

Multimodal Sensor Fusion for Context-Aware Ambient Intelligence Environments

Chandrasekar M^{1*}, Muthukumar S² and Krishnamurthy B³

¹Department of Intelligent Systems, Periyar Institute of Autonomous Systems Research, India Email:

meera.chandrasekar@periyar.edu

ORCID: <https://orcid.org/5288-2901-3462-4000>

²Department of Computational Intelligence, Achankovil Center for Applied Machine Learning, Bangladesh Email:

swathi.muthukumar@achankovil.edu

ORCID: <https://orcid.org/3512-3805-9380-3530>

³Department of Information Technology, Noyyal Center for Future Computing Architectures, South Korea Email:

bhavana.krishnamurthy@noyyal.edu

ORCID: <https://orcid.org/6863-7878-0088-3857>

ABSTRACT

Ambient intelligence environments promise proactive, unobtrusive support for everyday activity, but no single sensor modality is sufficient to characterize a real physical situation with confidence: motion sensors cannot identify who is present, microphones cannot see, cameras raise privacy concerns and fail in the dark, and wearables alone cannot capture the surrounding environment. This article synthesizes the literature on multimodal sensor fusion as the mechanism that turns heterogeneous, individually weak sensor signals into context estimates an ambient intelligence system can act on. We trace the foundational definitions of context and context-awareness through the ambient intelligence vision and its deployment in domains such as assisted living, then examine the fusion problem itself, from the classical levels-of-fusion framework predating deep learning through early, late, and intermediate fusion architectures to modern attention-based transformer designs. We survey applications spanning wearable activity recognition, audio-visual and affective computing, device-free WiFi sensing, and IoT environmental event detection, reporting concrete, source-verified performance figures rather than general claims. Read together, the evidence shows a technically maturing field in which independent research groups keep converging on structurally similar fusion designs, while leaving sensing-hardware heterogeneity, resource and communication constraints at the edge, privacy in continuously sensed environments, and the gap between curated benchmarks and real deployments as open problems that will determine how far the ambient intelligence vision travels from demonstration to everyday practice.

Keywords: multimodal sensor fusion; ambient intelligence; context-aware computing; activity recognition; Internet of Things sensing

Citation: Chandrasekar M, Muthukumar S and Krishnamurthy B [2023]. Multimodal Sensor Fusion for Context-Aware Ambient Intelligence Environments. DOI: <https://doi.org/10.5281/zenodo.22843864>

Copyright: © 2023 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 19 November 2022 Accepted: 13 February 2023 Published: 21 May 2023

Article Class: Research Article

1. Introduction

A room that dims its lights when someone falls asleep on the couch, a kitchen that notices a stove left unattended, a hallway that recognizes an elderly resident's unsteady gait before a fall happens — these scenarios have been the working vision of ambient intelligence for more than two decades, and they all rest on the same unglamorous prerequisite: a system that can take in signals from several different kinds of sensors at once and turn them into a coherent picture of what is actually happening in a physical space. Mark Weiser's founding essay on ubiquitous computing imagined computation that "weaves itself into the fabric of everyday life until it is indistinguishable from it" [1], and that vision has aged well as a description of where the field wants to go, even though the sensing and fusion machinery required to get there has taken far longer to mature than the essay implied. A single sensor modality is rarely sufficient for this kind of ambient reasoning: a passive infrared motion sensor cannot distinguish a person from a pet, a microphone cannot see through walls, a camera struggles in the dark or when privacy considerations rule it out entirely, and a wearable accelerometer alone cannot tell whether the person wearing it is alone or in company. Multimodal sensor fusion — the practice of combining measurements from heterogeneous sensors so that the combined estimate is more accurate, more robust, or simply more informative than any single sensor's output — is what turns a collection of individually weak signals into a system capable of context-aware behavior.

This article synthesizes the literature connecting three threads that are often discussed separately but that, taken together, define the engineering problem of context-aware ambient intelligence: the theory of context-awareness itself, which asks what "context" means and how a system should represent and reason about it [2,3]; the ambient intelligence research program, which specifies the kind of proactive, unobtrusive, human-centered environment this context should ultimately serve [4,5]; and the substantial body of work, spanning classical statistical estimation and modern deep learning, on how to actually fuse heterogeneous sensor streams into usable context estimates [6,7,8]. None of these three threads is new on its own. What has changed meaningfully in the last decade is the fusion layer: where early ambient intelligence systems leaned on hand-engineered rules

and relatively simple statistical combination, the field has shifted toward learned fusion architectures — convolutional and recurrent networks that jointly model multiple sensor streams [9], and more recently transformer-based architectures that use attention to decide, dynamically, which modality to trust and when [10]. This shift is not merely an accuracy improvement; it changes what "context-aware" can practically mean, because a learned fusion model can, in principle, adapt to a new combination of sensors or a new environment with far less manual re-engineering than a rule-based system ever could.

We organize the discussion around six themes. We first lay out the foundational definitions of context and context-awareness that the field has converged on, since a surprising amount of confusion in applied ambient intelligence work traces back to under-specified notions of what "context" even is. We then trace the ambient intelligence vision from its early articulation through its practical deployment in domains like assisted living, before turning to the fusion problem itself: the classical levels-of-fusion framework that predates deep learning by decades, and the architectural patterns — early, late, and hybrid fusion, and their modern attention-based descendants — that determine how multiple sensor streams are actually combined inside a model. A dedicated section surveys applications and the concrete, verifiable performance numbers the applied literature reports, because claims about sensor fusion are only useful to a practitioner when they come attached to a number and a dataset. We close with a discussion of open challenges — heterogeneity, energy and communication cost at the edge, privacy, and the persistent gap between benchmark datasets and messy real-world deployments — and a conclusion that tries to state plainly what the literature does and does not yet support.

