DOI: 10.65322/j.srd.05192026.112

FACE-BASED NON-VERBAL CUES FOR PEDESTRIAN CROSSING-INTENTION ESTIMATION IN AUTONOMOUS DRIVING: A CRITICAL REVIEW AND A RESOLUTION-AWARE FUSION FRAMEWORK

Zokhidov A., *Senior Lecturer,*
E-mail: zokhidovabdulla@gmail.com
University of Exact and Social Sciences

Abstracts. *Estimating whether a pedestrian is about to cross the road is a key perception task for autonomous vehicles (AVs). Most published estimators rely on trajectory, body pose and scene context, while face-based cues (head orientation, gaze and facial expression) are used rarely and inconsistently. This paper presents a structured review of face-based non-verbal cues for crossing-intention estimation and examines when such cues are physically observable from an on-board camera. Published work is organised into a taxonomy of four cue families, and public datasets are compared by the face-related annotations they provide. A pinhole-camera analysis then relates the face height in pixels to the pedestrian's distance and to the vehicle's stopping distance. For a 1920-pixel camera with a 60° horizontal field of view, an adult face falls below an assumed 48-pixel threshold for expression recognition at about 6.9 m and below a 20-pixel threshold for head-orientation estimation at about 16.6 m, whereas the stopping distance at 50 km/h is about 20.7 m. Head orientation is therefore the most usable face cue at urban speeds, while facial expression is mainly relevant at low speeds or with higher-resolution optics. On this basis, a quality-gated fusion framework is proposed in which the face branch contributes only when the face is resolvable, and recommendations are given for datasets, evaluation and privacy.*

Annotatsiya. *Piyodaning yo'lni kesib o'tish niyatini baholash avtonom transport vositalari (ATV) uchun asosiy idrok etish vazifalaridan biridir. Mavjud baholash modellarining aksariyati harakat trayektoriyasi, tana holati (pose) va sahna kontekstiga tayanadi, yuzga asoslangan belgilar (bosh orientatsiyasi, nigoh yo'nalishi va yuz mimikasi) esa kamdan-kam hamda tizimsiz qo'llaniladi. Ushbu maqolada piyodalarning yo'lni kesib o'tish niyatini baholashda yuzga asoslangan noverbal belgilarning tizimli sharhi keltiriladi hamda ushbu belgilar bort kamerasi orqali jismoniy jihatdan qachon aniqlanishi mumkinligi o'rganiladi. Ilmiy ishlar to'rtta belgi oilasidan iborat taksonomiyaga ajratildi, ommaviy ma'lumotlar to'plamlari (datasetlar) esa ulardagi yuzga oid annotatsiyalar bo'yicha qiyoslandi. Shundan so'ng, sodda kamera (pinhole-camera) tahlili asosida piksellardagi yuz balandligi, piyodagacha bo'lgan masofa va avtomobilning to'xtash masofasi o'zaro bog'landi. Gorizontal ko'rish maydoni* 60° *bo'lgan 1920 pikselli kamera uchun katta yoshli insonning yuzi mimikani aniqlash uchun qabul qilingan 48 pikselli chegaradan taxminan 6,9 m masofada, bosh orientatsiyasini baholash uchun 20 pikselli chegaradan esa taxminan 16,6 m masofada pastga tushadi; bunda 50 km/soat tezlikdagi to'xtash masofasi taxminan 20,7 m ni tashkil etadi. Shu sababli, shahar tezliklarida bosh orientatsiyasi yuzga oid eng foydali belgi hisoblanadi, yuz mimikasi esa faqat past tezliklarda yoki yuqori aniqlikdagi optika bilan ahamiyat kasb etadi. Shu asosda yuz tarmog'i faqat yuz yetarli aniqlikda ko'ringandagina hisobga olinadigan sifat filtri bilan boshqariluvchi integratsiyalash (quality-gated fusion) arxitekturasi taklif etildi, shuningdek, ma'lumotlar bazalari, baholash va maxfiylik bo'yicha amaliy tavsiyalar berildi.*

Аннотация. *Оценка намерения пешехода перейти дорогу является ключевой задачей восприятия для беспилотных транспортных средств (БТС). Большинство существующих алгоритмов оценки опираются на траекторию движения, позу тела и контекст сцены, в то время как признаки на основе лица (ориентация головы, направление взгляда и выражение лица) используются редко и бессистемно. В данной статье представлен структурированный обзор невербальных признаков лица для прогнозирования*

