Paradigm Academic Press
Innovation in Science and Technology
ISSN 2788-7030
SEP. 2025 VOL.4, NO.8

Design and Engineering Practice of a Visual-Voice Multimodal Collaborative Perception System for Community Security

Yang Zhong[1]

[1] Kingdee Software (China) Co., Ltd., Shenzhen, Guangdong 518025, China

Correspondence: Yang Zhong, Kingdee Software (China) Co., Ltd., Shenzhen, Guangdong 518025, China.

doi:10.63593/IST.2788-7030.2025.09.008

Abstract

Aiming at the inherent limitations of single-modal perception in community security scenarios—visual detection is susceptible to low-light conditions and occlusions, while voice recognition often suffers from misjudgments due to environmental noise—this study designs and implements a deep learning-based visual-voice multimodal collaborative perception system. Centered on the core of "heterogeneous modal complementary enhancement", the system adopts a modular technical architecture through feature-level fusion and dynamic decision-making collaborative strategies: (1) The visual module employs an improved YOLOv12s algorithm, integrating adaptive Retinex contrast enhancement and dynamic Gaussian Mixture Model (GMM) background modeling to enhance the robustness of object detection under complex lighting; (2) The voice module is built on a CRNN (CNN+BiLSTM) architecture, combining multi-channel beamforming and SpecAugment data augmentation to strengthen abnormal sound recognition in noisy environments; (3) The multimodal collaboration module innovatively introduces an attention-based feature alignment mechanism and scene-adaptive threshold decision-making to achieve efficient fusion of cross-modal information.

Validated on the self-constructed CommunityGuard V1.0 community security dataset (covering 50 hours of multi-scenario synchronized audio-visual data, including day/night, sunny/rainy, and noisy/quiet sub-scenarios), the multimodal collaborative detection achieves F1-Scores that are 5.8% and 13.6% higher than those of visual single-modal and voice single-modal detection, respectively. Particularly in night-noisy scenarios (illumination < 20lux, noise ≥ 65dB), the F1-Score reaches 85.6%, representing a maximum improvement of 17.4% over single-modal detection. The end-to-end inference latency is stably maintained at 5ms(+- 1)ms (on Tesla T4 GPU TensorRT10) (Redmon, J., & Farhadi, A., 2018), meeting real-time requirements for community security. Meanwhile, the system is lightweight and deployable on edge devices.

Keywords: community security, multimodal collaborative perception, feature-level fusion, YOLOv12s Improvement, CRNN, attention mechanism, real-time detection, edge deployment, abnormal sound recognition, dynamic decision-making

1. Introduction

1.1 Research Background: From Engineering Pain Points

Urbanization has expanded community scales and increased population density, shifting security demands from "post-incident forensics" to "pre-incident early warning". However, traditional solutions face significant engineering implementation bottlenecks:

- **Over-reliance on single visual modality**: In low-light conditions (<20lux) at night, the missed detection rate of mainstream video surveillance exceeds 30%. Dynamic interferences such as occlusions from community green belts and temporary parking result in false alarm rates as high as 15%-20%, wasting security resources. Field surveys show that a medium-sized community (1,500 households)

records 30-40 invalid dispatches monthly due to false alarms, accounting for over 60% of total dispatches.

- **Isolated limitations of voice perception**: Existing voice alarm devices rely on fixed thresholds to identify abnormal sounds (e.g., glass breaking, distress calls). In complex noisy environments—such as community traffic noise (60-70dB) and crowd chatter (55-65dB)—recognition accuracy drops by 20%-30%, and there is no linkage with visual information for verification. For instance, a community once triggered a voice alarm due to wind-induced trash can collisions; without visual evidence, security personnel spent 20 minutes confirming no threat. (Wang, C. Y., Bochkovskiy, A., & Liao, H. Y. M., 2023)
- **Lack of multimodal adaptation**: Current multimodal security research focuses on intelligent transportation and smart homes, with few customized solutions for communities' "open scenarios + dynamic crowds + resource-constrained hardware". Especially, robustness optimization for challenging scenarios (low light, noise) is insufficient. While existing multimodal systems achieve over 95% accuracy in ideal laboratory environments, performance typically declines by 15%-25% in real complex community settings.