It is worth being explicit, as with any literature synthesis, about what this article is not. It is not a report of new experiments, and none of the numbers quoted here were generated by the authors; every quantitative claim is drawn directly from a cited source and, where a source reports a comparison against a baseline, we say so rather than presenting the number as a general truth about sensor fusion. The honest state of this literature, as will become clear, is one of considerable architectural convergence — independent groups working on activity

recognition, IoT event detection, and affective computing keep arriving at structurally similar fusion designs — alongside a persistent and somewhat under-acknowledged gap between what gets published on curated benchmark datasets and what would be required to run continuously, reliably, and privately in someone's actual living room.

2. Foundations of Context-Aware Computing

"Context" is one of those terms that computing borrowed from ordinary language and then spent two decades trying to pin down precisely enough to build systems around. Dey's frequently cited definition treats context as "any information that can be used to characterize the situation of an entity," where an entity is a person, place, or object relevant to the interaction between a user and an application, including the user and application themselves [2]. The definition is deliberately broad, and Dey's accompanying argument is that this breadth is a feature rather than a flaw: earlier attempts to define context by enumerating example categories (location, identity, time, and so on) kept failing to generalize, because almost any piece of information can be contextual in the right situation. What matters operationally, in Dey's framing, is not a taxonomy of context types but an architecture — the Context Toolkit — that separates the acquisition of raw sensor data from the interpretation of that data into higher-level context abstractions that applications can consume without needing to know which physical sensor produced the underlying signal. This separation of concerns, acquisition versus interpretation, turns out to be exactly the architectural seam at which multimodal fusion has to happen: fusion is, in one sense, nothing more than a particularly sophisticated interpretation layer sitting between raw multi-sensor data and the context abstractions an ambient intelligence application actually reasons over.

Baldauf, Dustdar, and Rosenberg's survey of context-aware systems, published a few years after Dey's foundational work, is useful precisely because it treats context-awareness as a systems and architecture problem rather than a purely conceptual one [3]. Their survey lays out a now-standard layered architecture for context-aware systems — sensors at the bottom, a raw data retrieval layer, a preprocessing layer that fuses and filters raw sensor readings, a storage and management layer, and an application layer at the top — and this

layering has proven durable enough that most subsequent context-aware and ambient intelligence architectures, including recent IoT-focused ones, are recognizable variants of it. The survey's treatment of context reasoning strategies (rule-based, ontology-based, and fuzzy-logic approaches, among others) is a useful reminder that the "fusion" and "reasoning" problems, while related, are not identical: fusion produces a fused estimate of some contextual variable — a person is present, a room's occupancy is three, an activity is "cooking" — while reasoning decides what to do given that estimate. Multimodal sensor fusion research has, understandably, concentrated its effort on the fusion side of this pipeline, but the survey's insistence on treating the whole pipeline as an architecture worth designing explicitly, rather than an incidental byproduct of whichever sensors happened to be available, remains sound advice.

The move from context-awareness as a general ubiquitous-computing concern to context-awareness specifically for the Internet of Things is documented thoroughly in Perera, Zaslavsky, Christen, and Georgakopoulos's survey, which examined fifty context-aware IoT projects published between 2001 and 2011 and organized them along dimensions including context acquisition method, context modeling representation, and reasoning technique [11]. Their central observation, which has aged into something close to conventional wisdom, is that the sheer heterogeneity of IoT sensing hardware — differing in modality, sampling rate, precision, power budget, and communication protocol — makes context acquisition itself a nontrivial engineering problem well before any fusion or reasoning can begin, and that a disproportionate share of the systems they surveyed under-specified how they actually handled this heterogeneity in practice. Sezer, Dogdu, and Ozbayoglu's complementary survey extends this picture forward into the era where machine learning, and specifically deep learning, had become the default context-inference mechanism for IoT systems, cataloguing the shift from largely rule-based context reasoning toward learned models trained on the kind of large, heterogeneous sensor datasets that IoT deployments increasingly generate [12]. Read together, these three strands — Dey's conceptual grounding, Baldauf et al.'s architectural layering, and the IoT-specific surveys' documentation of sensing

heterogeneity — establish the vocabulary and the problem shape that the rest of this article's discussion of fusion techniques presupposes.

3. The Ambient Intelligence Vision and Its Deployment

Ambient intelligence, as a research program, is best understood as ambient-computing-with-a-specific-mission: environments that are sensitive to the presence of people, adaptive to their needs, and — crucially — largely invisible in their operation, closely paralleling Weiser's original ubiquitous computing vision but placing somewhat more emphasis on proactive, anticipatory behavior rather than simply making computation pervasively available [1]. Cook, Augusto, and Jakkula's survey of ambient intelligence technologies, applications, and opportunities remains a useful map of the field's scope a decade and a half after publication, precisely because it organizes ambient intelligence not as a single technology but as an integration problem spanning sensor networks, activity recognition, prediction and decision-making, and human-computer interaction, all coordinated toward the goal of supporting people in their everyday environments without demanding explicit interaction from them [4]. Their survey is candid that the integration problem, more than any single component technology, is what had historically limited ambient intelligence's move from laboratory testbeds to real deployments, and that observation has continued to hold: the sensing and fusion literature surveyed later in this article has matured considerably since 2009, but the systems-integration challenge of stitching a fusion pipeline into a full ambient intelligence application that a non-expert can install, trust, and maintain has not disappeared.

