, and others are normal pedestrians. The non-hidden following videos contain 4 to 6 pairs of acquaintances. If each sample refers to one behavior of a pedestrian



Ground truth label | Baseline: HFDF-GS | Ours: HFDF-TF

Figure 6: Visual comparison results. The HFDF-GS identifies pedestrians with similar trajectories as hidden followers.

in a 1-minute video, Real-HFD includes 160 samples of hidden followers ($\mathcal{H}$), 280 samples of acquaintances ($\overline{\mathcal{H}}_1$) and 370 samples of strangers ($\overline{\mathcal{H}}_2$). More details can be found in (Xu et al. 2021, 2022).

Sim-HFD. Due to the hidden following data in the real world is rare for training. Xu et al. (2022) established a simulation model to simulate the gaze behavior and movement behavior of four types of pedestrians: 1) The motion of ($\mathcal{H}$) is adjusted according to the principle of following ($\mathcal{T}$) and not being detected. 2) The motion of ($\overline{\mathcal{H}}_1$) is similar to that of ($\mathcal{T}$). 3) The motion of ($\overline{\mathcal{H}}_2$) is basically independent of ($\mathcal{T}$). After performing the simulation model 40 times, once putting 31 pedestrians into the scene: $1\times\mathcal{T}$, $10\times\mathcal{H}$, $10\times\overline{\mathcal{H}}_1$, and $10\times\overline{\mathcal{H}}_2$, there are 1200 series of spacing and gaze behaviors of normal pedestrians and hidden followers in the end. We trained with 4-fold cross-validation, of which the ratio of the training set and test set in both datasets is 7:3.

Experimental Setup

We evaluate the HFDF-TF model from the four aspects:

- How is the HFDF-TF model performance?
- How does the model perform with only gaze MFCCs or spacing MFCCs?
- Is the model applicable to different frequency-domain features?
- Will multimodal feature collaboration further improve HFD performance?

Comparison methods. In this paper, we compare HFDF-TF with the traditional following detection method based on trajectory (Li et al. 2013), the hidden follower detection model HFDF (Xu et al. 2021), and the SOTA HFD model HFDF-GS (Xu et al. 2022).

Parameter settings. The optimal gaze and spacing MFCC parameters in Fig. 1 are: gaze MFCCs with $dimension$ = 4, w = 10s, $overlap$ = 2s and spacing MFCCs with $dimension$ = 4, w = 8s, $overlap$ = 1s.

Evaluate metrics. Precision, Recall, F_1 score, Accuracy and AUC.

The Performance of HFDF-TF

On real-HFD and Sim-HFD, we evaluate the performance of the proposed HFDF-TF with other HFD methods.

Results. Table 3 reveals the performance comparison results on five evaluated metrics. Both on Real-HFD and Sim-HFD, the HFD performance of HFDF-TF is significantly better than the traditional trajectory-based method and the SOTA baseline method. Compared to HFDF-GS, on Sim-HFD, the proposed HFDF-TF is improved 2% to 3% in four evaluate metrics, and improved 3% to 4% on Sim-HFD. Besides, see Fig. 6 for the visual comparison results between HFDF-TF and HFDF-GS.

In order to test the effects of the gaze module, spacing module, and decision fusion strategy separately, the corresponding ablation model is defined as 1) HFDF-GF (gaze-only HFDF-TF), 2) HFDF-SF (spacing-only HFDF-TF), and 3) HFDF-ND (HFDF-TF without decision fusion, which based on the feature concatenation and only one TWIESN network). The ablation results are also displayed in Table 3. Compared to HFDF-G, HFDF-GF meets great improvement in gaze features, an increase of about 9% to 10%; Compared to HFDF-G, the improvement of HFDF-GF in spacing features is about 2% to 3%. Nevertheless, why does the spacing pattern in the frequency domain increase far less than the gaze pattern? The reasons are as follows:

(1) *Explanation from the behavior mode*. Relatively speaking, trajectory tracking is more accurate than the gaze angle. On the other hand, the distance law is a strong constraint, yet whether to gaze is a weak constraint: the target suddenly turns back is a small probability event, and the follower does not need to strictly control his gaze direction.