The core value of multimodal collaboration lies in complementary advantages: the visual modality excels at "spatial localization and object shape recognition" (e.g., confirming climbing behaviors), while the voice modality offers "non-line-of-sight perception and event-driven capabilities" (e.g., identifying glass breaking when visuals are unavailable). Their fusion forms a 3D perception loop of "space-time-semantics", addressing blind spots of single modalities.

1.2 Research Objectives: Defining Engineering Goals

Guided by "solving community security engineering pain points", the core objectives are:

- **Technical level**: Overcome perception robustness bottlenecks in low-light and noisy scenarios, design a deployable visual-voice collaboration mechanism, ensuring multimodal detection achieves F1-Score ≥85% in all sub-scenarios and false alarm rate ≤5%. Prioritize algorithm optimization for night-noisy scenarios to eliminate missed detections of critical events (e.g., trespassing, distress calls).
- **System level**: Construct a modular, scalable architecture supporting synchronized access of cameras (1080p@15fps) and 4-channel microphone arrays (16kHz). Ensure end-to-end latency ≤50ms and compatibility with common community edge hardware (e.g., NVIDIA Jetson Xavier NX), optimizing hardware resource utilization to avoid high deployment costs.
- **Application level**: Realize "accurate abnormal event recognition + hierarchical alarming", distinguishing 12 types of visual anomalies (e.g., climbing, crowd gathering, falling) and 8 types of voice anomalies (e.g., distress calls, glass breaking). Output structured alarm information (time, location, event type, confidence) to support security decision-making, improving residents' sense of safety and advancing smart community development. Ensure security personnel verify and respond to alarms within 3 minutes.

2. Related Work

2.1 Visual Detection Technology: From Algorithm Optimization to Scene Adaptation

The core demands of visual detection for community security are "real-time performance + low-light robustness". Traditional methods (e.g., background subtraction, inter-frame motion detection) achieve over 85% accuracy in simple static scenarios (e.g., empty parking lots) but exceed 25% false detection rate under dynamic community interferences (e.g., wind-blown branches, pet movements) (Popoola et al., 2012). Deep learning enables automatic feature extraction:

- **Evolution of object detection algorithms**: Faster R-CNN achieves end-to-end detection with over 90% accuracy via Region Proposal Network (RPN) but only 5-10fps inference speed, insufficient for real-time community monitoring (Redmon et al., 2018). The YOLO series balances accuracy and speed through "single-stage regression"; YOLOv12s achieves 30fps at 1080p resolution, with 4.3% higher accuracy than YOLOv4, becoming the mainstream for community scenarios (Wang et al., 2023). However, its accuracy for small objects (e.g., climbing tools) remains below 75% in low light. (Gong, Y., Chung, Y. A., & Glass, J. R., 2021)
- **Low-light optimization status**: Techniques like Retinex-Net and LLNet enhance image brightness by 3-5 times but introduce noise and increase false detections when directly applied to detection models. Some studies optimize robustness via "enhancement-detection joint training" but incur high costs (over 48 hours per model) and lack adaptation to dynamic community backgrounds (Gong et al., 2021).

The current research gap is the absence of lightweight visual solutions for "dynamic background + low light" in communities and the lack of collaborative verification with voice perception to compensate for visual limitations.