Assisted living for older adults is the application domain where the ambient intelligence vision has been pursued most concretely and for the longest sustained period, largely because the value proposition — supporting independent living and early detection of health decline without constant human supervision — is both socially significant and relatively well aligned with what sensor networks can plausibly deliver. Rashidi and Mihailidis's survey of ambient-assisted-living tools for older adults catalogues the sensing modalities this domain has relied on: passive infrared and pressure sensors for occupancy

and mobility tracking, wearable accelerometers for fall detection and gait analysis, RFID and contact sensors for object-interaction and activities-of-daily-living monitoring, and, increasingly, ambient audio and video sensing for richer activity and anomaly detection [5]. Their survey makes a point that is easy to overlook in a purely technical treatment of sensor fusion: fusion quality in this domain is not an abstract accuracy metric, it directly determines whether a system generates a manageable number of false alarms or, alternatively, either misses a genuine fall or nags its user into disabling it. A single-modality fall-detection system relying only on accelerometer thresholding, for instance, tends to confuse sitting down quickly with falling; combining accelerometer data with ambient occupancy and, where available, audio cues substantially narrows this class of false positive, which is precisely why fusion, rather than any single improved sensor, has been the dominant lever for improving assisted-living system reliability.

The IoT survey literature situates ambient intelligence and assisted-living systems within the broader architecture of Internet-connected sensing infrastructure, and Lin et al.'s survey of IoT architecture, enabling technologies, security, and privacy is useful here as a reminder that ambient intelligence deployments inherit essentially every systems challenge that general-purpose IoT deployments face — heterogeneous connectivity, constrained device resources, and a security and privacy surface that expands with every additional sensor added to an environment [13]. This inheritance matters for a simple reason: an ambient intelligence system's fusion algorithm can be arbitrarily sophisticated and still fail in deployment if the underlying sensor network cannot reliably deliver synchronized, timestamped data from its constituent devices, a systems-level precondition for fusion that the algorithmic fusion literature tends to assume rather than solve.

4. Levels and Architectures of Sensor Fusion

Long before deep learning gave sensor fusion its current architectural vocabulary, the data fusion community had already developed a framework for thinking about what fusion is and where in a processing pipeline it should happen. Hall and Llinas's introduction to multisensor data fusion, though written in 1997 for a readership working primarily on military and industrial sensing systems, laid out a levels-of-fusion framework —

commonly associated with the JDL (Joint Directors of Laboratories) data fusion model — that distinguishes low-level fusion of raw or lightly processed sensor signals, intermediate-level fusion of extracted features, and high-level fusion of independently derived decisions or situation assessments [6]. This vocabulary has proven remarkably durable: the modern deep learning literature's distinction between "early fusion" (combining raw or near-raw modality inputs before any significant modality-specific processing), "late fusion" (processing each modality independently through to a decision or embedding and combining only at the end), and "hybrid" or "intermediate" fusion (combining partially processed features at one or more intermediate layers) is, structurally, the same taxonomy Hall and Llinas described decades earlier, translated into the language of neural network layers rather than signal-processing stages.

Table 1: Representative sensor-fusion techniques spanning the classical and deep-learning eras of this literature, the fusion level each represents, and the effect each cited source reports. [6,9,10,19,22,17].

Technique	Source	Fusion Level / Setting	Reported Effect
JDL levels-of-fusion framework	Hall & Llinas [6]	Conceptual: low/intermediate/high-level fusion	Foundational taxonomy still used to classify modern deep fusion architectures
DeepConvLSTM (CNN + LSTM) wearable fusion	Ordonez, Roggen [9]	Intermediate fusion, multi-channel IMU streams	+4% avg over deep non-recurrent nets; up to +9% over prior best on OPPORTUNITY
Attention fusion-bottleneck transformer	Nagrani et al. [10]	Intermediate fusion, audio-visual	State-of-the-art on AudioSet/Epic-Kitchens/VGGSound with reduced compute vs. dense cross-attention
WiFi channel-state-information sensing	Ma et al. [19]	Device-free single/multi-modal radio sensing	Enables activity, gesture and vital-sign sensing without wearable or camera hardware
Hybrid RF + LSTM + ANN fusion	Cai et al. [22]	Feature-level fusion, temperature/gas/smoke IoT streams	96.8% accuracy, 0.982 ROC-AUC; beats best of 13 baselines by ≥ 2.4 points
Multimodal affect-detection meta-analysis	D'Mello, Kory [17]	Decision/feature-level fusion, multi-channel affect	Multimodal systems outperformed best unimodal system in the large majority of reviewed studies