(2) *Explanation from benefits of time-frequency transform*. The more accurate the time-domain parameters express the behavior, the smaller the additional benefits we can get in the frequency domain. It is obvious that the time-domain gaze parameters in HFDF-GS are not enough to express the following behavior (F_1 score: 71.5%).

The comparison between HFDF-ND and HFDF-TF in Table 3 shows that simple feature concatenation will destroy the unique frequency structure of gaze mode and spacing

Dataset	Model	Precision	Recall	F_1	Accuracy	AUC
Real-HFD	Trajectory	57.1	**97.5**	72.7	62.5	65.1
	HFDF	87.6	86.2	86.9	86.2	87.3
	HFDF-G	71.6	71.4	71.5	71.4	71.2
	HFDF-S	91.4	90.9	91.2	90.9	90.8
	HFDF-GS	92.0	91.4	91.7	91.4	91.6
	HFDF-GF	81.8	80.4	80.6	80.4	80.4
	HFDF-SF	94.0	93.8	93.9	93.8	93.7
	HFDF-ND	92.4	91.8	92.6	91.8	91.7
	HFDF-TF	**94.6**	94.4	**94.5**	**94.4**	**94.4**
Sim-HFD	Trajectory	61.0	**100.0**	75.8	78.7	79.4
	HFDF	86.3	87.9	87.1	87.1	87.0
	HFDF-G	86.3	87.9	87.1	87.1	87.0
	HFDF-S	87.2	88.4	87.8	87.7	87.3
	HFDF-GS	91.9	90.6	91.4	91.2	91.4
	HFDF-GF	88.3	89.5	88.6	88.4	88.4
	HFDF-SF	94.8	94.5	94.6	94.6	94.3
	HFDF-ND	93.9	94.7	93.4	94.8	94.8
	HFDF-TF	**95.5**	95.0	**95.2**	**95.1**	**95.0**

Table 3: Comparison of five evaluate metrics for each model on Real-HFD and Sim-HFD. HFDF-G and HFDF-S represent the gaze-only and the spacing-only model in HFDF-GS.

Feature	Model	Time	Precision	Recall	F_1	Accuracy	AUC
MFCC	HFDF-GF	0.75s	81.8	80.4	80.6	80.4	80.4
	HFDF-SF	0.73s	94.0	93.8	93.9	93.8	93.7
	HFDF-TF	1.22s	94.6	94.4	94.5	94.4	94.4
CQCC	HFDF-GF	1.39s	83.2	81.9	82.1	81.9	82.3
	HFDF-SF	1.33s	95.2	94.6	94.9	94.6	94.6
	HFDF-TF	2.36s	**96.1**	**95.4**	**94.9**	**95.4**	**95.2**

Table 4: Comparison of five evaluate metrics between MFCCs and CQCCs on Real-HFD, and the comparison of each iteration time during training.

mode in the frequency domain, so the performance will not increase but decrease. Therefore, decision fusion training is essential for our task. So far, we are convinced that the analysis of hidden following behavior in the frequency domain has made a significant contribution to HFD.

Different Frequency-Domain Features

On Real-HFD, we compared the performance of MFCCs and CQCCs on HFDF-GF and HFDF-SF, respectively.

Results. Table 4 shows the comparison between MFCCs and CQCCs on Real-HFD. Although CQCCs outperform MFCCs in all three models, it can be seen from the comparison of each iteration time that, the time complexity of CQCCs is almost twice that of MFCCs, which greatly limits the practical application of CQCCs. Yet the extraction process of MFCC involves many other typical frequency-domain features, such as FFT, spectrum, Fbank features, etc. In cases where all metrics are small differences, MFCCs may be a better feature selection.