DOI: 10.65322/j.srd.05192026.112

намерения перехода и проанализировано, в каких условиях такие признаки физически различимы с помощью бортовой камеры. Опубликованные работы систематизированы в таксономию из четырех групп признаков, а открытые датасеты сопоставлены по типам аннотаций, связанных с лицом. С использованием модели камеры-обскуры (pinhole-camera) установлена зависимость между высотой лица в пикселях, расстоянием до пешехода и остановочным путем автомобиля. Для камеры с разрешением 1920 пикселей и горизонтальным углом обзора 60° *размер лица взрослого человека опускается ниже порогового значения 48 пикселей (необходимого для распознавания мимики) на расстоянии около 6,9 м, а ниже порога в 20 пикселей (для оценки ориентации головы) — примерно на 16,6 м, тогда как остановочный путь при скорости 50 км/ч составляет около 20,7 м. Таким образом, ориентация головы выступает наиболее применимым лицевым признаком при городских скоростях движения, тогда как мимика лица актуальна преимущественно на малых скоростях или при наличии оптики высокого разрешения. На этой основе предложена структура объединения данных с контролем качества (quality-gated fusion), в которой ветвь анализа лица подключается только при достаточном разрешении, а также сформулированы рекомендации по формированию датасетов, методологии оценки и защите конфиденциальности.*

Keywords: *pedestrian crossing intention, autonomous vehicles, head orientation, gaze, facial expression, multi-cue fusion, image resolution.*

Kalit so‘zlar: *piyodaning kesib o‘tish niyati, avtonom transport vositalari, bosh orientatsiyasi, nigoh yo‘nalishi, yuz mimikasi, ko‘p belgili integratsiyalash, tasvir aniqligi.*

Ключевые слова: *намерение пешехода перейти дорогу, беспилотные транспортные средства, ориентация головы, направление взгляда, выражение.*

Introduction

Road traffic crashes kill about 1.19 million people every year, and pedestrians are among the road users with the highest risk of death or serious injury [1]. Autonomous vehicles (AVs) are expected to reduce this toll, but only if they can interpret the behaviour of the people around them rather than merely detect them [2, 3]. In urban driving, the most consequential of these judgements is whether a pedestrian standing near the kerb is about to step onto the road. A missed crossing is a direct safety hazard, while an unnecessary emergency stop reduces comfort, disturbs following traffic and erodes public trust in automated driving.

Early approaches treated the problem as short-term path prediction with recursive Bayesian filters, which extrapolate the observed motion of the pedestrian [4]. Such filters react only after the pedestrian has started to move. Kooij et al. [5] showed that the prediction horizon can be extended by adding context, including whether the pedestrian’s head is oriented towards the approaching vehicle, the distance to the kerb and the presence of a critical situation. The release of the JAAD [6] and PIE [7] datasets, recorded from vehicle-mounted cameras with behavioural annotations, shifted the field towards learning-based models. Body pose was shown to be a strong early cue [8, 9]; stacked recurrent networks fused pedestrian appearance, ego-vehicle speed and scene context [10]; and a common benchmark exposed large differences in evaluation practice between studies [11]. Later work demonstrated that even a well-designed model operating only on bounding-box sequences is a strong baseline [12], and attention-based fusion of several feature types improved results further [13]. Recent surveys summarise these developments in detail [14, 15].

Behavioural research suggests that pedestrians also communicate non-verbally. In naturalistic recordings, pedestrians frequently looked at the approaching vehicle before crossing [16], and human observers were able to judge crossing intentions from short video sequences of pedestrians at the kerb [17]. The evidence is not one-sided, however: in a field study, pedestrians

DOI: 10.65322/j.srd.05192026.112

relied mainly on the motion of the vehicle and seldom sought eye contact with the driver [18], and psychological research warns that facial movements are not a reliable readout of internal emotional states, let alone of intentions [19]. Face-based cues are therefore plausible but unproven contributors to intention estimation.

Two practical questions have received little systematic attention. First, it is not clear which face-based cue (head orientation, gaze or facial expression) carries the most useful information for the crossing decision. Second, and more fundamentally, an on-board camera can only exploit a face cue if the face covers enough pixels at the moment when the decision is needed. A cue that becomes observable only after the vehicle has passed its stopping distance cannot support a braking decision, however accurate the recogniser is.

The aim of this paper is to answer these questions through a structured review and a simple, reproducible analysis. The contributions are: (i) a taxonomy that places face-based cues among the other cue families used for crossing-intention estimation; (ii) a comparison of public datasets with respect to face-related annotations; (iii) a pinhole-camera analysis of the distance at which each face cue becomes observable, compared with the stopping distance at typical urban speeds; and (iv) a quality-gated fusion framework, together with recommendations for future datasets, evaluation protocols and privacy-preserving deployment.