Baltrusaitis, Ahuja, and Morency's survey of multimodal machine learning gives this taxonomy its most complete modern treatment, organizing the field's core technical challenges into five categories: representation (how to

encode information from multiple modalities in a way that captures their complementary structure), translation (mapping from one modality to another), alignment (identifying correspondences between elements of different modalities, such as matching a spoken word to the moment it appears in a video), fusion (the specific problem of joining information from multiple modalities to make a prediction), and co-learning (transferring knowledge between modalities, particularly useful when one modality is data-rich and another is data-poor) [8]. Their treatment of fusion specifically makes an argument that is easy to state and consequential in practice: the right fusion strategy is not a fixed architectural choice but depends on the relationship between the modalities being combined — modalities that are tightly temporally and semantically correlated (audio and lip movement in speech, for instance) tend to benefit from earlier, tighter fusion, while modalities that are more loosely related or that fail independently in different ways (a camera and a passive infrared sensor, say) often do better with fusion pushed later in the pipeline, where each modality's independent uncertainty can be represented explicitly before combination. Ramachandram and Taylor's complementary survey of deep multimodal learning covers much of the same architectural ground with a more explicitly deep-learning-centric lens, surveying multimodal autoencoders, deep belief networks, and multimodal convolutional and recurrent architectures, and making the case that a well-designed fusion architecture should be judged not only on task accuracy but on its robustness to a modality being missing or degraded at inference time — a property that matters enormously for ambient intelligence, where a sensor dropping offline mid-deployment is closer to the norm than the exception [14].

Atrey, Hossain, El Saddik, and Kankanhalli's earlier survey of multimodal fusion for multimedia analysis, predating the deep learning wave, is still worth reading as a taxonomy of fusion methodology on its own terms: they distinguish fusion by the combination strategy used (rule-based approaches such as weighted averaging, classification-based approaches such as support vector machines applied to concatenated features, and estimation-based approaches derived from the classical Kalman-filter and Bayesian-inference tradition), a categorization that maps reasonably cleanly onto how modern deep architectures are frequently described

even though the underlying mechanics — learned weights via backpropagation rather than hand-specified combination rules — are quite different [7]. Table 1 places several representative fusion techniques discussed across this literature side by side, alongside the level of fusion each represents and the concrete effect each source paper reports, to make the comparison across the classical and deep-learning-era literatures concrete rather than purely taxonomic.

A closely related but distinct strand of this literature has developed around embodied AI and robotics, where a physically situated agent — rather than a fixed ambient environment — must fuse camera, LiDAR, inertial, and proprioceptive sensor streams in real time to perceive and act in its surroundings. Ruan et al.'s recent survey of multi-sensor fusion perception for embodied AI covers much of the same fusion-level taxonomy discussed above, but with an added emphasis the fixed-environment ambient intelligence literature tends to underweight: the need for fusion to remain robust and low-latency as the sensing platform itself moves through, rather than sits still within, the environment being sensed [15]. The overlap in fusion methodology between embodied and ambient sensing is substantial enough that ambient intelligence system designers have begun borrowing architectural ideas — attention-weighted combination of modality-specific encoders, in particular — directly from the embodied AI and robotics fusion literature, even though the deployment contexts (a mobile robot versus a fixed smart-home sensor network) differ considerably.

Multi-sensor fusion for activity recognition specifically, the application area most central to ambient intelligence's day-to-day operation, has its own dedicated survey literature. Aguilera, Brena, Mayora, Molino-Minero-Re, and Trejo's survey of multi-sensor fusion for activity recognition works through the same levels-of-fusion taxonomy but grounds it specifically in the sensor modalities activity-recognition systems typically combine — inertial measurement units, ambient sensors, physiological sensors, and vision — and in the fusion algorithms (probabilistic graphical models, ensemble classifiers, and neural architectures) that have been applied to combine them [16]. Their survey's most durable contribution, in our reading, is less any single algorithmic recommendation than the observation that activity-recognition fusion research has

been slow to standardize its evaluation methodology, with different studies reporting accuracy on different, often small and idiosyncratic datasets in ways that make cross-paper comparison difficult — an observation that anticipates a theme we return to in the open-challenges discussion below.

5. Multimodal Deep Learning for Fusion

The architectural shift toward deep learning for sensor fusion is best illustrated concretely, rather than abstractly, and the wearable activity-recognition literature offers one of the clearest examples. Ordóñez and Roggen's DeepConvLSTM architecture combines convolutional layers, which learn to extract features directly from raw multimodal wearable-sensor streams (accelerometer and gyroscope data from multiple on-body locations) without hand-engineered feature extraction, with recurrent LSTM layers, which model the temporal dynamics of an activity across a sliding window of sensor readings [9]. This is intermediate fusion in the Baltrusaitis et al. taxonomy: the convolutional stage processes each sensor channel with shared or channel-specific filters before the representations are combined and passed to the recurrent stage, rather than concatenating raw sensor streams at the input (early fusion) or training entirely separate per-sensor classifiers and voting at the end (late fusion). On the OPPORTUNITY activity-recognition benchmark, the authors report that this architecture outperforms competing deep non-recurrent networks by 4% on average, and outperforms some previously reported results by up to 9% [9] — concrete evidence that jointly modeling the temporal structure of a multimodal wearable-sensor stream, rather than treating each time step's readings independently, captures information a purely feed-forward fusion architecture leaves on the table.