Model	Precision	Recall	F_1	Accuracy
($\mathrm{G_{mfcc}}$ + $\mathrm{G_{flow}}$, $\mathrm{S_{mfcc}}$)	92.4	90.9	91.3	90.9
($\mathrm{G_{mfcc}}$ + $\mathrm{S_{flow}}$, $\mathrm{S_{mfcc}}$)	92.4	90.9	91.3	90.9
($\mathrm{G_{mfcc}}$, $\mathrm{S_{mfcc}}$ + $\mathrm{G_{flow}}$)	91.1	90.9	90.8	90.9
($\mathrm{G_{mfcc}}$, $\mathrm{S_{mfcc}}$ + $\mathrm{S_{flow}}$)	89.3	88.6	88.2	88.6
($\mathrm{G_{mfcc}}$ + $\mathrm{G_{flow}}$, $\mathrm{S_{mfcc}}$ + $\mathrm{S_{flow}}$)	92.1	91.6	92.0	91.6
HFDF-TF: ($\mathrm{G_{mfcc}}$, $\mathrm{S_{mfcc}}$)	94.6	94.4	94.5	94.4
TF decision	**96.0**	**95.5**	**95.3**	**95.5**

Table 5: Performance of multimodal feature collaboration. Where ($\mathrm{G_{mfcc}}$ + $\mathrm{G_{flow}}$, $\mathrm{S_{mfcc}}$ + $\mathrm{S_{flow}}$) is a decision fusion model for gaze features (gaze MFCCs concatenate with gaze flow) and spacing features (spacing MFCCs concatenate with the spacing flow). "+" denotes the concatenation along specific dimensions, if there is no "+", it means no feature concatenation. TF decision is a decision fusion model for HFDF-GS and HFDF-TF.

The Performance of Multimodal Collaboration

On Real-HFD, we also explored the performance of collaborating time-frequency features. Moreover, we conducted decision fusion training on the HFDF-GS and HFDF-TF to integrate time-frequency domain information fully, the model is recorded as TF decision model.

Results. The gaze-spacing flow (Xu et al. 2022) combines with the gaze MFCCs and spacing MFCCs, respectively, or simultaneously (see Table 5). But compared to HFDF-TF, it doesn't seem to meet expectations. The decision fusion model of HFDF-GS and HFDF-TF achieved the best HFD.

Conclusions

Our paper seeks to open the minds of hidden follower detection (HFD) for a new research agenda. By studying the fidelity of source information recovery of time/frequency domain features, we found the frequency-domain features have better expression efficiency for time series, and establish a selection model for frequency-domain features based on the fidelity. Furthermore, by analyzing the motion of hidden following behavior in frequency domain, we propose a decision fusion hidden follower detection framework based on time-frequency transform (HFDF-TF) to achieve a more efficient HFD. The F_1 score on Real-HFD is improved by 2.8%, and the gaze-only module is improved by 10.1%, the multimodal collaboration performance is improved by 3.6%.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant U22A2035; in part by the Guangxi Natural Science Foundation Program under Grant 2021GXNSFDA075011. We thank Danni Xu for her guidance and help in our work.

References

Andersson, M.; Gudmundsson, J.; Laube, P.; and Wolle, T. 2008. Reporting Leaders and Followers among Trajectories of Moving Point Objects. *GeoInformatica*, 12: 497–528.

Bhattacharjee; Mrinmoy; Prasanna; Mahadeva, S. R.; Guha; and Prithwijit. 2020. Speech/Music Classification Using Features From Spectral Peaks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28: 1549–1559.

Bickel, P. J. 1969. A Distribution Free Version of the Smirnov Two Sample Test in the p-Variate Case. *Annals of Mathematical Statistics*, 40: 1–23.

Boeddeker, C.; Zhang, W.; Nakatani, T.; Kinoshita, K.; Ochiai, T.; Delcroix, M.; Kamo, N.; Qian, Y.; and Haeb-Umbach, R. 2021. Convolutive Transfer Function Invariant SDR Training Criteria for Multi-Channel Reverberant Speech Separation. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 8428–8432.