Methods

Review protocol

A structured narrative review was carried out on publications indexed in Scopus, IEEE Xplore and Google Scholar between 2013 and 2026. The search combined the terms “pedestrian” with “crossing intention”, “action prediction” or “behaviour prediction”, and with “head orientation”, “gaze”, “eye contact”, “facial expression” or “non-verbal communication”. Works were included if they (a) estimated pedestrian crossing intention or action from a vehicle-mounted camera, (b) introduced a dataset with crossing or attention annotations, or (c) proposed head-pose, gaze or expression estimators that are applicable to unconstrained street images. Studies based only on fixed surveillance cameras were included only when their findings concerned human behaviour rather than sensing geometry. Each work was coded by cue family, input representation, dataset and reported real-time considerations. Reference lists of the retrieved surveys [2, 14, 15] were used to find additional sources.

Observability model

To estimate when a face cue is observable, an ideal pinhole camera is assumed. For a sensor of W pixels across and a horizontal field of view θ, the focal length in pixels is

$$f_{\text{px}} = (W/2) / \tan(\theta/2). \quad (1)$$

An object of physical height H at distance Z along the optical axis then projects to

$$h(Z) = f_{\text{px}} \cdot H / Z \quad (2)$$

pixels. If a perception task c requires a face of at least $h_{\min}(c)$ pixels, the maximum distance at which the task can operate is

$$Z_{\max}(c) = f_{\text{px}} \cdot H / h_{\min}(c). \quad (3)$$

The physical face height is taken as $H = 0.20$ m (chin to hairline of an adult). Four illustrative thresholds are used. A minimum of 12 px corresponds to the input size of the first stage of the widely used MTCNN detector [20] and to the small-face regime of the WIDER FACE benchmark [21]. A threshold of 48 px corresponds to the image size of the FER2013 expression

DOI: 10.65322/j.srd.05192026.112

dataset [22], which is among the lowest resolutions at which expression classifiers are routinely trained. Thresholds of 20 px for coarse head orientation (towards or away from the vehicle) and 40 px for gaze or dense landmarks are engineering assumptions consistent with work on eye-contact detection at low resolution [23] and with the input sizes of common facial-behaviour toolkits [24–26]. Motion blur, compression and partial occlusion are ignored, so the resulting distances are optimistic upper bounds.

The reference against which these distances are judged is the stopping distance

$$d_s = v \cdot t_r + v^2 / (2a), \tag{4}$$

where v is the vehicle speed, t_r the total latency from perception to the onset of braking and a the deceleration. For an AV, $t_r = 0.5$ s and $a = 7$ m/s² (emergency braking on a dry road) are assumed; a value of $t_r = 1.0$ s is also reported as a conservative case. Three camera configurations are compared: a 1920-pixel sensor with a 60° horizontal field of view, a 3840-pixel sensor with the same field of view, and a 1920-pixel sensor with a narrower 30° lens.

Table 1

Cue families used for pedestrian crossing-intention estimation

Cue family	**Typical input**	**Representative works**	**Strengths**	**Limitations**
Kinematic / trajectory	Bounding-box sequence, velocity, time to collision, distance to kerb	[4, 5, 12]	Robust at long range; cheap to compute; available for every tracked pedestrian	Reacts late: motion changes only after the decision has been made
Body pose	2D skeleton keypoints, leg and torso orientation	[8, 9, 13]	Captures the onset of walking and body turning before large displacement	Needs a person height of tens of pixels; sensitive to occlusion
Scene context	Crosswalks, signals, road layout, ego-vehicle speed	[5, 7, 10, 11]	Explains why a pedestrian may cross; stable over time	Weak at the level of the individual; scene-specific bias
Face-based	Head orientation, gaze / eye contact, facial expression	[5, 16, 23]	Can signal awareness of the vehicle before any movement	Requires a visible, sufficiently large face; weak link to intention for expression [19]