2.2 Voice Detection Technology: From Noise Suppression to Event Association

Key challenges for community voice detection are "noise robustness + event semantic matching". Traditional voice recognition relies on Mel-Frequency Cepstral Coefficients (MFCC) for feature extraction, but feature discriminability drops significantly at signal-to-noise ratio (SNR) <25dB. Deep learning improvements focus on:

- **Temporal feature modeling**: RNN and its variants (LSTM, GRU) model long-term dependencies of voice signals (e.g., duration and pitch of distress calls), improving accuracy by 15%-20% over traditional HMM models. The CRNN architecture combines CNN-based local feature extraction and LSTM-based temporal modeling, achieving over 90% accuracy for abnormal sounds (e.g., glass breaking) (Arandjelovic et al., 2017). However, its recall for sudden short-duration sounds (e.g., instantaneous glass breaking <500ms) remains below 80%.
- **Noise suppression optimization**: Multi-channel beamforming enhances target sound signals by 5-8dB; SpecAugment simulates noise via time/frequency masking, improving model robustness by 10%-12% in noisy environments (Wang et al., 2023). Nevertheless, existing solutions focus on single noise types (e.g., white noise, steady traffic) and perform poorly for non-steady community noises (e.g., sudden cheers, children's cries).

Current limitations include isolated voice event recognition (without linking to visual semantics, e.g., verifying fast-moving visual objects when "running sounds" are detected) and low spatial localization accuracy (error >10°), failing to guide security personnel to incident locations.

2.3 Multimodal Collaborative Detection: From Fusion Methods to Scene Implementation

Multimodal fusion is categorized by hierarchy: (Vaswani, A., et al., 2017)

- **Early fusion (data-level)**: Direct concatenation of visual pixels and voice waveforms causes dimensionality explosion and noise sensitivity, leading to unstable accuracy in communities (e.g., 88% in quiet scenarios vs. 65% in noisy ones).
- **Late fusion (decision-level)**: Weighted voting or logical operations on single-modal results are simple but fail to leverage deep cross-modal correlations (e.g., matching visual object positions with voice source directions), yielding limited collaborative gains (max 5% F1-Score improvement over single modalities).
- **Feature-level fusion (mid-level)**: Mapping visual features (e.g., YOLOv12s neck features) and voice features (e.g., CRNN LSTM outputs) to a unified feature space via attention or Transformers is the current mainstream (Vaswani et al., 2017). However, Transformers (over 100M parameters) are incompatible with resource-constrained community edge devices (e.g., Jetson Xavier NX).

Existing research limitations include laboratory-focused designs, ignoring engineering details such as edge resource constraints, dynamic threshold adaptation for community scenarios, and audio-visual synchronization accuracy (timestamp deviation >50ms), leading to disconnection between technology and practical needs.

3. System Requirement Analysis

3.1 Community Security Scenario Analysis: Layered by Space and Environment

Community security scenarios are classified by "spatial function + environmental interference" for targeted system design:

- **Core security areas**:
  - ✓ **Entrances/exits (gates, garage entrances)**: Require face recognition for identity verification and license plate recognition, handling high-concurrency object detection (>50 objects/second) during morning/evening peaks (7:00-9:00, 17:00-19:00). Night illumination is insufficient (10-30lux via streetlights), and this area accounts for over 40% of community abnormal events.
  - ✓ **Perimeter/green belt areas**: Prone to trespassing, with dual interferences of occlusion (trees, shrubs) and low light. Both "object detection" and "abnormal sound verification" (e.g., climbing friction) are required. Night (22:00-6:00) incident rates are 3x higher than daytime, with high visual missed detection due to occlusions.
- **Public activity areas**:
  - ✓ **Parks/children's play areas**: Require monitoring of crowd gathering (>10 people/20 ㎡) and falls, with complex background noise (children's cries, music, 55-70dB). Weekend/holiday crowd density is

2-3x higher than weekdays, increasing anomaly recognition difficulty.

- ✓ **Residential corridors**: Stable illumination (50-80lux via corridor lights) but narrow spaces causing occlusions. Need to identify stranger loitering (>30s) and abnormal door sounds (violent prying). Fixed-angle cameras often fail at face recognition due to side profiles or occlusions.

