Computational Social Vision for Human-Agent Interaction

Hui Yu, Yiming Wang, Haodong Jin, and Schahram Dustdar

***Abstract*—Human-agent interaction (HAI) is an increasingly important research area that investigates how humans interact with intelligent agents. Effective HAI relies heavily on social signal understanding, particularly visual social signals that convey human attention, intention, and affect, enabling agents to perceive, interpret, and respond appropriately to human behaviour. Recent advances in artificial intelligence (AI) have substantially improved visual perception and representation learning, allowing machines to detect and model visual social behavioural cues with increasing accuracy. However, translating these cues into socially meaningful interpretations remains a significant challenge, often resulting in unnatural interactions and limited empathic responsiveness in HAI systems. In this paper, we coined the concept of computational social vision (CSV) spanning the domain of both psychology and computing science to understand social signals. We summarize visual-cognitive computing techniques for this purpose with an emphasis on how social vision and artificial intelligence technologies support the inference of latent social states from dynamic behavioural evidence. To structure the field, we organise representative methods according to the semantic roles of visual behaviours, including both head-centric and body-centric social vision approaches. Finally, we discuss socially interactive embodied systems, analyzing how agents, such as social robots and virtual humans, adaptively perceive their environments and interact with humans as well as the future trend.**

***Index Terms*—Computational social vision, human-agent interaction, social signal understanding, social vision, visual-cognitive computing, social robots, virtual humans**

I. Introduction

HUMAN-agent interaction (HAI) is a multidisciplinary research field concerned with how humans perceive, communicate with, and collaborate with intelligent agents. A central requirement for natural HAI is that agents should be able to interpret human social behaviours in a manner that is compatible with everyday social interaction. During face-to-face interaction, individuals continuously exchange nonverbal and paralinguistic cues [1–3], including facial expressions [4, 5], gaze directions [6, 7], body posture [8, 9], and gestures [10, 11]. Together with audio and contextual information, these cues constitute social signals through which interlocutors convey latent states such as emotions [12], intentions [13], and social relationships [14]. Since such states are not directly observable as discrete class labels, social signal understanding refers to the inference of underlying social meanings from complex behavioural evidence [1]. Among the channels through which social signals are conveyed, vision is particularly important because many cues related to attention, affect, engagement, and interpersonal orientation are visually grounded. The ability to interpret visual social signals therefore underpins a broad range of HAI applications, including social robotics, assistive healthcare, and mental-health support.

Recent advances in computer vision [15–18] have substantially improved the automatic detection and representation of visual behavioural cues. However, progress in visual perception has not yet resulted in comparable progress in social understanding. Social meaning is not determined by visual appearance alone, but emerges from the interaction between observable behaviour and its scene, cultural, relational, conversational, and historical context. The same behavioural pattern may therefore indicate different social meanings under different interactional conditions. This makes social signal understanding fundamentally different from conventional visual recognition, where the objective is often to assign an input to a predefined visual category. For HAI, the key challenge is not only to identify what is visible, but also to infer what the visible behaviour implies for the ongoing interaction. Addressing this challenge requires models that go beyond representation learning and incorporate interpretive principles grounded in social psychology, behavioural science, and cognitive accounts of human perception.

In psychology and cognitive science, the study of how humans derive socially relevant information from visual input is known as social vision [19]. This perspective treats visual behavioural evidence not merely as patterns to be classified, but as information from which observers infer mental states, social relations, and interactional intentions. Social vision therefore provides a theoretical foundation for the interpretive layer that current computational models often lack. At the same time, existing social-vision research has largely emphasized behavioural, perceptual, and cognitive explanations, and has not yet provided a systematic computational pathway for embedding such principles in artificial agents.

Manuscript received: 13 May 2026; revised: 19 May 2026; accepted: 20 May 2026. *(Corresponding author: Hui Yu.)*

Citation: H. Yu, Y. Wang, H. Jin, and S. Dustdar, Computational social vision for human-agent interaction, *Int. J. Intell. Control Syst.*, 2026, 31(2), 193–204.

Hui Yu and Yiming Wang are with School of Psychology and Neuroscience, University of Glasgow, Glasgow G12 8QB, UK (e-mail: hui.yu@glasgow.ac.uk; yiming.wang@glasgow.ac.uk).

Haodong Jin is with Department of Control Science and Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China (e-mail: 231260086@st.usst.edu.cn).

Schahram Dustdar is with Research Division of Distributed Systems, Vienna University of Technology, Vienna 1040, Austria (e-mail: dustdar@dsg.tuwien.ac.at).

Digital Object Identifier 10.62678/IJICS202606.10373

To address this need, interdisciplinary knowledge and methodologies that bridge low-level visual perception with high-level social inference are required, allowing agents to automatically perceive interpersonal cues, relationships, and intentions from visual input [20]. In this paper, we coined this as computational social vision (CSV). This can then empower agents to engage in socially aware, natural, and adaptive interactions with humans. The high-level social interface here means the layer where raw pixels are translated into meaningful social knowledge that a computer or agent can use to interact appropriately with humans. To clarify, social signal understanding denotes the target problem of inferring latent social meanings from behavioural evidence. Social vision provides the psychological and cognitive foundation for understanding how humans visually perceive socially relevant information, such as faces, gaze, posture, gesture, motion, and interpersonal configurations. Visual-cognitive computing denotes the computational pathway that transforms visual input into structured and cognitively meaningful representations. Computational social vision unifies these perspectives by leveraging visual-cognitive computing to model socially observable behaviors and enable socially meaningful inference in human-agent interaction. Therefore, CSV is positioned as a bridge between low-level visual cue analysis and high-level social understanding, rather than as a simple extension of conventional visual recognition.

Based on this conceptual scope, this paper organizes representative research in connection with CSV studies into three categories: head-centric method, body-centric method, and agent-centric method. Head-centric methods focus on cues that are primarily observed from the head and face, including head pose, gaze direction, eye behaviour, facial movements, and expression. These cues are dense, high-frequency, and closely related to attention, emotion, affect, and interpersonal orientation. Body-centric methods focus on posture, motion, gesture, and sign, which often express intention, engagement, activity, and interaction structure over longer temporal intervals. Agent-centric methods further examine how visual-cognitive pipelines are integrated into social robots and virtual humans, where perception, social inference, decision-making, and response generation must be coordinated in closed-loop interaction. This organization allows us to compare different cue types, modelling assumptions, and deployment contexts within a unified CSV framework.

In summary, this paper discusses research advances in visual-cognitive computing for social signal understanding aligning with the concept of computational social vision. We develop a head-centric and body-centric technical map to support structured comparison across cues, tasks, and modelling paradigms, and we further discuss intelligent agents as a representative deployment context that connects visual perception, inference, and interaction. Compared with existing surveys, our main contributions are as follows:

(1) We present an analysis of methods closely related to computational social vision for social signal understanding. In contrast to providing a general overview of visual perception techniques, we focus specifically on CSV approaches that examine how visual signals cognitively influence social interactions.

(2) We summarize the field of CSV into head-centric method, body-centric method, and agent-centric method, and discuss how different branches model and understand social behaviors.

(3) We outline open challenges and promising directions for future research on CSV, with an emphasis on evaluation metrics.

II. Preliminaries: Social Vision

Social vision provides the psychological and cognitive foundation for understanding how humans derive social meaning from visual information. Rather than treating vision as a passive process of recognizing objects, social vision emphasizes how observers infer others' attention, affect, intention, disposition, and interaction tendencies from visually available cues [19, 21]. In this view, faces, eyes, bodies, motion, interpersonal distance, and scene context are not merely visual patterns. They are socially meaningful signals that support adaptive perception and behaviour. A central idea in social vision is that visual perception is functionally organized for social interaction. The social-functional account suggests that observers selectively attend to adaptively relevant cues and use them to make behavioural forecasts about others [22, 23]. This perspective is closely related to ecological theories of perception, where visual cues are understood through the actions and responses they afford. Therefore, the meaning of a facial expression, gaze direction, or body posture cannot be fully determined by its local appearance alone. It depends on what the cue implies for the observer in a given social context.

Social vision is also inherently integrative. Earlier studies often examined facial identity, expression, gaze, gender, race, body language, and biological motion as separate sources of information. However, social perception usually depends on compound cues that are combined rapidly and functionally [24]. For example, gaze direction and facial expression jointly indicate the source and target of another person's intention, while faces and bodies can convey congruent or conflicting affective meanings [25]. Biological motion further shows that dynamic body movement is shaped by social interpretation rather than being perceived as purely geometric motion [26]. Developmental and neural evidence also suggests that face processing and gaze processing are specialized through the interaction between early predispositions and social experience [27]. These findings suggest that social vision concerns not only what is visually detected, but also how multiple visual cues are organized into socially meaningful interpretations.