DOI: 10.65322/j.srd.05192026.112

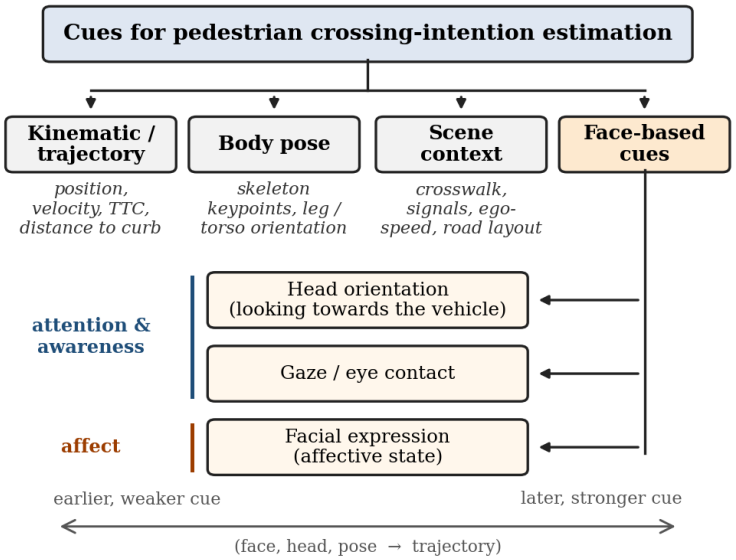

Figure 1. Taxonomy of cues for pedestrian crossing-intention estimation. Face-based cues divide into attention-related cues (head orientation and gaze) and affect-related cues (facial expression).

Results and discussion

Taxonomy of cues. Table 1 and Fig. 1 summarise the four cue families found in the reviewed literature. Kinematic cues remain the backbone of practical systems because they are available for every tracked pedestrian at almost any range. Their weakness is latency: a trajectory reveals the crossing decision only once the pedestrian has started to move. Body pose partially closes this gap by capturing the first step or a turn of the torso [8, 9]. Scene context does not describe the individual but explains why a crossing is likely, for example at a marked crosswalk or a red pedestrian signal [10, 11].

Face-based cues form a fourth family with an important internal distinction. Head orientation and gaze describe *attention*: whether the pedestrian has noticed the vehicle. Facial expression describes *affect*, such as surprise or neutral calm. Attention cues have a clear causal role in the crossing decision, because a pedestrian who has not seen the vehicle cannot negotiate with it; this is why head orientation already appeared as a latent variable in early context-based models [5]. The link between affect and intention is indirect and is mediated by culture, situation and individual differences [19].

Face-based cues in the literature

Head orientation. Among face-based cues, head orientation is the most established. JAAD and PIE annotate whether a pedestrian is looking towards the vehicle [6, 7], and the analysis of JAAD showed that looking frequently precedes crossing [16]. Coarse orientation can be estimated from very few pixels, and fine-grained head-pose regression without landmarks is available when the face is larger [26].

Gaze and eye contact. True gaze estimation requires visible eyes and is mostly developed for close-range settings [25]. For street scenes, Belkada et al. [23] introduced the task of detecting eye contact in the wild and showed that body keypoints can serve as a proxy when the face is too small for crop-based methods. This observation is important for AV perception, because it allows an attention cue to be estimated at distances where the face itself is not resolvable.

Facial expression. Expression recognition has matured on large in-the-wild datasets such as AffectNet [27], and expressions can be described compactly through action units of the Facial Action Coding System [28] or through blendshape coefficients derived from dense face meshes [29, 30]. Its use for crossing intention is still exploratory. No public intention dataset provides

DOI: 10.65322/j.srd.05192026.112

expression labels, so studies must transfer recognisers trained on web images to street imagery, and the psychological evidence cautions against reading intentions directly from facial movements [19]. Expression is therefore best regarded as a complementary, short-range signal (for instance, a startled reaction to an approaching vehicle) rather than as a primary predictor.

Datasets. Table 2 compares the datasets most relevant to face-based intention estimation. JAAD and PIE are the standard benchmarks for crossing prediction and contain attention annotations in the form of looking labels [6, 7, 31]. LOOK provides eye-contact labels for pedestrians in driving scenes but no crossing labels [23]. AffectNet provides expression labels but no traffic context [27]. To the author's knowledge, no public dataset provides crossing labels, attention labels and expression labels for the same pedestrians. This gap makes it impossible to measure the additional value of expression beyond attention, and closing it is a prerequisite for convincing results.