Brown, C.; Chauhan, J.; Grammenos, A.; Han, J.; Hasthanasombat, A.; Spathis, D.; Xia, T.; Cicuta, P.; and Mascolo, C. 2020. Exploring Automatic Diagnosis of COVID-19 from Crowdsourced Respiratory Sound Data. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, 3474–3484. New York, NY, USA. ISBN 9781450379984.

Decorsière, R.; Søndergaard, P. L.; MacDonald, E. N.; and Dau, T. 2015. Inversion of Auditory Spectrograms, Traditional Spectrograms, and Other Envelope Representations. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(1): 46–56.

Duan, H.; Zhao, Y.; Xiong, Y.; Liu, W.; and Lin, D. 2020. Omni-Sourced Webly-Supervised Learning for Video Recognition. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV*, 670–688. Berlin, Heidelberg. ISBN 978-3-030-58554-9.

Güler, R. A.; Neverova, N.; and Kokkinos, I. 2018. DensePose: Dense Human Pose Estimation in the Wild. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7297–7306.

Ionescu, R. T.; Smeureanu, S.; Popescu, M.; and Alexe, B. 2019. Detecting Abnormal Events in Video Using Narrowed Normality Clusters. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 1951–1960.

Jiang, J.; Nan, Z.; Chen, H.; Chen, S.; and Zheng, N. 2020. Predicting Short-Term Next-Active-Object Through Visual Attention and Hand Position. *Neurocomputing*, 433.

Jiang, N.; Bai, S.; Xu, Y.; Xing, C.; Zhou, Z.; and Wu, W. 2018. Online Inter-Camera Trajectory Association Exploiting Person Re-Identification and Camera Topology. In *Proceedings of the 26th ACM International Conference on Multimedia*, MM '18, 1457–1465. New York, NY, USA: Association for Computing Machinery. ISBN 9781450356657.

Kellnhofer, P.; Recasens, A.; Stent, S.; Matusik, W.; and Torralba, A. 2019. Gaze360: Physically Unconstrained Gaze Estimation in the Wild. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 6911–6920.

Kjærgaard, M.; Blunck, H.; Wüstenberg, M.; Grønbæk, K.; Wirz, M.; Roggen, D.; and Tröster, G. 2013. Time-lag Method for Detecting Following and Leadership Behavior of Pedestrians from Mobile Sensing Data. *IEEE International Conference on Pervasive Computing and Communications (PerCom)*, 18: 22.

Kugler, J. 2022. [Power spectrum]. *Deutsche medizinische Wochenschrift*, 103 38: 1471–2.

Lai, S.-C.; Hung, Y.-H.; Zhu, Y.-C.; Wang, S.-T.; Huang, Q.-X.; Sheu, M.-H.; and Juang, W.-H. 2022. Hardware Accelerator Design of DCT Algorithm With Unique-Group Cosine Coefficients for Mel-Scale Frequency Cepstral Coefficients. *IEEE Access*, 10: 79681–79688.

Lee, C. S.; Li, M.; Lou, Y.; and Dahiya, R. 2022. Restoration of Lung Sound Signals Using a Hybrid Wavelet-Based Approach. *IEEE Sensors Journal*, 22(20): 19700–19712.

Lee, Y.; Min, J.; Han, D.; and ko, H. 2019. Spectro-Temporal Attenton-Based Voice Activity Detection. *IEEE Signal Processing Letters*, PP: 1–1.

Li, S.; Zhang, L.; and Diao, X. 2020. Deep-learning-based human intention prediction using RGB images and optical flow. *Journal of Intelligent & Robotic Systems*, 97(1): 95–107.

Li, Z.; Fei, W.; Crofoot; and Margaret. 2013. Mining Following Relationships in Movement Data. *Proceedings - IEEE International Conference on Data Mining, ICDM*, 458–467.