At the same time, social vision is not universal or context-free. Facial expressions and other social cues are shaped by culture, observer experience, and communication norms. Cross-cultural evidence shows that mental representations of basic facial expressions can differ between Western and East Asian observers, challenging the assumption that emotional expressions are represented by identical facial movement patterns across cultures [28]. Perception-driven facial expression synthesis further demonstrates that computational models of facial expression can be constructed from human perceptual judgements, linking visual synthesis with observer-dependent social meaning [20]. These findings indicate that social

vision involves both visually observable signals and the perceiver's learned social-cognitive interpretation.

Based on these foundations, we define computational social vision as a computational approach rooted in social vision that connects low-level visual perception to high-level social inference. Specifically, CSV aims to enable agents to automatically perceive interpersonal cues, relationships, and intentions from visual input, thereby supporting socially aware, natural, and adaptive interactions with humans. This definition distinguishes CSV from conventional computer vision: the goal is not only to detect faces, poses, actions, or scenes, but also to transform such visual evidence into social understanding. Recent benchmarking work on dynamic social vision shows that current image, video, and language models still struggle to match human behavioural and neural responses to complex social interactions [29]. This gap further motivates CSV as a visual-cognitive computing paradigm for human-agent interaction.

III. Head-Centric Computational Social Vision

Head-centric cues are among the most frequently observed visual signals in social interaction. They are closely associated with attention allocation, interaction intent, affective expression, and interpersonal orientation. This section reviews three core cue types, namely head pose, gaze direction and eye behaviour, and facial movement and expression. For each cue type, we summarize representative computational methods and discuss how they contribute to computational social vision.

A. Head Pose

Head pose [30, 31] is commonly represented by three Euler angles, yaw, pitch, and roll. It characterizes head orientation in three-dimensional (3D) space and relative alignment, which often reflects attention direction and interaction intent. In intelligent driving scenarios, head pose is frequently used for driver monitoring to improve road safety. Bisogni et al. [32] present HYDE-F, a head pose estimation framework that relies on a depth camera only for Internet of Things enabled advanced driver assistance systems. This method integrates fractal encoding with keypoint intensity analysis through an optimization based fusion scheme, and it reports improved mean absolute error with competitive cross dataset generalization. Ye et al. [33] introduce a vision-based driver fatigue detection framework that combines facial state recognition with 3D head pose estimation to handle large head deflections. Their pipeline uses RetinaFace to localize the face and extract five landmarks, then applies a residual channel attention network to classify eye and mouth states. Head pose is computed from the five landmarks, and an over angle criterion is used to flag excessive head deviation. These cues are further integrated into fatigue related decision metrics. Zhao et al. [34] design a deep-learning head pose estimator without explicit preprocessing, use it to characterize pose separability between safe and distracted driving on a virtual 3D driver head dataset for distracted driver detection, and extend the analysis to continuous real-driving videos to estimate viewpoint-dependent safety thresholds.

Beyond individual monitoring, head pose at the group level can also reveal collective social signals, especially in smart classroom analytics. Liu et al. [35] propose asymmetric relation-aware head pose estimation (ARHPE) to address noisy or incorrect labels by replacing hard angle targets with a two-dimensional (2D) Lorentz distribution based soft label. They further show that estimated student head orientations can support attention or interest inference in classroom videos, and they provide qualitative evidence of robustness in challenging cases through stable head orientation localization and visualization. Dai et al. [36] investigate classroom video analytics for education by developing an end-to-end, landmark free head pose estimator that predicts Euler angles using a multi-loss network that combines classification and regression objectives. Building on this estimator, they design an attention and status assessment scheme that maps head angle dynamics, such as state change frequency and amplitude, to student attention. Their results suggest that diversified criteria outperform single criterion baselines in classroom attention assessment.

B. Gaze Direction and Eye Behaviour

During social communication, gaze often reflects objects of interest of a person [37]. Gaze direction and eye behaviour [38–40] therefore provides a finer grained attention cue than head orientation. The psychological research has shown that gaze and eye movement, as important social signals, interact with emotion perception in face processing [41–43]. Consequently, these findings from psychology have driven the development of visual computing approaches for gaze and eye movement detection.

Senarath et al. [44] design a three-stage convolutional model with three attention mechanisms for remote gaze estimation in retail environments. The model first infers 3D gaze from monocular images with depth estimation, then refines gaze target prediction using field of view, depth range, and object channel attentions. It finally produces a gaze saliency heatmap. They also introduce Retail Gaze to support adaptation to real store settings and report consistent gains over baselines. Yang et al. [45] propose vGaze, a mobile continuous gaze tracking system that performs implicit calibration with saliency awareness, without requiring explicit user cooperation. Modi and Singh [46] develop a low-cost webcam-based gaze tracking framework using a convolutional model to enable real time gaze point recording under unconstrained head motion, and they demonstrate utility through a social media advertising case study and a gaze driven interactive interface for virtual tasks.

In virtual reality (VR) or augmented reality (AR) scenarios, gaze has become a key signal for capturing user intent and internal decision processes. David-John et al. [47] study intent to interact prediction in virtual reality using gaze dynamics, with the goal of reducing reliance on environment specific gaze mapping and strict calibration. Using eye tracking data from 15 participants performing an item selection task, they train a logistic regression model to predict interaction onset. The most informative cues include gaze velocity, ambient versus focal attention, and saccade dynamics, which suggests

that pre-interaction eye movement patterns can enable low friction predictive interfaces. Xu et al. [48] exploit common mode features suppression network (CMFS-Net) for robust appearance-based gaze point estimation in extended reality (XR), aiming to reduce cross subject variance. The framework fuses eye and face regions, estimates gaze bias between an input image and a template image using a differential amplifier inspired design, and aggregates candidate predictions with a point mean filtering strategy. Chen et al. [49] present EX-Gaze, an on-device eye tracking system based on event cameras for extended reality. It targets kilohertz level tracking with low latency on embedded hardware by introducing a sparse event patch representation and a corresponding sparse transformer to reduce computation.

C. Facial Movement and Expression

Facial movements and expressions [50, 51] provide a direct channel for affective communication and convey rich social signals. Classic findings [52] associate specific facial muscle activation patterns with widely studied categories such as happiness, anger, sadness, surprise, fear, and disgust. Modern emotion theories advise that emotional facial expressions can be personalised, different groups of people (e.g., different ages, different cultural backgrounds) may perceive expressions differently [53].

Recent research has revisited facial expression recognition with stronger representation learning and multimodal modeling. Zhou et al. [54] propose CEPrompt, a cross-modal emotion aware prompting framework that adapts vision language pre-trained models to facial expression recognition by injecting emotion semantics from text into fine grained facial appearance modeling. The method introduces an emotion conception guided visual adapter to increase sensitivity to subtle expression relevant cues, and it uses knowledge distillation to preserve category invariant representations learned by the backbone. Tsai et al. [55] develop a behavioral signal-based pain level rating approach for emergency triage that models facial expressions together with vocal characteristics from audio-video recordings, motivated by the limited reliability of self-reported numerical scales in clinical practice. Xia et al. [56] present relation convolutional neural network (ReCNN), a relation aware model for in-the-wild facial expression recognition. It explicitly models relationships between key facial regions and expression categories, rather than treating the whole face as a uniform feature source. This design helps the model focus on discriminative regions under pose variation, occlusion, and illumination changes.

In cognitive science, the human face is regarded not merely as a static object, but as a high-dimensional and dynamic communication system that transmits socially relevant information over time. Jack and Schyns [57] argue that facial communication involves both information transmission and decoding: a sender encodes social information through facial morphology, complexion, and dynamics, while a receiver interprets such information based on perceptual knowledge and cultural context. This view indicates that facial expressions should be understood as temporally unfolding signals rather than fixed visual templates. From a computer science perspective, dynamic 3D facial expressions provide richer spatiotemporal evidence for understanding affective states. Tie and Guan [58] exploit an automatic emotion recognition framework based on dynamic 3D facial expressions. Their method tracks 26 facial fiducial points using multiple particle filters, drives a generic mesh via elastic body spline warping, and extracts spatiotemporal deformation features. The descriptors are then embedded into a low dimensional manifold using discriminative isometric mapping (Isomap) and classified with the nearest class center rule. Beyond recognition, Jack et al. [59] further show that dynamic facial expressions transmit an evolving hierarchy of signals over time. Using generative face grammar, reverse correlation, information theory, and Bayesian classifiers, they model how observers expect facial action units to unfold during emotional communication. Their results suggest that early facial movements tend to convey simpler and biologically rooted information, such as broad danger-related or approach-avoidance signals, whereas later movements provide more complex and socially specific cues that support finer discrimination among classic emotion categories. This temporal hierarchy highlights why dynamic facial modeling is important for computational social vision: social meaning is not only encoded in facial appearance, but also in the order, timing, and evolution of visible facial movements.