Table 2

Public datasets relevant to face-based crossing-intention estimation

Dataset	**Year**	**Source of images**	**Crossing / intention labels**	**Looking / eye-contact labels**	**Expression labels**
JAAD [6, 31]	2017	Vehicle-mounted camera, urban streets	Yes	Yes (looking)	No
PIE [7]	2019	Vehicle-mounted camera, urban streets	Yes (with intention probability)	Yes (looking)	No
LOOK [23]	2021	Driving and robot datasets in the wild	No	Yes (eye contact)	No
AffectNet [27]	2019	Web images	No	No	Yes

Table 3

Maximum distance (m) at which each face cue is observable, from Eq. (3) with H = 0.20 m

Perception task	**h_{min} (px)**	**1920 px, 60° HFOV**	**3840 px, 60° HFOV**	**1920 px, 30° HFOV**
Face detection	12	27.7	55.4	59.7
Coarse head orientation	20	16.6	33.3	35.8
Gaze / dense landmarks	40	8.3	16.6	17.9
Facial expression	48	6.9	13.9	14.9

Table 4

Stopping distance (m) from Eq. (4) with a = 7 m/s²

Latency t_r	**30 km/h**	**40 km/h**	**50 km/h**	**60 km/h**
0.5 s (automated system)	9.1	14.4	20.7	28.2
1.0 s (conservative)	13.3	19.9	27.7	36.5

DOI: 10.65322/j.srd.05192026.112

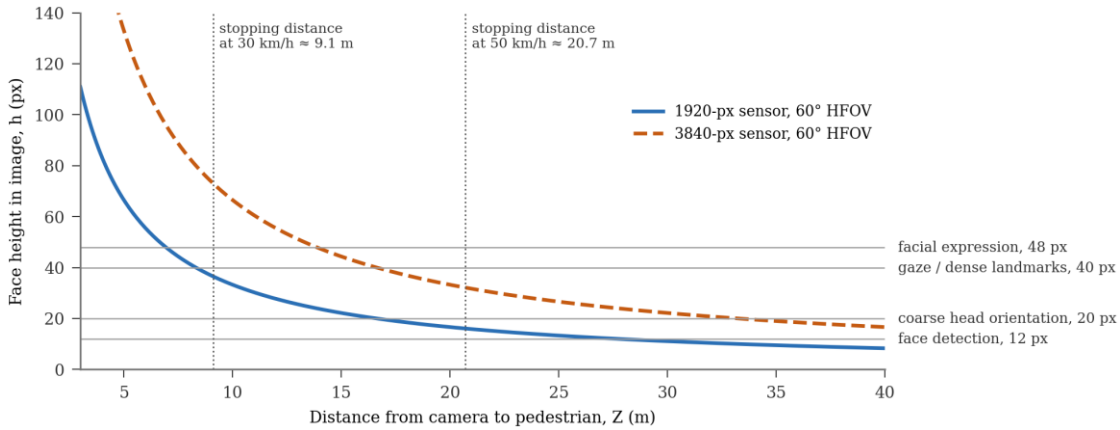

Figure 2. Face height in the image as a function of distance for two sensors with a 60° horizontal field of view. Horizontal lines mark the assumed minimum face sizes for each task; dotted vertical lines mark the stopping distance at 30 and 50 km/h (t_r = 0.5 s, a = 7 m/s²)

Observability of face cues. Tables 3 and 4 and Fig. 2 give the central quantitative result of the review. With a typical 1920-pixel camera and a 60° field of view, a face is detectable up to about 28 m, its coarse orientation can be estimated up to about 17 m, and its expression can be classified only within about 7 m. At 50 km/h the stopping distance is about 21 m even with a fast automated reaction. In this configuration, expression and even head orientation become observable only after the point at which the vehicle could still stop comfortably, and therefore cannot inform the braking decision for an approaching pedestrian.

The picture changes with speed and optics. At 30 km/h, a common limit in residential streets and near schools, the stopping distance falls to about 9 m and coarse head orientation is available well in advance even with a 1920-pixel camera. Doubling the sensor resolution or halving the field of view roughly doubles every range in Table 3: head orientation then remains usable at 50 km/h, and expression becomes observable at about 14–15 m, which is sufficient at 30 km/h but still not at 50 km/h. Because stopping distance is a minimum requirement (comfortable deceleration of 2–3 m/s² requires two to three times more distance), these conclusions are, if anything, optimistic.

Three practical implications follow. First, among face-based cues, head orientation offers the best trade-off between information and range, and should be the first face cue added to an intention estimator. Second, facial expression is relevant mainly in low-speed scenarios, such as approaching a crosswalk, manoeuvring in car parks or interacting with pedestrians who are already close, and for interfaces that communicate the vehicle's intent to pedestrians. Third, face-based modules should be paired with high-resolution or narrow-field cameras, or with region-of-interest crops from a high-resolution stream, rather than run on the same wide-angle image used for general object detection.