Attention-based fusion, and transformer architectures more specifically, represent the most recent substantial architectural shift in this literature, and Nagrani et al.'s "fusion bottleneck" architecture for audio-visual video classification is a representative and influential example [10]. Rather than fusing audio and visual modalities either very early (concatenating raw or lightly processed features, which the authors note tends to let one modality's noise or irrelevant information dominate the combined representation) or very late (independent processing until a final decision layer, which limits how

much the modalities can inform each other's intermediate representations), their architecture forces cross-modal attention to flow through a small, deliberately restricted set of "bottleneck" tokens at several intermediate layers of a transformer. The stated effect of this restriction is to force the model to compress and prioritize only the most relevant cross-modal information at each fusion point, rather than allowing unrestricted pairwise attention between every audio and every visual token, which the authors report both improves fusion performance and reduces computational cost relative to unrestricted cross-modal attention, with the architecture achieving results competitive with or exceeding prior work on audio-visual classification benchmarks including AudioSet, Epic-Kitchens, and VGGSound [10]. The bottleneck idea generalizes naturally beyond audio-visual video classification to the broader multi-sensor ambient intelligence setting: an environment instrumented with several heterogeneous sensors faces exactly the same problem the fusion-bottleneck architecture was designed to address, namely that unrestricted attention or concatenation across many modalities scales poorly and risks letting a noisy or irrelevant sensor stream dominate the fused representation.

Affective computing is a second application area, adjacent to but distinct from activity recognition, where multimodal fusion has an unusually long and well-documented research history, largely because emotional and affective states are inherently expressed across multiple channels simultaneously — facial expression, vocal prosody, physiological signals, and linguistic content all carry partially redundant and partially complementary affective information. D'Mello and Kory's meta-analysis of multimodal affect detection systems is particularly valuable here because it is one of the few pieces of this literature to quantify the fusion benefit in aggregate across many independent studies rather than reporting a single system's result: synthesizing accuracy results across the multimodal affect-detection studies in their review, they find that multimodal systems combining two or more channels achieved higher classification accuracy than the best corresponding unimodal system in the substantial majority of the studies reviewed, providing meta-analytic evidence, rather than a single anecdote, that multimodal fusion delivers a real and fairly

consistent accuracy benefit over single-channel affect sensing across a wide range of system designs and channel combinations [17]. Poria, Cambria, Bajpai, and Hussain's complementary review of affective computing traces the field's progression from unimodal sentiment and emotion analysis toward multimodal fusion in more architectural detail, covering feature-level, decision-level, and model-level (hybrid) fusion strategies specifically as applied to combining text, audio, and visual affective signals, and arguing that the choice of fusion level interacts strongly with how tightly synchronized and how individually reliable each affective channel is in a given deployment context [18] — an argument that echoes, from the affective-computing side, the same modality-relationship-dependent fusion recommendation that Baltrusaitis et al. make for multimodal machine learning generally.

6. Sensing Modalities, Applications, and Reported Performance

Ambient intelligence systems draw on a wide range of sensing modalities beyond the wearable inertial sensors and audio-visual channels discussed above, and radio-frequency sensing has emerged as a particularly active area because it offers device-free sensing — inferring human presence, location, and activity from how a person's body perturbs an existing wireless signal, without requiring the person to wear or carry anything. Ma, Zhou, and Wang's survey of WiFi sensing using channel state information catalogues this line of work in detail, covering how fine-grained amplitude and phase information extracted from commodity WiFi hardware's channel state information can be used to infer activity, gesture, gait, vital signs, and localization, generally by treating the time series of channel state information as a signal to be fused, in the fusion-architecture sense, with complementary information from motion models or from other sensing modalities [19]. Their survey is candid that WiFi-CSI sensing's principal practical appeal — no dedicated sensor hardware required, since it can piggyback on WiFi infrastructure already present in most indoor environments — comes with a corresponding fragility, since channel state information is sensitive to environmental changes (furniture rearrangement, multiple people present simultaneously, different WiFi hardware) in ways that a purpose-built sensor is not, which is precisely the kind of single-modality weakness that multimodal fusion with a second, more

environment-robust sensing channel is well suited to compensate for.

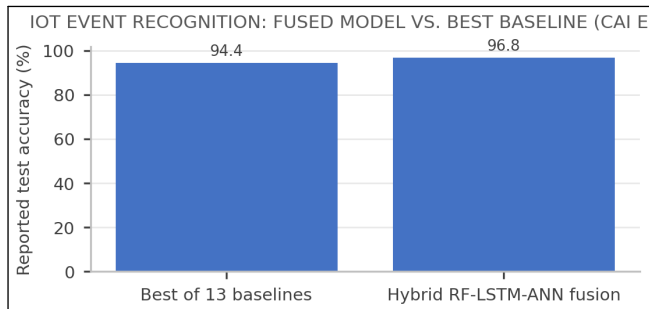


Figure 1: Test accuracy of Cai et al.'s hybrid random-forest/LSTM/ANN multimodal fusion model for IoT environmental event recognition (temperature, gas, and smoke sensor streams) against the strongest of thirteen evaluated baseline models -- the reported fusion model outperforms the best baseline by at least 2.4 percentage points [22].