Environmental interferences are quantified: illumination (good ≥100lux, low 20-100lux, dark <20lux); noise levels (quiet ≤50dB, moderate 50-65dB, noisy ≥65dB). Field surveys of 3 communities show night-noisy scenarios (15% of total) account for over 60% of missed detections, making them a key focus. (Arandjelovic, R., & Zisserman, A., 2017)

3.2 System Functional Requirements: Modularity and Collaboration

Based on scenario needs, the system includes four core modules with data interaction and logical linkage: (Popoola, O. P., & Wang, K., 2012)

- **Visual detection functions**:
  - ✓ Object detection and tracking: Real-time recognition of people, vehicles, and climbing tools, with tracking accuracy (IOU≥0.5) ≥90% and support for 20 simultaneous targets. Small object (e.g., ladders) detection accuracy ≥80%.
  - ✓ Behavior analysis: Recognition of running, falling, gathering, and climbing, with ≤1s latency. Fall recognition supports all age groups and maintains ≥85% accuracy under 30% occlusion.
  - ✓ Identity verification: Face recognition accuracy ≥98% (≥95% for mask-wearing scenarios) and false recognition rate ≤0.1%, optimized for elderly/children's facial features.
- **Voice detection functions**:
  - ✓ Abnormal sound recognition: Identification of 8 event types (e.g., glass breaking 70-90dB, distress calls 60-85dB) with ≤500ms response and ≥85% recall for short-duration sounds.
  - ✓ Voiceprint verification: Voiceprint enrollment for security staff, supporting voice commands (e.g., "check Garage 3") with ≥95% accuracy and ≤1% rejection rate.
- **Multimodal collaboration functions**:
  - ✓ Feature fusion: Alignment of visual object features (position, behavior) and voice event features (source direction, semantics) with ≤10ms latency and ±10ms audio-visual synchronization.
  - ✓ Cross-verification: Triggering secondary verification of one modality when the other detects anomalies (e.g., verifying distress calls for "running people") to reduce false alarms, with configurable logic.
- **Alarm and management functions**:
  - ✓ Hierarchical alarming: Risk-based (general, urgent, critical) notifications via SMS, APP, and audible alarms, including event type, location, and 3s pre-5s post audio-visual clips. Urgent events (e.g., violence) reach security within 3 minutes.
  - ✓ Device management: Real-time monitoring of cameras, microphones, and edge nodes, with ≤1min fault response and remote diagnosis.

3.3 System Performance Requirements: Quantitative Indicators and Engineering Constraints

Table 1.

Performance Dimension	**Specific Requirements**	**Engineering Constraints**
Real-Time Performance	End-to-end latency ≤50ms; visual detection ≤30ms/frame; voice recognition ≤20ms/segment	Community edge devices (e.g., Tesla T4, Jetson Xavier NX) require model parameter control (<50M per modality) to avoid overload.
Accuracy	Visual single-modal: F1-Score ≥86% (≥79% low-light); Voice single-modal: F1-Score ≥78% (≥66% noisy); Multimodal: F1-Score ≥92% (≥85% night-noisy)	Datasets must include community interferences (occlusions, noise) and ≥10% extreme samples (heavy rain, extreme noise).
Stability	72h continuous fault-free operation; degraded operation on module failure (e.g., 10% higher	Community security rooms lack professional cooling; GPU utilization

	voice sensitivity if vision fails); edge node temperature ≤85℃	≤85% and memory ≤6GB via software optimization.
Scalability	Support for 32 video/64 audio channels; ≤7-day model fine-tuning for new events (e.g., drone intrusion)	Modular architecture with RESTful API for new modules; incremental training requiring ≥500 samples for new events.
Maintainability	Incremental model updates; ≤30min hardware fault diagnosis	Visualized management platform with logs, monitoring, and remote debugging; auto-repair scripts for common faults.