Proposed quality-gated fusion framework. The analysis shows that the face branch of an intention estimator will often receive faces that are too small, turned away or occluded. A model that treats face features as always valid will learn from noise in these cases, and its errors will be hard to diagnose. We therefore propose the quality-gated framework shown in Fig. 3. Trajectory, pose and context branches encode their inputs into feature vectors e_k. The face branch detects the face, estimates landmarks and computes head-pose, gaze and action-unit or blendshape descriptors. Before these descriptors enter the fusion stage, a quality gate checks whether the face is usable:

$$g = 1[h \geq h_{\min}] \cdot 1[s \geq \tau] \cdot 1[|\psi| \leq \psi_{\max}], \quad (5)$$

DOI: 10.65322/j.srd.05192026.112

where $1[\cdot]$ is the indicator function, h the face height in pixels, s the detector confidence with threshold τ, and ψ the estimated head yaw with limit ψ_{max}. The fusion weights are computed by masked attention,

$$\alpha_k = m_k \exp(\mathbf{w}^T e_k) / \Sigma_j m_j \exp(\mathbf{w}^T e_j), \quad (6)$$

with $m_k = 1$ for the trajectory, pose and context branches and $m_{face} = g$ for the face branch. The fused representation and the crossing probability are

$$z = \Sigma_k \alpha_k e_k, \quad p = \sigma(w_o^T z + b_o). \quad (7)$$

When the face is not usable, the face branch receives zero weight and the model falls back to the remaining cues without retraining. Because the branches are independent, the face module can be replaced (for example, by a faster detector or by a keypoint-based eye-contact estimator [23]) without changing the rest of the system. To support safe planning, the framework outputs an uncertainty estimate together with the probability, for instance by Monte Carlo dropout [32]:

$$\bar{p} = (1/T) \Sigma_t p_t, \quad u = (1/T) \Sigma_t (p_t - \bar{p})^2, \quad (8)$$

where T stochastic forward passes produce predictions p_t. A planner can then slow down when $\bar{p} + \lambda\sqrt{u}$ exceeds a threshold, so that uncertain cases near the road are handled conservatively.

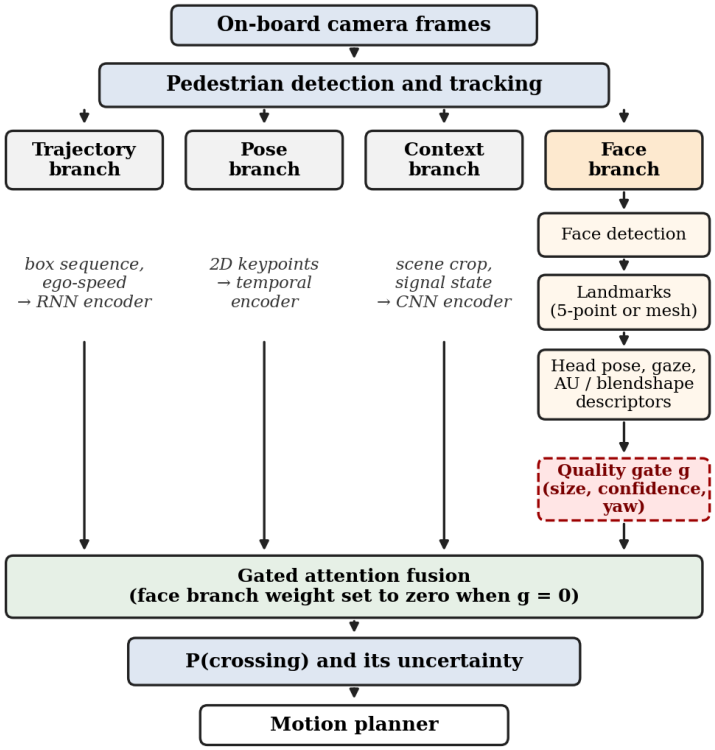

Figure 3. Proposed quality-gated fusion framework. The face branch contributes to the crossing probability only when the quality gate g indicates a usable face; otherwise the estimator relies on trajectory, pose and context

Open challenges

Visibility bias. A face is visible to a forward-facing camera mainly when the pedestrian is facing the vehicle, which is itself a sign of attention. Models may therefore learn that "face detected" predicts crossing, independently of what the face shows. Evaluations should report

DOI: 10.65322/j.srd.05192026.112

whether gains come from face content or merely from face visibility, for example by comparing against a baseline that receives only the gate value *g*.

Occlusion and appearance. Hoods, sunglasses, face masks, umbrellas, night-time illumination and motion blur all reduce the reliability of face cues. These conditions are common in real traffic and are under-represented in expression datasets built from web images [27].