Table 1 summarizes representative head-centric CSV methods according to the subsection structure of this survey. Head pose methods provide relatively robust but coarse evidence for orientation and attention. Gaze-based methods offer finer-grained cues for visual attention and interaction intent, especially in mobile, XR, and multi-person scenarios. Facial movement and expression methods further support affective and interpersonal inference, but their interpretation depends strongly on temporal dynamics, context, and cultural variability. These comparisons show that head-centric CSV should move beyond isolated cue detection and focus on transforming visual cues into socially meaningful inferences for HAI.

IV. Body-Centric Computational Social Vision

Compared with head-centric cues, the human body serves as a primary substrate for social behaviour, where multiple parts often coordinate in a structured manner, leading to richer but more complex visual evidence. Body postures, gestures, and motion patterns are also more strongly shaped by situational context and cultural conventions, and thus may convey culturally specific social meanings. Accordingly, this section reviews visual computing of full-body pose and motion, and further discusses hand behaviours as one of the most social expressive sources of nonverbal communication.

A. Body Pose and Motion

Vision based modeling of body pose and action estimation [60, 61] typically follows a pipeline that detects keypoints or body meshes and then learns spatiotemporal representations for recognition or inference. These cues can support predictions of affect, intention, and interaction dynamics, and they provide a computational basis for describing body language. Müller et al. [62] introduce MPIIEmo, a multimodal dataset for emotion recognition from bodily expressions and speech in

Table 1 Representative head-centric CSV methods.

Category	Representative work	Social cue or target	Critical idea
Head pose	Ye et al. [33]	Driver fatigue and head deviation	Combines facial state recognition with 3D head pose estimation for fatigue-related decision metrics.
	Zhao et al. [34]	Distracted driving	Uses deep head pose estimation to characterize pose differences between safe and distracted driving.
	Bisogni et al. [32]	Head orientation and attention	Performs depth-based head pose estimation with fractal encoding and keypoint intensity fusion.
Gaze direction and eye behaviour	Massé et al. [37]	Visual focus of attention	Tracks gaze and visual focus of people involved in social interaction.
	David-John et al. [47]	Intent to interact in VR	Uses gaze dynamics to predict interaction onset in virtual environments.
	Yang et al. [45]	Mobile visual attention	Introduces saliency-aware implicit calibration for continuous mobile gaze tracking.
	Chen et al. [49]	Low-latency XR gaze	Combines event-frame sensing and sparse transformer modelling for high-frequency gaze tracking.
Facial movement and expression	Tie and Guan [58]	Dynamic emotional state	Extracts spatiotemporal deformation features from dynamic 3D facial expressions.
	Jack et al. [28]	Cultural emotion perception	Models culture-specific facial expression representations using dynamic 3D facial movements.
	Xia et al. [56]	In-the-wild expression recognition	Models relationships between facial regions and expression categories.
	Zhou et al. [54]	Facial emotion recognition	Injects textual emotion semantics into visual prompt learning for expression recognition.

naturalistic dyadic interactions. These data are captured unobtrusively with multi-view cameras and microphones in a real kitchen environment. The dataset contains 224 synchronized audio-video clips from 8 actor couples and provides both categorical emotion labels and continuous affect annotations. Romeo et al. [63] study apparent personality prediction for adaptive social human-robot interaction (HRI) by modeling body language cues from short side view dyadic interaction videos. Their work targets the limitation of face dominated designs and highlights the value of body signals when facial evidence is weak or unavailable. Recent work has also explored generative modeling for socially meaningful pose behavior. Martin-Ozimek et al. [64] adapt diffusion-based behavior cloning to synthesize facilitator pose behavior in multi-party social interactions. Their study tests whether pose-only cues are sufficient to generate realistic nonverbal signals for socially assistive agents. They compare two conditioning schemes. One uses raw 360° scene images, and the other uses pose keypoint renderings. They evaluate the trade-off between accuracy and efficiency using mean per-joint position error (MPJPE), training cost, and inference latency. Their results indicate that pose rendered conditioning can guide diffusion models toward plausible social motion while improving computational characteristics. Body pose can further serve as a strong prior for attention and activity inference in complex scenes. Guan and Lau [65] propose PoseSOR, a pose aware framework for salient object ranking. The method models attention shifts beyond low level and semantic saliency by treating body pose as an indicator of ongoing activity and human object interaction. A pose aware interaction module is introduced that injects pose knowledge into salient object queries to learn joint level and instance level interactions. They also propose a pose driven ranking module that uses pose derived directional cues to predict the next attended object. For fine grained social analysis, Balazia et al. [66] introduce BBSI, an annotation set for subtle bodily behaviors embedded in continuous group interactions. Motivated by psychology, they annotate 26 h of spontaneous interaction from MPIIGroupInteraction with 15 body language classes. They also benchmark temporal action detection by adapting pyramid dilated attention network (PDAN) with multiple spatiotemporal backbones.

Recently, 3D human pose estimation has attracted growing attention because it can represent articulated motion with richer geometric constraints and improved viewpoint robustness. Kratzer et al. [67] propose an affordance-conditioned framework for full-body motion prediction in grasping and placing, indicating effective intention anticipation from learned affordances. This framework learns graspability and placeability as probabilistic densities conditioned on 3D scene geometry and human state, combines these with an RNN-based short-term dynamics model, and generates long-horizon predictions via constraint-based trajectory optimization under obstacle and goal-set constraints. Peng et al. [68] propose kinematics and trajectory prior knowledge-enhanced transformer (KTPFormer) for 3D human pose estimation. The model injects kinematics and motion trajectory priors into self-attention to address the limitations of purely linear query Q, key K, and value V projections. It introduces kinematics prior attention to encoding anatomical connectivity through a learnable kinematic topology, and trajectory prior attention to model cross frame joint motion with a learnable trajectory topology. This design supports joint spatial and temporal correlation modeling within a unified transformer. Feng et al. [69] introduce ChatPose, a multimodal large language model (LLM) that outputs skinned multi-person linear (SMPL) parameters through a dedicated pose token and a specialized projection layer that maps language embeddings to 3D pose parameters. The framework supports both text conditioned and image conditioned pose reasoning, and it enables interactive pose specification through dialogue. Despite these advances, the use of 3D pose for social signal understanding remains relatively underexplored, leaving

substantial room for deeper modeling of interaction structure and social meaning.

B. Hand Gesture and Sign

Hand gesture and sign [70, 71] often convey concise and highly discriminative social meanings, which are one of the most essential central to human-robot interaction. Peral et al. [72] present gesture recognition key frame (EUREKA), an efficient vision-based gesture recognition pipeline for human-robot interaction that avoids intrusive sensors. The framework first detects 2D hand landmarks with MediaPipe, then performs gesture classification from a temporal representation of these landmarks using a densely connected network. Hu et al. [73] propose SignBERT, a self-supervised hand model aware pretraining framework for isolated sign language recognition. The method treats hand pose sequences extracted by an off-the-shelf estimator as visual tokens and enriches them with gesture state, temporal position, and hand chirality embeddings. It is pretrained with a mask and reconstructed objective over hand tokens using multiple masking strategies to learn temporal context at different scales. A hand model aware decoder further injects hand priors to stabilize contextual modeling. Zuo et al. [74] propose NLA-SLR, a natural language assisted framework that leverages gloss semantics to reduce confusion among visually similar signs. They introduce language aware label smoothing and inter modality mixup to improve discriminability. Their video keypoint network combines RGB and keypoints with multi-scale temporal modeling and achieves strong results on common benchmarks.

In industrial robotics, hand gestures provide an intuitive interface for issuing control commands and enabling low-friction human supervision. Hubert et al. [75] investigate vision-based hand gesture recognition for human-robot interaction in cobotic platforms, specifically addressing challenges like visual occlusions and varying camera viewpoints in real-world industrial finishing tasks. Sign language functions as a speech-free communication system with a rich linguistic structure. Its forms and conventions vary across cultures and languages, which motivates dedicated datasets and evaluation protocols. Rao et al. [76] study Indian sign language recognition from mobile front camera selfie mode videos, targeting scenarios where deaf or hard-of-hearing users can operate a recognition system independently. They construct a dataset of 200 common signs performed by 5 subjects and captured from multiple viewing angles under varied backgrounds. They evaluate several convolutional neural network architectures under different training scales and test protocols, including robustness under seen and unseen conditions. Their results demonstrate strong recognition accuracy on the constructed benchmark and support the feasibility of mobile sign recognition under realistic capture settings. Yang et al. [77] revisit the dynamic time warping (DTW)-based level building framework, which can jointly segment and recognize Chinese hand signs. They replace DTW with a hidden Markov model (HMM)-based similarity function, motivated by the ability of HMMs to better characterize sign dynamics. They also propose a fast-HMM strategy that precomputes frame-prefix decoding probabilities using the Viterbi algorithm once per sign model. This enables efficient scoring of candidate segments during level building through lightweight arithmetic. Recently, Hu et al. [78] propose correlation network (CorrNet), a hybrid modeling framework for continuous sign language recognition (CSLR) that explicitly captures cross-frame body trajectories from both manual cues (hand and arm) and non-manual cues (face and head). In contrast to pipelines that extract frame-wise features independently using a shared two-dimensional convolutional neural network, CorrNet introduces a correlation module that computes correlation maps between the current frame and its neighboring frames to model patch-level motion and trajectory patterns.