Wang, Chen, Hao, Peng, and Hu's survey of deep learning for sensor-based activity recognition, and Chen, Zhang, Yao, Guo, Yu, and Liu's later and more comprehensive survey covering the same territory, together document how thoroughly deep learning has displaced hand-crafted feature engineering as the default approach to turning raw wearable and ambient sensor data into activity labels over the past decade [20,21]. Chen et al.'s survey in particular organizes the activity-recognition literature around the specific deep architectures used — convolutional networks for spatial feature extraction across sensor channels, recurrent networks for temporal modeling, autoencoders for unsupervised or semi-supervised feature learning where labeled activity data is scarce, and, increasingly, attention and transformer-based architectures — and highlights multimodal, multi-sensor fusion as one of the field's most active and still-evolving subareas rather than a solved problem, echoing the earlier-noted evaluation-standardization concern from Aguilera et al.'s narrower survey [16].

The most concrete, end-to-end demonstration of multimodal fusion's applied value in this article's specific IoT-sensing context comes from Cai, Huang, and Chen's recent work on multimodal sensor data fusion for event recognition in IoT environments [22]. Their system fuses temperature, gas, smoke, and other environmental sensor streams drawn from three real-world public datasets (UCI Gas Sensor, Intel Berkeley Lab, and a Kaggle environmental-sensing dataset) into a combined dataset of 47,234 samples across 18 features, and applies a hybrid architecture combining random forest feature selection, LSTM-based

temporal modeling, and an artificial neural network classifier to recognize environmental events. On a held-out test split, the reported hybrid fusion model achieves 96.8% accuracy, 95.9% precision, 96.3% recall, 96.1% F1-score, and a 0.982 ROC-AUC, and the authors report that this outperforms all thirteen classical machine learning and deep learning baseline models they evaluated against by at least 2.4 percentage points in test accuracy [22]. Figure 1 visualizes this reported gap between the fused multimodal model and the strongest evaluated baseline. Two things about this result are worth stating plainly rather than glossing over: first, it is one of relatively few papers in this literature that both fuses genuinely heterogeneous sensor modalities (rather than, say, multiple accelerometer axes, which are technically multi-sensor but not strongly heterogeneous) and reports a full, direct comparison against a substantial baseline set, which is exactly the kind of evidence the field needs more of; second, a 2.4-point accuracy margin over the best of thirteen baselines, while genuinely meaningful, is a more modest claim than the framing of "multimodal fusion versus unimodal" comparisons sometimes implies, since several of those thirteen baselines were themselves presumably given access to some subset of the same fused feature set rather than being restricted to a single raw sensor stream.

7. Open Challenges

Several problems recur across this literature clearly enough to be worth naming directly rather than leaving implicit. The first, and in our reading the most consequential, is the persistent gap between benchmark-dataset evaluation and real deployment conditions. Most of the quantitative results synthesized in this article — the DeepConvLSTM improvements on OPPORTUNITY [9], the fusion-bottleneck results on AudioSet and VGGSound [10], the hybrid fusion model's accuracy on the merged UCI/Intel/Kaggle dataset [22] — are reported on datasets collected under relatively controlled conditions, with a fixed and known sensor configuration, largely clean labels, and no sustained exposure to the sensor drift, hardware failure, and adversarial or simply messy real-world conditions that Perera et al.'s and Sezer et al.'s IoT-focused surveys both flag as under-addressed in the broader context-aware computing literature [11,12]. Aguilera et al.'s activity-recognition survey makes essentially the same point from a narrower angle, noting the field's

inconsistent and often small-scale evaluation datasets as a standing obstacle to comparing fusion methods meaningfully across studies [16].

Second, sensor and modality heterogeneity remains a harder engineering problem than the algorithmic fusion literature tends to assume. A fusion architecture like DeepConvLSTM or a fusion-bottleneck transformer presupposes that the input modalities are already synchronized, calibrated, and delivered in a consistent format; Lin et al.'s and Perera et al.'s surveys both make clear that achieving this precondition across heterogeneous IoT hardware, with differing sampling rates, clock drift, and intermittent connectivity, is itself a substantial and only partially solved systems problem that sits upstream of any fusion algorithm's ability to perform well [13,11]. Third, and related, resource and energy constraints at the sensing edge shape what fusion architectures are even feasible to run continuously in an ambient intelligence deployment; Zhou et al.'s survey of edge intelligence documents the broader tension between pushing inference (and, increasingly, learning) onto resource-constrained edge hardware close to where sensor data is generated, versus the greater model capacity and easier multimodal fusion available when data is aggregated and processed centrally, a tension that becomes sharper, not milder, as the number of fused sensor modalities and their combined data rate grows [23].

Fourth, privacy deserves more direct treatment than the applied fusion literature typically gives it. An ambient intelligence system that fuses audio, video, WiFi channel state information, and wearable data about a person inside their own home is, definitionally, building a substantially more complete and more sensitive picture of that person's behavior than any single sensor modality could on its own — multimodal fusion's core technical advantage (more information, more robustly combined) is simultaneously its core privacy liability. Rashidi and Mihailidis's assisted-living survey raises this tension explicitly in the context of older-adult monitoring, where the population most likely to benefit from ambient intelligence support is also a population for whom sustained, largely passive, multimodal home sensing raises real autonomy and consent questions that the fusion-architecture literature, focused as it typically is on accuracy metrics, tends not to engage with directly [5]. Fifth, and more mundane but no less real: evaluation

and reporting standards across this literature are inconsistent enough — different studies reporting accuracy, F1, AUC, or some ad hoc combination, on datasets of wildly varying size and realism — that comparing fusion techniques across papers, as this article's own Table 1 illustrates by necessity, means comparing numbers that were not designed to be directly comparable in the first place, a concern that both the activity-recognition and the general multimodal-machine-learning survey literature raise from their respective angles [16,8].