Liu, Y.; Zhang, H.; Xu, D.; and He, K. 2022. Graph transformer network with temporal kernel attention for skeleton-based action recognition. *Knowledge-Based Systems*, 240(108146).

Lu; Wen-kai; Zhang; and Qiang. 2009. Deconvolutive Short-Time Fourier Transform Spectrogram. *IEEE Signal Processing Letters*, 16(7): 576–579.

Martin, R.; Zhang, J.; and Liu, C. 2010. Dempster–Shafer Theory and Statistical Inference with Weak Beliefs. *Statistical Science*, 25: 72–87.

Murty, K.; and Yegnanarayana, B. 2006. Combining evidence from residual phase and MFCC features for speaker recognition. *IEEE Signal Processing Letters*, 13(1): 52–55.

S.B., S.; and Rao, K. 2016. Voice/Non-voice Detection Using Phase of Zero Frequency Filtered Speech Signal. *Speech Communication*, 81.

Shah, A.; Chen, S.; Zhou, K.; Chen, Y.; and Raj, B. 2023. Approach to Learning Generalized Audio Representation Through Batch Embedding Covariance Regularization and Constant-Q Transforms. *ArXiv*, abs/2303.03591.

Siqueira; Fernando; Bogorny; and Vania. 2011. Discovering Chasing Behavior in Moving Object Trajectories. *Transactions in GIS*, 15: 667–688.

Tanisaro, P.; and Heidemann, G. 2016. Time Series Classification Using Time Warping Invariant Echo State Networks. In *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 831–836.

Todisco; Massimiliano; Delgado; Héctor; Evans; and Nicholas. 2016. A New Feature for Automatic Speaker Verification Anti-Spoofing: Constant Q Cepstral Coefficients. In *Odyssey 2016 - The Speaker and Language Recognition Workshop*.

Van Segbroeck, M.; Tsiartas, A.; and Narayanan, S. 2013. A robust frontend for VAD: Exploiting contextual, discriminative and spectral cues of human voice. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 704–708.

Wang, J.; Chen, Y.; Hao, S.; Peng, X.; and Hu, L. 2017. Deep Learning for Sensor-based Activity Recognition: A Survey. *Pattern Recognition Letters*, 119.

Wu, J.; Cao, J.; Song, L.; Wang, Y.; Yang, M.; and Yuan, J. 2021. Track to Detect and Segment: An Online Multi-Object Tracker. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 12347–12356.

Xie, W.; Ren, G.; and Liu, S. 2020. Video Relation Detection with Trajectory-Aware Multi-Modal Features. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, 4590–4594. New York, NY, USA: Association for Computing Machinery. ISBN 9781450379885.

Xu, D.; Hu, R.; Wang, Z.; Luo, L.; Li, D.; and Zeng, W. 2022. Gaze- and Spacing-Flow Unveil Intentions: Hidden Follower Discovery. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, 2115–2123. New York, NY, USA. ISBN 9781450392037.

Xu, D.; Hu, R.; Xiong, Z.; Wang, Z.; Luo, L.; and Li, D. 2021. Trajectory is Not Enough: Hidden Following Detection. In *Proceedings of the 29th ACM International Conference on Multimedia*, MM '21, 5373–5381. New York, NY, USA. ISBN 9781450386517.

Yadav, S.; and Rai, A. 2020. Frequency and Temporal Convolutional Attention for Text-Independent Speaker Recognition. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6794–6798.

Yu, Y.; Zhang, D.; Li, Y.; and Zhang, Z. 2022. Multi-Proxy Learning from an Entropy Optimization Perspective. In Raedt, L. D., ed., *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, 1594–1600. Main Track.

Zhou, J. T.; Du, J.; Zhu, H.; Peng, X.; Liu, Y.; and Goh, R. 2019. AnomalyNet: An Anomaly Detection Network for Video Surveillance. *IEEE Transactions on Information Forensics and Security*, 14: 2537–2550.