Table 2 compares representative body-centric CSV methods following the subsection structure of body pose, motion, hand gesture, and sign. Compared with head-centric cues, body-centric cues provide broader interaction-level evidence. Body posture and motion are useful for modelling affect, personality, attention, and group dynamics, especially when facial cues are unavailable or unreliable. Hand gestures and sign language provide more discrete and semantically structured signals, but their meanings are shaped by task setting, language, and cultural convention. Therefore, body-centric

Table 2 Representative body-centric CSV methods.

Category	Representative work	Social cue or target	Critical idea
Body pose and motion	Müller et al. [62]	Emotion from body and speech	Provides multimodal dyadic interaction data for recognizing affect from bodily expressions and speech.
	Romeo et al. [63]	Apparent personality in HRI	Uses body language cues to support adaptive social human-robot interaction.
	Balazia et al. [66]	Group body language	Provides annotations and benchmarks for subtle bodily behaviours in continuous group interaction.
	Guan and Lau [65]	Human attention and saliency	Uses body pose as a prior to model attention shifts and human-object interaction.
	Martin-Ozimek et al. [64]	Facilitator pose behaviour	Applies diffusion-based behaviour cloning to synthesize nonverbal pose behaviour in multi-party interaction.
Hand gesture and sign	Hu et al. [73]	Isolated sign recognition	Introduces hand-model-aware self-supervised pretraining for sign language representation.
	Peral et al. [72]	Gesture-based HRI	Uses hand landmarks and temporal modelling for efficient vision-based gesture recognition.
	Zuo et al. [74]	Semantically assisted sign recognition	Uses gloss semantics, language-aware label smoothing, and inter-modality mixup.
	Hu et al. [78]	Continuous sign recognition	Captures cross-frame trajectories from manual and non-manual sign cues.
	Hubert et al. [75]	Industrial HRI gesture	Studies robust gesture recognition under occlusion and viewpoint variation in cobotic platforms.

CSV requires temporal modelling, contextual grounding, and culturally sensitive interpretation rather than isolated pose or gesture classification.

V. Agent-Centric Computational Social Vision

Agent-centric artificial intelligence (AI) has been explored in interactive and decision-oriented systems, including computing continuum environments [79, 80], recommender systems [81, 82], and AI assistants [83, 84]. In this section, we focus on its social-vision form, where embodied or virtual agents perceive human behaviours, infer social states, and generate appropriate responses. Unlike head-centric and body-centric methods that mainly analyze human cues from an observer's perspective, agent-centric CSV emphasizes closed-loop perception-action coupling. Agents extract visual and multimodal cues, interpret attention, affect, intention, or engagement, and adapt their behaviours through human feedback. This section reviews two representative settings: social robots in physical environments and virtual social embodiment in immersive or digital environments.

A. Social Robots

Social robots are a typical example of agent-centric systems. They represent a system-level application domain in which visual-cognitive computing is integrated into closed-loop perception, social inference, and adaptive response. As a core component of social vision, robotic vision provides the perceptual foundation for such closed-loop interactions. Robinson et al. [85] review robotic vision for human-robot interaction and collaboration, showing that visual information can support action and gesture recognition, navigation in human spaces, object handover, social communication, and learning from demonstration. Unlike general computer vision, robotic vision directly informs robot behaviour and interaction control. In human social environments, this requirement becomes more complex because social robots must process heterogeneous and asynchronous cues in real time. Duncan et al. [86] summarize multimodal perception systems that combine pose, gaze, gesture, speech, and contextual information to infer human communication, status, and intent. From a complementary communication perspective, Urakami and Seaborn [87] further emphasize that nonverbal cues should be translated into robot design patterns, because they shape the perceived naturalness, accessibility, and social agency of robots. These studies establish social robots as an application setting where visual cues must be interpreted together with multimodal context and embodied response.

Recent studies further highlight the importance of robot-centred perception. Zhong et al. [88] improve the visual perception of a Pepper robot by integrating human pose estimation and tracking models and evaluate the system in both controlled and in-the-wild interaction scenarios. Their results indicate that improved perception can support target-user following and proactive robot behaviours, but interaction quality still depends on embodied behaviour and user experience. Avogaro et al. [89] propose HARPER, a dataset for 3D human pose estimation and forecasting from a quadruped robot's perspective. By emphasizing partial visibility, unusual viewpoints, camera motion, physical contact, and collision prediction, HARPER shows that social robot perception differs fundamentally from conventional third-person visual analysis.

Beyond perceiving users, social robots must anticipate social intent and adapt their behaviours accordingly. Bian et al. [90] formulate this problem as joint egocentric forecasting of intent to interact, attitude toward the agent, and upcoming social actions, using whole-body skeletons to support proactive interaction preparation. To address deployment constraints, Bian et al. [91] propose a multimodal distillation framework in which a rich teacher model learns from body pose, face, hand gestures, gaze, and raw images, while a lightweight student model performs robust inference from corrupted body pose. Finally, social robots must also handle group interaction. Nigro et al. [92] identify key computational challenges in social group HRI, including group detection, engagement estimation, conversation information, approaching behaviours, and conversational behaviours. Song et al. [93] use a vision-language model (VLM) based navigation approach to perceive human social signals and take actions that are socially compliant with human expectations. Overall, social robotics connects visual perception, multimodal social inference, action planning, and adaptive response, making it an important application domain for evaluating visual-cognitive social signal understanding under embodied and real-world constraints.

B. Virtual Social Embodiment

Unlike social robots, which act as real-world social entities, virtual or mixed-reality agents operate through social perception and intelligent interaction within immersive digital environments. Immersive VR environment provides a platform for humans to communicate and interact with virtual agents. As a result, virtual social embodiment is becoming increasingly important in virtual reality, augmented reality, metaverse platforms, embodied conversational agents, and socially interactive simulation environments [94, 95].

In virtual environments, users are expected to interact with highly realistic virtual humans, including both user-controlled avatars and computer-controlled agents. For user-controlled avatars, existing techniques primarily focus on transferring social behaviours from real-world users to their virtual representations. In Ref. [96], a personalised facial expression reconstruction method is proposed to enable users to display realistic facial expressions in VR environments. Similar studies on head avatar creation and facial expression transfer can be found in Refs. [97, 98]. Beyond facial animation, related research has also explored avatar eye-gaze transfer [47, 48, 99, 100] and body gesture transfer [101, 102].

For computer-controlled agents, the core technologies are generative AI approaches that enable artificial agents to produce realistic social behaviours. Owing to the strong performance of modern generative language models, many existing studies focus on conversational agent generation [103, 104]. Additional work has investigated body gesture generation [105, 106] and facial expression generation [51, 100, 107, 108]. Recent approaches increasingly utilise conversational context to guide the generation of socially meaning-

ful gestures and facial expressions. For example, in Ref. [109], a large language model-driven body gesture generation framework is proposed, where LLM-generated speech is used to guide virtual humans in producing corresponding body gestures, enabling more natural body language during conversation.

Table 3 summarizes representative agent-centric CSV systems according to the two main subsections of this survey. Unlike head-centric and body-centric analyses, agent-centric CSV couples visual social perception with agent action. Social robots require real-time perception, multimodal integration, action planning, and feedback-based adaptation in physical environments. Virtual social embodiment, by contrast, emphasizes social signal reconstruction, transfer, and generation, including facial animation, gaze behaviour, and gesture synthesis. Across both settings, the central question is not only whether an agent can recognize social cues, but whether it can use them to produce socially appropriate responses in closed-loop interaction.

VI. Challenge and Future Trend

Recent advances in CSV have been strongly driven by the rapid development of AI and computer vision. However, many AI techniques were originally designed for mathematical optimization, pattern recognition, or engineering-oriented prediction, rather than for modelling socially situated human behaviour. As a result, AI-grounded social vision approaches may not always be well aligned with how humans perceive, interpret, and respond to social signals. Kyrlitsias and Michael-Grigoriou [94] also argue that social meaning is not inherently embedded in classical AI and computer vision pipelines. A key challenge for CSV is therefore to move beyond visual cue detection and develop computational models that can better capture the psychological, contextual, and interactive nature of social understanding.