None of these challenges strikes us as a reason to doubt that multimodal sensor fusion is the right general approach for context-aware ambient intelligence; the consistency with which independent research groups working on wearable activity recognition, audio-visual classification, affective computing, and IoT event detection have converged on structurally similar fusion designs — combine modalities at an intermediate representational level where possible, let attention or a comparable learned mechanism decide how much to trust each modality rather than fixing the combination rule by hand, and treat missing or degraded modalities as a design constraint rather than an edge case — is itself fairly strong evidence that the field has found something real. What the challenges above do suggest is that the next wave of progress in this area is less likely to come from a fundamentally new fusion architecture and more likely to come from taking deployment realism — heterogeneous, drifting, resource-constrained, privacy-sensitive, genuinely uncontrolled sensing environments — as seriously as the field has, over the past decade, learned to take fusion accuracy on curated benchmarks.

8. Conclusion

The literature reviewed here supports a reasonably specific set of claims about multimodal sensor fusion for context-aware ambient intelligence. Context, properly understood in Dey's operational sense, is broad enough that almost any sensor stream can be contextually relevant, which is exactly why single-modality sensing so reliably falls short of what ambient intelligence applications need [2]. The ambient intelligence vision itself — proactive, unobtrusive, human-centered environments, in the spirit of Weiser's original ubiquitous-computing essay — has been pursued concretely enough, particularly in assisted-living

applications, that its remaining obstacles are now considerably more about systems integration, deployment realism, and trust than about a shortage of underlying sensing or fusion technology [1,4,5]. The fusion techniques themselves have matured along a fairly legible historical arc, from the classical levels-of-fusion framework that predates deep learning entirely [6], through hand-specified rule-based and estimation-based combination strategies [7], to today's learned intermediate and attention-based fusion architectures that consistently outperform both earlier fusion designs and unimodal baselines across activity recognition, audio-visual classification, affective computing, and IoT event detection [9,10,17,22].

What the evidence does not yet support, and what we think it is important to say plainly rather than paper over, is that this technology is ready to run unattended, indefinitely, in the genuinely uncontrolled conditions of an arbitrary person's home. The strongest quantitative results synthesized in this article come from carefully constructed benchmark datasets and relatively short evaluation windows, and the surveys that examine deployment realism most directly — on IoT context-awareness generally, on assisted living specifically — are consistent in flagging heterogeneity, resource constraints, and privacy as substantially open problems rather than engineering details left to be filled in [11,12,13]. Taken together, we read the literature as describing a field with a solid and increasingly well-evidenced technical core — how to fuse heterogeneous sensor modalities so that the combination genuinely outperforms any single sensor — wrapped in a still-unfinished layer of systems, privacy, and deployment engineering that will determine whether the ambient intelligence vision articulated more than three decades ago finally becomes an ordinary feature of everyday environments, or remains, as it has for most of those three decades, a compelling demonstration that has not quite made the leap to ubiquity.

9. References

1. Weiser, M. (1991) The Computer for the 21st Century. ACM SIGMOBILE Mobile Computing and Communications Review, 3(3): 3-11. DOI: [10.1145/329124.329126](https://doi.org/10.1145/329124.329126)
2. Dey, A.K. (2001) Understanding and Using Context. Personal and Ubiquitous Computing, 5(1): 4-7. DOI: [10.1007/s007790170019](https://doi.org/10.1007/s007790170019)
3. Baldauf, M., Dustdar, S., Rosenberg, F. (2007) A Survey on Context-Aware Systems. International Journal of Ad Hoc and Ubiquitous Computing, 2(4): 263-277. DOI: [10.1504/IJAHUC.2007.014070](https://doi.org/10.1504/IJAHUC.2007.014070)
4. Cook, D.J., Augusto, J.C., Jakkula, V.R. (2009) Ambient Intelligence: Technologies, Applications, and Opportunities. Pervasive and Mobile Computing, 5(4): 277-298. DOI: [10.1016/j.pmcj.2009.04.001](https://doi.org/10.1016/j.pmcj.2009.04.001)
5. Rashidi, P., Mihailidis, A. (2013) A Survey on Ambient-Assisted Living Tools for Older Adults. IEEE Journal of Biomedical and Health Informatics, 17(3): 579-590. DOI: [10.1109/JBHI.2012.2234129](https://doi.org/10.1109/JBHI.2012.2234129)
6. Hall, D.L., Llinas, J. (1997) An Introduction to Multisensor Data Fusion. Proceedings of the IEEE, 85(1): 6-23. DOI: [10.1109/5.554205](https://doi.org/10.1109/5.554205)
7. Atrey, P.K., Hossain, M.A., El Saddik, A., Kankanhalli, M.S. (2010) Multimodal Fusion for Multimedia Analysis: A Survey. Multimedia Systems, 16(6): 345-379. DOI: [10.1007/s00530-010-0182-0](https://doi.org/10.1007/s00530-010-0182-0)
8. Baltrusaitis, T., Ahuja, C., Morency, L.P. (2019) Multimodal Machine Learning: A Survey and Taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(2): 423-443. DOI: [10.1109/TPAMI.2018.2798607](https://doi.org/10.1109/TPAMI.2018.2798607)
9. Ordóñez, F.J., Roggen, D. (2016) Deep Convolutional and LSTM Recurrent Neural Networks for Multimodal Wearable Activity Recognition. Sensors, 16(1): 115. DOI: [10.3390/s16010115](https://doi.org/10.3390/s16010115)
10. Nagrani, A., Yang, S., Arnab, A., Jansen, A., Schmid, C., Sun, C. (2021) Attention Bottlenecks for Multimodal Fusion. Advances in Neural Information Processing Systems (NeurIPS) 34. DOI: [10.48550/arXiv.2107.00135](https://doi.org/10.48550/arXiv.2107.00135)
11. Perera, C., Zaslavsky, A., Christen, P., Georgakopoulos, D. (2014) Context Aware Computing for The Internet of Things: A Survey. IEEE Communications Surveys & Tutorials, 16(1): 414-454. DOI: [10.1109/SURV.2013.042313.00197](https://doi.org/10.1109/SURV.2013.042313.00197)
12. Sezer, O.B., Dogdu, E., Ozbayoglu, A.M. (2018) Context-Aware Computing, Learning, and Big Data in Internet of Things: A Survey. IEEE Internet of Things Journal, 5(1): 1-27. DOI: [10.1109/JIOT.2017.2773600](https://doi.org/10.1109/JIOT.2017.2773600)
13. Lin, J., Yu, W., Zhang, N., Yang, X., Zhang, H., Zhao, W. (2017) A Survey on Internet of Things: Architecture, Enabling Technologies, Security and Privacy, and Applications. IEEE Internet of Things Journal, 4(5): 1125-1142. DOI: [10.1109/JIOT.2017.2683200](https://doi.org/10.1109/JIOT.2017.2683200)
14. Ramachandram, D., Taylor, G.W. (2017) Deep Multimodal Learning: A Survey on Recent Advances and Trends. IEEE Signal Processing Magazine, 34(6): 96-108. DOI: [10.1109/SPM.2017.2683200](https://doi.org/10.1109/SPM.2017.2683200)