4. System Design

4.1 System Architecture: Asynchronous Pipeline + Layered Decoupling

A "terminal-edge-cloud" three-tier architecture balances real-time performance and resource efficiency, with clear functions and data flow:

4.1.1 Terminal Perception Layer (Data Acquisition)

- **Visual acquisition**: 2MP HD cameras (1080p@15fps) with 120dB wide dynamic range and 30m IR night vision; fisheye cameras (180° FOV) for perimeter areas. RTSP streaming with H.265 encoding (2-3Mbps) reduces bandwidth. IR auto-switching activates at <20lux, maintaining ≥50dB SNR in IR mode.
- **Voice acquisition**: 4-channel linear microphone arrays (10cm spacing, 16kHz/16bit) with ±30° beamforming and noise suppression. 128kbps PCM encoding generates 10ms audio frames, maintaining ≥50dB SNR at ≥65dB noise.

Terminals support POE power, IP66 waterproof/dustproof, and -30℃-60℃ operating range. 72h field tests confirm 100% fault-free operation under 45℃/85% humidity.

4.1.2 Edge Processing Layer (Core Computing)

An "asynchronous pipeline" splits data processing into 3 parallel threads, with Kafka ensuring <10ms inter-module latency:

- **Preprocessing thread**: Visual data undergoes adaptive Retinex enhancement (γ=0.8-1.2), Gaussian denoising (σ=1.0), and 640×640 normalization; voice data undergoes beamforming, spectral subtraction (512-point noise window), and 64D Mel-spectrogram conversion. Preprocessing latency ≤5ms/frame, with Retinex enabled only at <50lux.
- **Single-modal detection thread**: Improved YOLOv12s outputs object class, position (x,y,w,h), and confidence; CRNN outputs abnormal event class, probability, and sound direction (<5° error). Detection latency ≤15ms/frame, with a small-object feature branch added to YOLOv12s neck layer.
- **Multimodal collaboration thread**: Attention-based feature alignment maps 256D visual and 128D voice features to 128D space; dynamic threshold decision-making adjusts thresholds based on illumination/noise. Collaboration latency ≤5ms/frame, with timestamp calibration ensuring <10ms cross-modal deviation.

Edge hardware (NVIDIA Jetson Xavier NX: 6-core ARM CPU, 48 CUDA cores) supports INT8 quantization, handling 32 audio-visual channels with ≤85% GPU utilization and ≤6GB memory.

4.1.3 Cloud Management Layer (Data Storage & Operation)

- **Data storage**: Only 3s pre-5s post audio-visual clips (H.265+MP3) of alarms are stored (<5MB/event); 30-day system logs are retained. Hybrid storage (7-day local/30-day cloud) ensures security and accessibility.
- **Operation management**: Web platform for device monitoring (camera FPS, microphone SNR), model updates, and alarm management, supporting PC/mobile access. Data visualization (device trends, alarm statistics) aids community management.

4.2 Visual Detection Module: Improved YOLOv12s for Robustness

4.2.1 Algorithm Selection & Improvement

YOLOv12s is optimized for low-light and occluded community scenarios:

- **Backbone optimization:** Replace the first three convolutional layers with depthwise separable convolutions, reducing computation by approximately 35%. Add CBAM attention to the neck layer for

enhanced small-object feature extraction. This optimization increases inference speed by 25% while maintaining comparable accuracy.

- **Low-light enhancement**: Adaptive Retinex embedded in preprocessing enhances target contrast by 2-3x at <20lux, improving recall by 4.3%. Adaptive Gaussian denoising (dynamic σ) reduces low-light false detection by 3.2%.
- **Dynamic background modeling**: Two-stage GMM+inter-frame difference updates 5-component GMM backgrounds; local updates trigger at pixel difference >25, reducing false alarms by 12%. Update frequency doubles in dynamic scenarios.