One important direction concerns evaluation. In computer vision and machine learning, quantitative measures such as accuracy, F1-score, mean squared error, and structural similarity index measure (SSIM) are widely used to assess model performance. Although these metrics are necessary for engineering evaluation, they are insufficient for measuring whether a system produces socially meaningful, natural, or appropriate interpretations. For example, a facial expression generator may achieve low pixel-level error but still produce expressions that appear unnatural or socially inappropriate to human observers. Similarly, a gaze or pose estimator may obtain high numerical accuracy while failing to support correct inference of attention, engagement, or intention in real interaction. Future CSV evaluation should therefore incorporate human-centred criteria, such as perceived naturalness, social appropriateness, trust, engagement, and interaction quality. Insights from psychology and cognitive science are needed to reshape evaluation frameworks so that computational measures can be better aligned with human perception of social behaviour [110, 111].

Another challenge is the contextual and cultural variability of social signals. Visual cues such as facial expressions, gaze direction, posture, interpersonal distance, and gestures do not carry fixed meanings across situations [112–114]. Their interpretation may vary according to cultural norms, individual differences, social relationships, interaction history, and task context. Therefore, CSV systems trained on limited datasets may produce biased or culturally insensitive interpretations when deployed in diverse populations. Future research should pay more attention to social bias, cultural sensitivity, and personalization. Instead of assuming universal mappings from visual cues to social states, CSV models should learn adaptive representations that account for individual expressivity, user preference, and long-term interaction patterns.

Finally, CSV should be further developed as a multimodal and closed-loop framework for HAI. In real social interaction, visual evidence rarely operates alone. Facial movements,

Table 3 Representative agent-centric CSV systems and applications.

Category	Representative work	Social cue or target	Critical idea
Social robot	Robinson et al. [85]	Robotic visual perception	Reviews robotic vision for gesture recognition, navigation, handover, social communication, and learning from demonstration.
	Urakami and Seaborn [87]	Robot social agency	Translates nonverbal cues into robot design patterns from a communication perspective.
	Duncan et al. [86]	Human communication and intent	Surveys multimodal perception systems combining pose, gaze, gesture, speech, and context.
	Zhong et al. [88]	Target-user following	Improves Pepper robot perception using human pose estimation and tracking.
	Avogaro et al. [89]	3D human pose forecasting	Provides robot-perspective data with partial visibility, camera motion, and physical contact.
	Bian et al. [90]	Intent, attitude, and action	Forecasts future social actions from an agent-centred viewpoint.
	Bian et al. [91]	Robust social HRI	Distills rich multimodal teacher knowledge into a lightweight pose-based student model.
	Nigro et al. [92]	Group engagement and behaviour	Reviews computational challenges in social group interaction with robots.
	Song et al. [93]	Socially compliant navigation	Uses vision-language model scoring to interpret social cues for robot navigation.
Virtual social embodiment	Lou et al. [96]	Facial expression reconstruction	Reconstructs realistic facial expressions for head-mounted display users in virtual environments.
	Bhattacharya et al. [106]	Emotive gesture generation	Uses transformer-based text-to-gesture generation for virtual agents.
	Gunawardhana et al. [107]	Multimodal agent behaviour	Generates user-aware multimodal behaviours for interactive virtual agents.
	Xu et al. [109]	Speech-driven gesture	Uses LLM-generated speech information to guide body gesture generation.

gaze, body motion, gesture, speech, language, and environmental context jointly shape social meaning. Moreover, in HAI, perception and action are mutually coupled: an agent's response can change human behaviour, which in turn affects subsequent perception and decision-making. Future CSV systems should therefore integrate multimodal signals, model temporal dependencies, and evaluate social understanding in interactive settings rather than only on static datasets. Such progress will help CSV move from isolated recognition tasks toward socially aware agents capable of robust, adaptive, and ethically responsible interaction with humans.

VII. Conclusion

In this paper, we summarized recent advances in visual-cognitive computing based on the computational social vision concept and discussed how they could support social signal understanding by enabling inference of latent social states from visual behavioural evidence. Related concepts and methods were organized into head-centric and body-centric aspects, which provided a structured perspective on how to model human behaviours based on visual cues. We further outlined social robotics and virtual humans as a representative domain where HAI approaches are comprehensively discussed. This paper provided a unified technical map that can guide systematic comparison and help shape future research toward more robust social understanding in human-agent systems.

Acknowledgment

This work was supported by the UK Research and Innovation (No. EP/Z000025/1) and the Horizon Europe Programme under the Marie Skłodowska-Curie Actions grant for Affective Computing Models (No. 101130271).

References

[1] A. Vinciarelli, H. Salamin, and M. Pantic, Social signal processing: Understanding social interactions through nonverbal behavior analysis, in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, Miami, FL, USA, 2009, 42–49.

[2] H. Joo, T. Simon, M. Cikara, and Y. Sheikh, Towards social artificial intelligence: Nonverbal social signal prediction in a triadic interaction, in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Long Beach, CA, USA, 2019, 10865–10875.

[3] R. M. Krauss, Y. Chen, and P. Chawla, Nonverbal behavior and nonverbal communication: What do conversational hand gestures tell us?, *Adv. Exp. Soc. Psychol.*, 1996, 28, 389–450.

[4] L. E. Ishii, J. C. Nellis, K. D. Boahene, P. Byrne, and M. Ishii, The importance and psychology of facial expression, *Otolaryngol. Clin. North Am.*, 2018, 51(6), 1011–1017.

[5] Y. Wang, H. Yu, W. Gao, Y. Xia, and C. Nduka, MGEED: A multimodal genuine emotion and expression detection database, *IEEE Trans. Affect. Comput.*, 2024, 15(2), 606–619.

[6] A. Frischen, A. P. Bayliss, and S. P. Tipper, Gaze cueing of attention: Visual attention, social cognition, and individual differences, *Psychol. Bull.*, 2007, 133(4), 694–724.

[7] G. Ziv, Gaze behavior and visual attention: A review of eye tracking studies in aviation, *Int. J. Aviat. Psychol.*, 2016, 26(3&4), 75–104.

[8] A. Gaschler, S. Jentzsch, M. Giuliani, K. Huth, J. de Ruiter, and A. Knoll, Social behavior recognition using body posture and head pose for human-robot interaction, in *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems*, Vilamoura-Algarve, Portugal, 2012, 2128–2133.

[9] C. Pabba, V. Bhardwaj, and P. Kumar, A visual intelligent system for students' behavior classification using body pose and facial features in a smart classroom, *Multimed. Tools Appl.*, 2024, 83(12), 36975–37005.

[10] A. D. Wilson and A. F. Bobick, Learning visual behavior for gesture analysis, in *Proc. International Symposium on Computer Vision*, Coral Gables, FL, USA, 1995, 229–234.

[11] N. Mendes, J. Ferrer, J. Vitorino, M. Safeea, and P. Neto, Human behavior and hand gesture classification for smart human-robot interaction, *Procedia Manuf.*, 2017, 11, 91–98.

[12] F. Z. Canal, T. R. Müller, J. C. Matias, G. G. Scotton, A. R. de Sa Junior, E. Pozzebon, and A. C. Sobieranski, A survey on facial emotion recognition techniques: A state-of-the-art literature review, *Inf. Sci.*, 2022, 582, 593–617.

[13] J. Schwarz, C. C. Marais, T. Leyvand, S. E. Hudson, and J. Mankoff, Combining body pose, gaze, and gesture to determine intention to interact in vision-based interfaces, in *Proc. SIGCHI Conference on Human Factors in Computing Systems*, Toronto, Canada, 2014, 3443–3452.

[14] J. Li, Y. Wong, Q. Zhao, and M. S. Kankanhalli, Visual social relationship recognition, *Int. J. Comput. Vis.*, 2020, 128(6), 1750–1764.

[15] X. Zhao, L. Wang, Y. Zhang, X. Han, M. Deveci, and M. Parmar, A review of convolutional neural networks in computer vision, *Artif. Intell. Rev.*, 2024, 57(4), 99.

[16] G. Zhao, Y. Li, and Q. Xu, From emotion AI to cognitive AI, *Int. J. Netw. Dyn. Intell.*, 2022, 1(1), 65–72.

[17] X. Liu, Y. Xia, H. Yu, J. Dong, M. Jian, and T. D. Pham, Region based parallel hierarchy convolutional neural network for automatic facial nerve paralysis evaluation, *IEEE Trans. Neural Syst. Rehabil. Eng.*, 2020, 28(10), 2325–2332.