10.1109/MSP.2017.2738401

15. Ruan, S., Wang, R., Shen, X., Liu, H., Xiao, B., Shi, J., Zhang, K., Huang, Z., Liu, Y., Chen, E., He, Y. (2025) A Survey of Multi-sensor Fusion Perception for Embodied AI: Background, Methods, Challenges and Prospects. arXiv preprint. DOI: [10.48550/arXiv.2506.19769](https://doi.org/10.48550/arXiv.2506.19769)
16. Aguilera, A.A., Brena, R.F., Mayora, O., Molino-Minero-Re, E., Trejo, L.A. (2019) Multi-Sensor Fusion for Activity Recognition-A Survey. *Sensors*, 19(17): 3808. DOI: [10.3390/s19173808](https://doi.org/10.3390/s19173808)
17. D'Mello, S.K., Kory, J. (2015) A Review and Meta-Analysis of Multimodal Affect Detection Systems. *ACM Computing Surveys*, 47(3): 1-36. DOI: [10.1145/2682899](https://doi.org/10.1145/2682899)
18. Poria, S., Cambria, E., Bajpai, R., Hussain, A. (2017) A Review of Affective Computing: From Unimodal Analysis to Multimodal Fusion. *Information Fusion*, 37: 98-125. DOI: [10.1016/j.inffus.2017.02.003](https://doi.org/10.1016/j.inffus.2017.02.003)
19. Ma, Y., Zhou, G., Wang, S. (2019) WiFi Sensing with Channel State Information: A Survey. *ACM Computing Surveys*, 52(3): 1-36. DOI: [10.1145/3310194](https://doi.org/10.1145/3310194)
20. Wang, J., Chen, Y., Hao, S., Peng, X., Hu, L. (2019) Deep Learning for Sensor-based Activity Recognition: A Survey. *Pattern Recognition Letters*, 119: 3-11. DOI: [10.1016/j.patrec.2018.02.010](https://doi.org/10.1016/j.patrec.2018.02.010)
21. Chen, K., Zhang, D., Yao, L., Guo, B., Yu, Z., Liu, Y. (2021) Deep Learning for Sensor-based Human Activity Recognition: Overview, Challenges, and Opportunities. *ACM Computing Surveys*, 54(4): 1-40. DOI: [10.1145/3447744](https://doi.org/10.1145/3447744)
22. Cai, Z., Huang, M., Chen, M. (2026) Multimodal Sensor Data Fusion and Event Recognition in IoT. *Discover Artificial Intelligence*, 6(1): 1112. DOI: [10.1007/s44163-026-02189-w](https://doi.org/10.1007/s44163-026-02189-w)
23. Zhou, Z., Chen, X., Li, E., Zeng, L., Luo, K., Zhang, J. (2019) Edge Intelligence: Paving the Last Mile of Artificial Intelligence With Edge Computing. *Proceedings of the IEEE*, 107(8): 1738-1762. DOI: [10.1109/JPROC.2019.2918951](https://doi.org/10.1109/JPROC.2019.2918951)

10. Declarations

Funding

This literature synthesis was not supported by any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflict of Interest

The authors declare that they have no competing interests relevant to this work.

Data Availability Statement

This article is a literature synthesis and did not generate new primary data; all figures and results discussed are drawn from the cited publications, which readers should consult for their own underlying data and code availability statements.

Ethical Approval

Not applicable, as this work involved no human participants, animal subjects, or collection of primary data by the authors.