From expression to intention. Facial movements vary across cultures, contexts and individuals and do not map one-to-one onto emotions [19]. Intention models should therefore use expression as a weak, context-dependent signal and should be validated on data from the region where they are deployed, including Central Asian traffic, for which no public intention dataset currently exists.

Evaluation. Crossing datasets are imbalanced, and the reviewed studies use different splits, observation lengths and prediction horizons [11]. Accuracy alone is therefore misleading. Reports should include F1-score, precision and recall for the crossing class, and performance as a function of time to the crossing event and of face size in pixels, so that the range-dependent value of face cues becomes visible.

Privacy. Face analysis in public space raises legitimate privacy concerns. Processing should take place on board, store no face images or identity features, and be limited to the descriptors needed for the safety function, in line with data-minimisation principles such as those of the General Data Protection Regulation [33].

Conclusion. This paper reviewed face-based non-verbal cues for pedestrian crossing-intention estimation and analysed when such cues can actually be observed from an on-board camera. The review shows that kinematic, pose and context cues dominate current estimators, while face-based cues are used rarely and without a common rationale. The observability analysis shows that their usefulness depends strongly on range: with a common 1920-pixel, 60° camera, facial expression is resolvable only within about 7 m and head orientation within about 17 m, both shorter than the stopping distance at 50 km/h. Head orientation is therefore the most practical face cue, and facial expression is mainly relevant at low speeds or with higher-resolution optics.

Based on these findings, the following recommendations are made: (1) add head orientation or eye contact before facial expression when extending an intention estimator with face cues; (2) gate the face branch by face size, detection confidence and head yaw, as in the proposed framework; (3) use high-resolution or narrow-field cameras, or high-resolution regions of interest, for face analysis; (4) build datasets that annotate crossing, attention and expression for the same pedestrians, including data from Central Asian cities; and (5) report results as a function of face size and time to the event. Future work will implement the proposed framework and quantify the contribution of each face cue across distance ranges on such data.

References

1. World Health Organization. *Global Status Report on Road Safety 2023*. – Geneva: World Health Organization, 2023.

2. Rasouli A., Tsotsos J. K. Autonomous vehicles that interact with pedestrians: A survey of theory and practice // *IEEE Transactions on Intelligent Transportation Systems*. – 2020. – Vol. 21, No. 3. – P. 900–918.

3. Badue C., Guidolini R., Carneiro R. V., et al. Self-driving cars: A survey // *Expert Systems with Applications*. – 2021. – Vol. 165. – Art. 113816.

4. Schneider N., Gavrila D. M. Pedestrian path prediction with recursive Bayesian filters: A comparative study // *Proceedings of the German Conference on Pattern Recognition (GCPR)*. – 2013. – P. 174–183.

DOI: 10.65322/j.srd.05192026.112

5. Kooij J. F. P., Schneider N., Flohr F., Gavrila D. M. Context-based pedestrian path prediction // *Proceedings of the European Conference on Computer Vision (ECCV)*. – 2014. – P. 618–633.

6. Rasouli A., Kotseruba I., Tsotsos J. K. Are they going to cross? A benchmark dataset and baseline for pedestrian crosswalk behavior // *Proceedings of the IEEE International Conference on Computer Vision Workshops*. – 2017. – P. 206–213.

7. Rasouli A., Kotseruba I., Kunic T., Tsotsos J. K. PIE: A large-scale dataset and models for pedestrian intention estimation and trajectory prediction // *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. – 2019. – P. 6262–6271.

8. Fang Z., López A. M. Is the pedestrian going to cross? Answering by 2D pose estimation // *Proceedings of the IEEE Intelligent Vehicles Symposium (IV)*. – 2018. – P. 1271–1276.

9. Ghori O., Mackowiak R., Bautista M., et al. Learning to forecast pedestrian intention from pose dynamics // *Proceedings of the IEEE Intelligent Vehicles Symposium (IV)*. – 2018. – P. 1277–1284.

10. Rasouli A., Kotseruba I., Tsotsos J. K. Pedestrian action anticipation using contextual feature fusion in stacked RNNs // *Proceedings of the British Machine Vision Conference (BMVC)*. – 2019.

11. Kotseruba I., Rasouli A., Tsotsos J. K. Benchmark for evaluating pedestrian action prediction // *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. – 2021. – P. 1258–1268.

12. Achaji L., Moreau J., Fouqueray T., Aioun F., Charpillet F. Is attention to bounding boxes all you need for pedestrian action prediction? // *Proceedings of the IEEE Intelligent Vehicles Symposium (IV)*. – 2022. – P. 895–902.