[18] K. Wang, C. Gou, N. Zheng, J. M. Rehg, and F.-Y. Wang, Parallel vision for perception and understanding of complex scenes: Methods, framework, and perspectives, *Artif. Intell. Rev*., 2017, 48(3), 299–329.

[19] R. B. Jr. Adams, N. Ambady, K. Nakayama, and S. Shimojo, *The Science of Social Vision*, New York, NY, USA: Oxford University Press, 2010.

[20] H. Yu, O. G. B. Garrod, and P. G. Schyns, Perception-driven facial expression synthesis, *Comput. Graph.*, 2012, 36(3), 152–162.

[21] L. A. Zebrowitz, P. M. Bronstad, and J. M. Montepare, An ecological theory of face perception, in *The Science of Social Vision*. New York, NY, USA: Oxford University Press, 2010, 3–30.

[22] R. B. Jr. Adams and K. Kveraga, Social vision: Functional forecasting and the integration of compound social cues, *Rev. Philos. Psychol.*, 2015, 6(4), 591–610.

[23] R. B. Jr. Adams, D. N. Albohn, and K. Kveraga, Social vision: Applying a social-functional approach to face and expression perception, *Curr. Dir. Psychol. Sci.*, 2017, 26(3), 243–248.

[24] R. B. Jr. Adams, R. G. Jr. Franklin, A. J. Nelson, and M. T. Stevenson, Compound social cues in human face processing, in *The Science of Social Vision*. New York, NY, USA: Oxford University Press, 2010, 90–107.

[25] B. de Gelder and M. Tamietto, Faces, bodies, social vision as agent vision, and social consciousness, in *The Science of Social Vision*. New York, NY, USA: Oxford University Press, 2010, 51–74.

[26] K. L. Johnson, F. E. Pollick, and L. S. McKay, Social constraints on the visual perception of biological motion, in *The Science of Social Vision*. New York, NY, USA: Oxford University Press, 2010, 264–277.

[27] T. Farroni and A. Senju, Specialized brain for the social vision: Perspectives from typical and atypical development, in *The Science of Social Vision*. New York, NY, USA: Oxford University Press, 2010, 421–444.

[28] R. E. Jack, O. G. B. Garrod, H. Yu, R. Caldara, and P. G. Schyns, Facial expressions of emotion are not culturally universal, *Proc. Natl. Acad. Sci. USA*, 2012, 109(19), 7241–7244.

[29] K. Garcia, E. McMahon, C. Conwell, M. F. Bonner, and L. Isik, Modeling dynamic social vision highlights gaps between deep learning and humans, in *Proc. 13th International Conference on Learning Representations*, Singapore, Singapore, 2025, 71337–71362.

[30] R. Algabri, A. Abdu, and S. Lee, Deep learning and machine learning techniques for head pose estimation: A survey, *Artif. Intell. Rev.*, 2024, 57(10), 288.

[31] A. Asperti and D. Filippini, Deep learning for head pose estimation: A survey, *SN Comput. Sci.*, 2023, 4(4), 349.

[32] C. Bisogni, L. Cascone, M. Nappi, and C. Pero, IoT-enabled biometric security: Enhancing smart car safety with depth-based head pose estimation, *ACM Trans. Multimed. Comput. Commun. Appl.*, 2024, 20(6), 155.

[33] M. Ye, W. Zhang, P. Cao, and K. Liu, Driver fatigue detection based on residual channel attention network and head pose estimation, *Appl. Sci.*, 2021, 11(19), 9195.

[34] Z. Zhao, S. Xia, X. Xu, L. Zhang, H. Yan, Y. Xu, and Z. Zhang, Driver distraction detection method based on continuous head pose estimation, *Comput. Intell. Neurosci.*, 2020, 2020(1), 9606908.

[35] H. Liu, T. Liu, Z. Zhang, A. K. Sangaiah, B. Yang, and Y. Li, ARHPE: Asymmetric relation-aware representation learning for head pose estimation in industrial human-computer interaction, *IEEE Trans. Ind. Inform.*, 2022, 18(10), 7107–7117.

[36] Z. Dai, W. Zheng, Z. Zhao, and Y. Yang, Research on learning status evaluation in the classroom based on end-to-end head pose estimation, in *Proc. 4th International Conference on Educational Technology*, Wuhan, China, 2024, 491–495.

[37] B. Massé, S. Ba, and R. Horaud, Tracking gaze and visual focus of attention of people involved in social interaction, *IEEE Trans. Pattern Anal. Mach. Intell.*, 2018, 40(11), 2711–2724.

[38] X. Wang, J. Zhang, H. Zhang, S. Zhao, and H. Liu, Vision-based gaze estimation: A review, *IEEE Trans. Cogn. Dev. Syst.*, 2022, 14(2), 316–332.

[39] S. Ghosh, A. Dhall, M. Hayat, J. Knibbe, and Q. Ji, Automatic gaze analysis: A survey of deep learning based approaches, *IEEE Trans. Pattern Anal. Mach. Intell.*, 2024, 46(1), 61–84.

[40] H. Admoni and B. Scassellati, Social eye gaze in human-robot interaction: A review, *J. Hum.-Robot Interact.*, 2017, 6(1), 25–63.

[41] J. Liang, Y.-Q. Zou, S.-Y. Liang, Y.-W. Wu, and W.-J. Yan, Emotional gaze: The effects of gaze direction on the perception of facial emotions, *Front. Psychol*., 2021, 12, 684357.

[42] R. B. Jr. Adams and R. E. Kleck, Perceived gaze direction and the processing of facial displays of emotion, *Psychol. Sci.*, 2003, 14(6), 644–647.

[43] M. W. Schurgin, J. Nelson, S. Iida, H. Ohira, J. Y. Chiao, and S. L. Franconeri, Eye movements during emotion recognition in faces, *J. Vis.*, 2014, 14(13), 14.

[44] S. Senarath, P. Pathirana, D. Meedeniya, and S. Jayarathna, Customer gaze estimation in retail using deep learning, *IEEE Access*, 2022, 10, 64904–64919.

[45] S. Yang, M. Jin, and Y. He, Continuous gaze tracking with implicit saliency-aware calibration on mobile devices, *IEEE Trans. Mob. Comput.*, 2023, 22(10), 5816–5828.

[46] N. Modi and J. Singh, Real-time camera-based eye gaze tracking using convolutional neural network: A case study on social media website, *Virtual Real.*, 2022, 26(4), 1489–1506.

[47] B. David-John, C. Peacock, T. Zhang, T. S. Murdison, H. Benko, and T. R. Jonker, Towards gaze-based prediction of the intent to interact in virtual reality, in *Proc. ACM Symposium on Eye Tracking Research and Applications*, Virtual Event, 2021, 2.

[48] X. Xu, J. Chen, C. Li, C. Fu, L. Yang, Y. Yan, and Z. Lyu, Robust gaze point estimation for metaverse with common mode features suppression network, *IEEE Trans. Consum. Electron.*, 2024, 70(1), 2090–2098.

[49] N. Chen, Y. Shen, T. Zhang, Y. Yang, and H. Wen, EX-Gaze: High-frequency and low-latency gaze tracking with hybrid event-frame cameras for on-device extended reality, *IEEE Trans. Vis. Comput. Graph.*, 2025, 31(5), 2299–2309.

[50] H. Yu and H. Liu, Regression-based facial expression optimization, *IEEE Trans. Hum.-Mach. Syst.*, 2014, 44(3), 386–394.

[51] H. Jin, M. Jian, D. Ding, and H. Yu, HMamba-3DFT: A hierarchical mamba framework for emotion-driven semantic 3D facial tracking, *Pattern Recognit.*, 2026, 178, 113415.

[52] Y. Wang, S. Yan, Y. Liu, W. Song, J. Liu, Y. Chang, X. Mai, X. Hu, W. Zhang, and Z. Gan, A survey on facial expression recognition of static and dynamic emotions, arXiv preprint arXiv: 2408.15777, 2024.

[53] S. Denham, H. Warren, M. von Salisch, O. Benga, J.-C. Chin, and E. Geangu, Emotions and social development in childhood, in *The Wiley-Blackwell Handbook of Childhood Social Development*, 2nd ed. Chichester, UK: Wiley-Blackwell, 2011, 413–433.

[54] H. Zhou, S. Huang, F. Zhang, and C. Xu, CEPrompt: Cross-modal emotion-aware prompting for facial expression recognition, *IEEE Trans. Circuits Syst. Video Technol.*, 2024, 34(11), 11886–11899.