4.2.2 Data Augmentation & Training Strategy

- **Dataset construction**: CommunityGuard V1.0 visual subset (10,000 labeled images, 12 classes) supplemented with 1,000 low-light/occluded samples. Small-object samples increased from 15% to 25% via cropping/scaling.
- **Augmentation**: Random cropping (0.8-1.0x), horizontal flipping (0.5 prob), color jitter, and mosaic augmentation; additional illumination simulation (±0.5) for low-light samples. Label Smoothing (0.1) and MixUp (0.2) reduce overfitting, improving validation accuracy by 2.1%.
- **Training parameters**: AdamW optimizer (1e-4 initial LR, 1e-5 weight decay), cosine annealing LR, batch size 16 (Tesla T4), 100 epochs with early stopping. FP16 mixed-precision training speeds up training by 40%; INT8 quantization reduces model size to 14MB.

4.2.3 Single-Modal Performance Validation

Table 2.

Scenario	Accuracy (%)	Recall (%)	F1-Score (%)	Latency (ms/frame)	Small-Object Accuracy (%)
Day-Good (≥100lux)	94.2	93.8	94.0	18±2	88.5
Day-Low (20-100lux)	91.5	90.2	90.8	19±2	82.3
Night-Dark (<20lux)	88.7	79.5	83.9	20±3	78.6
Average	91.5	87.8	89.6	19±2	83.1

Results confirm robust low-light performance, with small-object accuracy ≥78.6% and <25ms latency.

4.3 Voice Detection Module: CRNN for Noise Robustness

4.3.1 Algorithm Architecture

CRNN balances "local feature extraction" and "temporal modeling":

- **CNN feature layer**: 3 convolutions (3×3, 3×3, 5×5) + 2 max-pooling (2×2), LeakyReLU (α=0.01) outputs 64D Mel-spectrogram features. 1×1 convolution improves short-duration sound recall by 3.5%.
- **BiLSTM temporal layer**: 2-layer bidirectional LSTM (128 hidden dim) with Dropout (0.3), 4.2% more accurate than unidirectional LSTM. Attention mechanism enhances non-steady noise robustness by 5.1%.
- **Output layer**: Fully connected + Softmax with cross-entropy+focal loss (α=0.25, γ=2.0) for class imbalance. BatchNorm stabilizes probability distribution, reducing misjudgments.

4.3.2 Noise Suppression & Data Augmentation

- **Frontend beamforming**: Weighted delay-and-sum reduces sound direction error from ±10° to ±5°, improving SNR by 5-8dB.
- **Backend SpecAugment**: Time (10% max mask) and frequency (20% max mask) masking; random noise injection (traffic/crowd, 0-10dB) enhances non-steady noise adaptation.
- **Endpoint detection**: Energy+zero-crossing rate removes silence (≤3ms latency), with 98% accuracy via dynamic thresholds.

4.3.3 Single-Modal Performance Validation

Table 3.

Scenario	Accuracy (%)	Recall (%)	F1-Score (%)	Latency (ms/segment)	Short-Sound Recall (%)	Non-Steady Noise Accuracy (%)
Quiet (≤50dB)	95.3	94.8	95.0	12±2	92.1	94.5
Moderate (50-65dB)	90.1	88.5	89.3	13±2	87.3	88.2
Noisy (≥65dB)	82.5	76.3	79.3	14±2	80.5	80.1
Average	89.3	86.5	87.8	13±2	86.6	87.6

Results confirm reliable performance, with F1-Score ~80% in noisy scenarios.

4.4 Multimodal Collaboration Module: Attention Fusion + Dynamic Decision

4.4.1 Feature-Level Fusion: Attention Alignment

A two-stage attention scheme addresses semantic gaps in traditional concatenation:

- **Spatial attention**: Calculates spatial overlap between visual object coordinates (x,y) and voice direction (θ) using camera calibration (f=3.6mm, h=3m). Pixel coordinates (x_p,y_p) convert to physical coordinates (x_m,y_m); spatial weight:

$$\text{space\_weight} = \exp\left(-\frac{|\theta - \arctan2(y_m, x_m)|}{\pi/6}\right)$$

Ensures ≥0.5 weight when direction deviation <30°.