13. Yang D., Zhang H., Yurtsever E., Redmill K. A., Özgüner Ü. Predicting pedestrian crossing intention with feature fusion and spatio-temporal attention // *IEEE Transactions on Intelligent Vehicles*. – 2022. – Vol. 7, No. 2. – P. 221–230.

14. Sharma N., Dhiman C., Indu S. Pedestrian intention prediction for autonomous vehicles: A comprehensive survey // *Neurocomputing*. – 2022. – Vol. 508. – P. 120–152.

15. Ridel D., Rehder E., Lauer M., Stiller C., Wolf D. A literature review on the prediction of pedestrian behavior in urban scenarios // *Proceedings of the IEEE International Conference on Intelligent Transportation Systems (ITSC)*. – 2018. – P. 3105–3112.

16. Rasouli A., Kotseruba I., Tsotsos J. K. Agreeing to cross: How drivers and pedestrians communicate // *Proceedings of the IEEE Intelligent Vehicles Symposium (IV)*. – 2017. – P. 264–269.

17. Schmidt S., Färber B. Pedestrians at the kerb – Recognising the action intentions of humans // *Transportation Research Part F: Traffic Psychology and Behaviour*. – 2009. – Vol. 12, No. 4. – P. 300–310.

18. Dey D., Terken J. Pedestrian interaction with vehicles: Roles of explicit and implicit communication // *Proceedings of the International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI)*. – 2017. – P. 109–113.

19. Barrett L. F., Adolphs R., Marsella S., Martinez A. M., Pollak S. D. Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements // *Psychological Science in the Public Interest*. – 2019. – Vol. 20, No. 1. – P. 1–68.

20. Zhang K., Zhang Z., Li Z., Qiao Y. Joint face detection and alignment using multitask cascaded convolutional networks // *IEEE Signal Processing Letters*. – 2016. – Vol. 23, No. 10. – P. 1499–1503.

DOI: 10.65322/j.srd.05192026.112

21. Yang S., Luo P., Loy C. C., Tang X. WIDER FACE: A face detection benchmark // *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. – 2016. – P. 5525–5533.

22. Goodfellow I. J., Erhan D., Carrier P. L., et al. Challenges in representation learning: A report on three machine learning contests // *Proceedings of the International Conference on Neural Information Processing (ICONIP)*. – 2013. – LNCS 8228. – P. 117–124.

23. Belkada Y., Bertoni L., Caristan R., Mordan T., Alahi A. Do pedestrians pay attention? Eye contact detection in the wild // *arXiv preprint*. – 2021. – arXiv:2112.04212.

24. Baltrušaitis T., Zadeh A., Lim Y. C., Morency L.-P. OpenFace 2.0: Facial behavior analysis toolkit // *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition (FG)*. – 2018. – P. 59–66.

25. Zhang X., Sugano Y., Fritz M., Bulling A. Appearance-based gaze estimation in the wild // *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. – 2015. – P. 4511–4520.

26. Ruiz N., Chong E., Rehg J. M. Fine-grained head pose estimation without keypoints // *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. – 2018. – P. 2074–2083.

27. Mollahosseini A., Hasani B., Mahoor M. H. AffectNet: A database for facial expression, valence, and arousal computing in the wild // *IEEE Transactions on Affective Computing*. – 2019. – Vol. 10, No. 1. – P. 18–31.

28. Ekman P., Friesen W. V. *Facial Action Coding System*. – Palo Alto: Consulting Psychologists Press, 1978.

29. Kartynnik Y., Ablavatski A., Grishchenko I., Grundmann M. Real-time facial surface geometry from monocular video on mobile GPUs // *arXiv preprint*. – 2019. – arXiv:1907.06724.

30. Lugaresi C., Tang J., Nash H., et al. MediaPipe: A framework for building perception pipelines // *arXiv preprint*. – 2019. – arXiv:1906.08172.

31. Kotseruba I., Rasouli A., Tsotsos J. K. Joint attention in autonomous driving (JAAD) // *arXiv preprint*. – 2016. – arXiv:1609.04741.

32. Gal Y., Ghahramani Z. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning // *Proceedings of the International Conference on Machine Learning (ICML)*. – 2016. – P. 1050–1059.

33. European Parliament and Council of the European Union. Regulation (EU) 2016/679 (General Data Protection Regulation) // *Official Journal of the European Union*. – 2016. – L 119. – P. 1–88.