[55] F.-S. Tsai, Y.-L. Hsu, W.-C. Chen, Y.-M. Weng, C.-J. Ng, and C.-C. Lee, Toward development and evaluation of pain level-rating scale for emergency triage based on vocal characteristics and facial expressions, in *Proc. 17th Annual Conference of the International Speech Communication Association*, 2016, 92–96.

[56] Y. Xia, H. Yu, X. Wang, M. Jian, and F.-Y. Wang, Relation-aware facial expression recognition, *IEEE Trans. Cogn. Dev. Syst*., 2022, 14(3), 1143–1154.

[57] R. E. Jack and P. G. Schyns, The human face as a dynamic tool for social communication, *Curr. Biol.*, 2015, 25, R621–R634.

[58] Y. Tie and L. Guan, A deformable 3D facial expression model for dynamic human emotional state recognition, *IEEE Trans. Circuits Syst. Video Technol.*, 2013, 23(1), 142–157.

[59] R. E. Jack, O. G. B. Garrod, and P. G. Schyns, Dynamic facial expressions of emotion transmit an evolving hierarchy of signals over time, *Curr. Biol.*, 2014, 24(2), 187–192.

[60] C. Zheng, W. Wu, C. Chen, T. Yang, S. Zhu, J. Shen, N. Kehtarnavaz, and M. Shah, Deep learning-based human pose estimation: A survey, *ACM Comput. Surv.*, 2024, 56(1), 11.

[61] L. K. Topham, W. Khan, D. Al-Jumeily, and A. Hussain, Human body pose estimation for gait identification: A comprehensive survey of datasets and models, *ACM Comput. Surv.*, 2023, 55(6), 120.

[62] P. M. Müller, S. Amin, P. Verma, M. Andriluka, and A. Bulling, Emotion recognition from embedded bodily expressions and speech during dyadic interactions, in *Proc. International Conference on Affective Computing and Intelligent Interaction*, Xi'an, China, 2015, 663–669.

[63] M. Romeo, D. Hernández García, T. Han, A. Cangelosi, and K. Jokinen, Predicting apparent personality from body language: Benchmarking deep learning architectures for adaptive social human-robot interaction, *Adv. Robot.*, 2021, 35(19), 1167–1179.

[64] A. Martin-Ozimek, I. Jayarathne, S. L. Mon, and J. Y. Chew, Diffusion-based imitation learning for social pose generation, in *Proc. 20th ACM/IEEE International Conference on Human-Robot Interaction*, Melbourne, Australia, 2025, 1488–1492.

[65] H. Guan and R. W. H. Lau, PoseSOR: Human pose can guide our attention, in *Proc. 18th European Conference on Computer Vision*, Milan, Italy, 2024, 350–366.

[66] M. Balazia, P. Müller, Á. L. Tánczos, A. von Liechtenstein, and F. Brémond, Bodily behaviors in social interaction: Novel annotations and state-of-the-art evaluation, in *Proc. 30th ACM International Conference on Multimedia*, Lisboa, Portugal, 2022, 70–79.

[67] P. Kratzer, N. B. Midlagajni, M. Toussaint, and J. Mainprice, Anticipating human intention for full-body motion prediction in object

grasping and placing tasks, in *Proc. 29th IEEE International Conference on Robot and Human Interactive Communication*, Naples, Italy, 2020, 1157–1163.

[68] J. Peng, Y. Zhou, and P. Y. Mok, KTPFormer: Kinematics and trajectory prior knowledge-enhanced transformer for 3D human pose estimation, in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, WA, USA, 2024, 1123–1132.

[69] Y. Feng, J. Lin, S. K. Dwivedi, Y. Sun, P. Patel, and M. J. Black, ChatPose: Chatting about 3D human pose, in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, WA, USA, 2024, 2093–2103.

[70] N. Mohamed, M. B. Mustafa, and N. Jomhari, A review of the hand gesture recognition system: Current progress and future directions, *IEEE Access*, 2021, 9, 157422–157436.

[71] R. Rastgoo, K. Kiani, and S. Escalera, Sign language recognition: A deep survey, *Expert Syst. Appl.*, 2021, 164, 113794.

[72] M. Peral, A. Sanfeliu, and A. Garrell, Efficient hand gesture recognition for human-robot interaction, *IEEE Robot. Autom. Lett.*, 2022, 7(4), 10272–10279.

[73] H. Hu, W. Zhao, W. Zhou, Y. Wang, and H. Li, SignBERT: Pre-training of hand-model-aware representation for sign language recognition, in *Proc. IEEE/CVF International Conference on Computer Vision*, Montreal, Canada, 2021, 11067–11076.

[74] R. Zuo, F. Wei, and B. Mak, Natural language-assisted sign language recognition, in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Vancouver, Canada, 2023, 14890–14900.

[75] C. Hubert, N. Odic, M. Noel, S. Gharib, S. H. H. Zargarbashi, and L. Séoud, MuViH: Multi-view hand gesture dataset and recognition pipeline for human-robot interaction in a collaborative robotic finishing platform, *Robot. Comput.-Integr. Manuf.*, 2025, 94, 102957.

[76] G. A. Rao, K. Syamala, P. V. V. Kishore, and A. S. C. S. Sastry, Deep convolutional neural networks for sign language recognition, in *Proc. Conference on Signal Processing and Communication Engineering Systems*, Vijayawada, India, 2018, 194–197.

[77] W. Yang, J. Tao, and Z. Ye, Continuous sign language recognition using level building based on fast hidden Markov model, *Pattern Recognit. Lett.*, 2016, 78(2), 28–35.

[78] L. Hu, L. Gao, Z. Liu, and W. Feng, Continuous sign language recognition with correlation network, in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Vancouver, Canada, 2023, 2529–2539.

[79] J. L. Herrera, B. Sedlak, and S. Dustdar, Active inference for sustainable computing continuum systems, *IEEE Internet Comput.*, 2026, 30(1), 81–90.

[80] Y. Dong, S. Zhang, X. Yao, J. Zhang, S. Huang, and Y. Li, Active collaboration in IoT: Cognitive-agent oriented resource-centric architecture, in *Proc. 33rd IEEE International Conference on Computer Communications and Networks*, Kailua-Kona, HI, USA, 2024.

[81] I. da Silva Portugal, P. Alencar, and D. Cowan, Agentic recommender systems: A systematic literature review, *IEEE Trans. Softw. Eng.*, 2026, 52(4), 1234–1264.

[82] Y. Shang, P. Liu, Y. Yan, Z. Wu, L. Sheng, Y. Yu, C. Jiang, A. Zhang, F. Xu, Y. Wang, M. Zhang, and Y. Li, AgentRecBench: Benchmarking LLM agent-based personalized recommender systems, in *Proc. 39th Conference on Neural Information Processing Systems*, Vancouver, Canada, 2025.

[83] K. R. Chirumamilla, Multi-agent AI assistant for real-time business insights, *J. Recent Trends Comput. Sci. Eng.*, 2025, 13(3), 17–36.

[84] H. Luo, G. Sun, Y. Liu, D. Niyato, H. Yu, M. Atiquzzaman, and S. Dustdar, A trustworthy agentic multi-LLM network: Challenges, solutions, and a use case, *IEEE Wirel. Commun*., to be published.

[85] N. Robinson, B. Tidd, D. Campbell, D. Kulić, and P. Corke, Robotic vision for human-robot interaction and collaboration: A survey and systematic review, *ACM Trans. Hum.-Robot Interact.*, 2023, 12(1), 12.

[86] J. A. Duncan, F. Alambeigi, and M. W. Pryor, A survey of multimodal perception methods for human-robot interaction in social environments, *ACM Trans. Hum.-Robot Interact.*, 2024, 13(4), 46.

[87] J. Urakami and K. Seaborn, Nonverbal cues in human-robot interaction: A communication studies perspective, *ACM Trans. Hum.-Robot Interact.*, 2023, 12(2), 22.

[88] W. Zhong, L. Tian, D. T. Le, and H. Rezatofighi, Improving visual perception of a social robot for controlled and in-the-wild human-robot interaction, in *Proc. ACM/IEEE International Conference on Human-Robot Interaction*, Boulder, CO, USA, 2024, 1199–1203.

[89] A. Avogaro, A. Toaiari, F. Cunico, X. Xu, H. Dafas, A. Vinciarelli, E. Li, and M. Cristani, Exploring 3D human pose estimation and forecasting from the robot's perspective: The HARPER dataset, in *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems*, Abu Dhabi, United Arab Emirates, 2024, 5828–5835.

[90] T. Bian, Y. Ma, M. Chollet, V. Sanchez, and T. Guha, Interact with me: Joint egocentric forecasting of intent to interact, attitude and social actions, in *Proc. IEEE International Conference on Multimedia and Expo*, Nantes, France, 2025, 1–6.

[91] T. Bian, M. Chollet, and T. Guha, Robust understanding of human-robot social interactions through multimodal distillation, in *Proc. 33rd ACM International Conference on Multimedia*, Dublin, Ireland, 2025, 5726–5734.

[92] M. Nigro, E. Akinrintoyo, N. Salomons, and M. Spitale, Social group human-robot interaction: A scoping review of computational challenges, in *Proc. 20th ACM/IEEE International Conference on Human-Robot Interaction*, Melbourne, Australia, 2025, 468–478.

[93] D. Song, J. Liang, A. Payandeh, A. H. Raj, X. Xiao, and D. Manocha, VLM-Social-Nav: Socially aware robot navigation through scoring using vision-language models, *IEEE Robot. Autom. Lett.*, 2025, 10(1), 508–515.

[94] C. Kyrlitsias and D. Michael-Grigoriou, Social interaction with agents and avatars in immersive virtual environments: A survey, *Front. Virtual Real.*, 2022, 2, 786665.

[95] S. Lim, R. Schmälzle, and G. Bente, Artificial social influence via human-embodied AI agent interaction in immersive virtual reality (VR): Effects of similarity-matching during health conversations, *Comput. Hum. Behav. Artif. Hum.*, 2025, 5, 100172.

[96] J. Lou, Y. Wang, C. Nduka, M. Hamedi, I. Mavridou, F.-Y. Wang, and H. Yu, Realistic facial expression reconstruction for VR HMD users, *IEEE Trans. Multimed*., 2020, 22(3), 730–743.

[97] Q. Cao, H. Yu, P. Charisse, S. Qiao, and B. Stevens, Is high-fidelity important for human-like virtual avatars in human computer interactions?, *Int. J. Netw. Dyn. Intell.*, 2023, 2(1), 15–23.

[98] Z. Chang, H. Jin, Y. Song, Y. Sun, and H. Yu, GAT-NeRF: Geometry-aware-transformer-enhanced neural radiance fields for high-fidelity 4D facial avatars, *ACM Trans. Multimed. Comput. Commun. Appl.*, 2026, 22(2), 48.

[99] W. Steptoe, R. Wolff, A. Murgia, E. Guimaraes, J. Rae, P. Sharkey, D. Roberts, and A. Steed, Eye-tracking for avatar eye-gaze and interactional analysis in immersive collaborative virtual environments, in *Proc. ACM Conference on Computer Supported Cooperative Work*, San Diego, CA, USA, 2008, 197–200.

[100] A. Richard, C. Lea, S. Ma, J. Gall, F. de la Torre, and Y. Sheikh, Audio- and gaze-driven facial animation of codec avatars, in *Proc. IEEE Winter Conference on Applications of Computer Vision*, Waikoloa, HI, USA, 2021, 41–50.

[101] J. Kang, T. Kim, H. Kim, and S.-H. Lee, Real-time translation of upper-body gestures to virtual avatars in dissimilar telepresence environments, *IEEE Trans. Vis. Comput. Graph*., 2025, 31(10), 8711–8725.

[102] J. Jiang, P. Streli, H. Qiu, A. Fender, L. Laich, P. Snape, and C. Holz, AvatarPoser: Articulated full-body pose tracking from sparse motion sensing, in *Proc. 17th European Conference on Computer Vision*, Tel Aviv, Israel, 2022, 443–460.

[103] J. S. Park, J. O'Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein, Generative agents: Interactive simulacra of human behavior, in *Proc. 36th Annual ACM Symposium on User Interface Software and Technology*, San Francisco, CA, USA, 2023, 2.

[104] X. Liu, Q. Wu, H. Zhou, Y. Du, W. Wu, D. Lin, and Z. Liu, Audio-driven co-speech gesture video generation, in *Proc. 36th Conference on Neural Information Processing Systems*, New Orleans, LA, USA, 2022, 1554.

[105] X. Wang, N. Cao, Q. Chen, and S. Cao, The interaction design of 3D virtual humans: A survey, *Comput. Sci. Rev.*, 2024, 53, 100653.

[106] U. Bhattacharya, N. Rewkowski, A. Banerjee, P. Guhan, A. Bera, and D. Manocha, Text2Gestures: A transformer-based network for generating emotive body gestures for virtual agents, in *Proc. IEEE Virtual Reality and 3D User Interfaces*, Lisboa, Portugal, 2021, 1–10.

[107] B. S. Gunawardhana, Y. Zhang, Q. Sun, and Z. Deng, Toward user-aware interactive virtual agents: Generative multi-modal agent behaviors in VR, in *Proc. IEEE International Symposium on Mixed and Augmented Reality*, Bellevue, WA, USA, 2024, 1068–1077.

[108] T. T. N. Nguyen and D. T. Tran, Realistic interactive facial reaction for simulated patient, in *Proc. 14th IEEE Global Conference on Consumer Electronics*, Osaka, Japan, 2025, 529–530.

[109] C. Xu, G. Lyu, J. Yan, M. Yang, and C. Deng, LLM knows body language, too: Translating speech voices into human gestures, in *Proc. 62nd Annual Meeting of the Association for Computational Linguistics* (*Volume 1: Long Papers*), Bangkok, Thailand, 2024, 5004–5013.

[110] J. Xu, M. Jiang, S. Wang, M. S. Kankanhalli, and Q. Zhao, Predicting human gaze beyond pixels, *J. Vis.*, 2014, 14(1), 28.

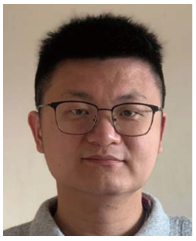
[111] S. Fan, Z. Shen, M. Jiang, B. L. Koenig, J. Xu, M. S. Kankanhalli, and Q. Zhao, Emotional attention: A study of image sentiment and visual attention, in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018, 7521–7531.

[112] H. Zhao, Z. Wang, G. Cheng, W. Qian, P. Chen, J. Yin, S. Dustdar, and S. Deng, Online workload scheduling for social welfare maximization in the computing continuum, *IEEE Trans. Serv. Comput.*, 2025, 18(4), 2267–2280.

[113] T. Peroutka, I. Murturi, P. K. Donta, and S. Dustdar, A graph-based approach to human activity recognition, in *Proc. 6th IEEE International Conference on Cognitive Machine Intelligence*, Washington, DC, USA, 2024, 117–126.

[114] D. Keltner, D. Sauter, J. Tracy, and A. Cowen, Emotional expression: Advances in basic emotion theory, *J. Nonverbal Behav.*, 2019, 43(2), 133–160.

Hui Yu received the PhD degree from Brunel University of London, UK, in 2009. He is currently a professor of visual and cognitive computing and leads the Visual Computing and Social Robotics Group, School of Psychology and Neuroscience, University of Glasgow, UK. His research interests include visual cognitive computing as well as machine learning with applications to image/video understanding, four-dimensional facial analysis, social robot, human-machine interaction, and intelligent vehicle. He has been awarded the Industrial Fellowship by the Royal Academy of Engineering. He served as the associate vice president of IEEE Systems, Man and Cybernetics Society, and an associate editor for *IEEE Transactions on Human-Machine Systems*, *IEEE Transactions on Intelligent Vehicles*, and *IEEE Transactions on Computational Social Systems.*



Yiming Wang received the PhD degree in the field of computer vision from University of Portsmouth, UK, in 2018. He is currently a research scientist at School of Psychology and Neuroscience, University of Glasgow, UK. His research interests include machine learning, computer vision, and human-machine interaction.

Haodong Jin received the MS degree from Shandong University of Finance and Economics, China, in 2023. He is currently pursuing the PhD degree at Department of Control Science and Engineering, University of Shanghai for Science and Technology, Shanghai, China. His research interests include 3D face reconstruction and medical image processing.

Schahram Dustdar received the PhD degree in business informatics from University of Linz, Austria, in 1992. He is currently a full professor of computer science and heads Research Division of Distributed Systems, Vienna University of Technology, Austria. He is also a part time research professor at Catalan Institute for Research and Advanced Studies, Pompeu Fabra University, Barcelona, Spain. His research interests include distributed systems, edge intelligence, and complex and autonomic software systems. He is the editor-in-chief of *Computing*, associate editor of *ACM Transactions on the Web*, *ACM Transactions on Internet Technology*, *IEEE Transactions on Cloud Computing*, and *IEEE Transactions on Services Computing*. He is also on the editorial boards of *IEEE Internet Computing and IEEE Computer*. He has received the ACM Distinguished Scientist Award, Distinguished Speaker Award, and IBM Faculty Award. He is an elected member of Academia Europaea, where he was Informatics Section chairman from 2015 to 2022. He is an IEEE Fellow, AAIA Fellow, and AIIA Fellow.