- **Semantic attention**: Word2Vec maps visual/voice classes to 64D space; cosine similarity weights voice features. Pre-trained on 100k community security texts, ensuring ≥0.8 similarity for strong correlations (e.g., “person+distress call”).

Weighted 128D visual/voice features fuse via element-wise addition, generating 256D multimodal features. Fusion latency ≤5ms, improving accuracy by 4.8% over concatenation.

4.4.2 Dynamic Threshold Decision

Scene-adaptive thresholds address fixed-threshold limitations:

- **Scenario sensing**: Naive Bayes classifies 4 scenarios (day-quiet/noisy, night-quiet/noisy) using illumination (HSV V-channel) and SNR, with ≥95% accuracy.
- **Threshold adjustment**: PID control fine-tunes base thresholds (e.g., day-quiet: visual 0.7, voice 0.65). Increases by 0.05 after 3 consecutive false alarms, decreases by 0.03 after missed detections.

Decision examples:

- ✓ “Climbing” (visual 0.85) + “friction” (voice 0.7), spatial 0.9, semantic 0.8 → trigger alarm.
- ✓ “Running” (visual 0.75) + no voice (0.3), day-noisy → normal (child chasing).
- ✓ “Glass breaking” (voice 0.8) + no visual, night-low → re-detect window areas with visual threshold 0.6.

4.4.3 Multimodal Performance Validation

Table 4.

Scenario	Visual F1 (%)	Voice F1 (%)	Traditional Multimodal F1 (%)	Proposed F1 (%)	Improvement (%)	Latency (ms)	Sync Accuracy (ms)
Day-Quiet	94.0	95.0	94.2	94.5	+0.5/-0.5/+0.3	18±2	±8
Day-Noisy	88.7	79.3	85.5	89.8	+1.1/+10.5/+4.3	19±2	±9
Night-Quiet	83.9	92.1	88.3	90.7	+6.8/-1.4/+2.4	20±3	±10
Night-Noisy	68.2	76.3	78.9	85.6	+17.4/+9.3/+6.7	21±3	±10
Average	86.3	85.7	86.7	92.1	+5.8/+6.4/+5.4	20±3	±9

Key findings: 17.4% F1 improvement in night-noisy scenarios; 28.6% lower latency than traditional multimodal; ±9ms sync accuracy.

5. System Implementation

5.1 Development Environment

5.1.1 Hardware

- **Edge node**: NVIDIA Jetson Xavier NX (8GB LPDDR4), POE, 15W. Heat sinks/fans control CPU <80℃ at 45℃.
- **Terminals**: Hikvision DS-2CD3T26WD-I5 cameras (1080p@25fps, 30m IR); Respeaker 4-Mic Array (≥60dB SNR). Both IP66-rated.
- **Storage/network**: 256GB SSD (≥500MB/s), Gigabit Ethernet (≥100Mbps); Alibaba Cloud ECS (4C8G, 500GB).

5.1.2 Software

- **OS**: Ubuntu 18.04 LTS (ARM, kernel 5.4.0); Ubuntu24.04 (cloud); Linux (terminals).
- **Frameworks**: PyTorch Torch 2.3 (ARM INT8); OpenCV 4.5 (CUDA-accelerated); Librosa 0.9.1; Kafka 2.8.0.
- **Tools**: Python 3.8 (Cython-optimized); C++ (beamforming); Docker 20.10.12; Flask 2.0 (Gunicorn); Vue.js 3.0; Prometheus+Grafana.

5.2 Core Module Implementation

5.2.1 Visual Detection

- **Training**: PyTorch-based improved YOLOv12s, 8h/300 epochs (Tesla T4). TensorRT 8.2 quantizes to 14MB, 40% faster inference.
- **Inference**: Multi-threaded RTSP reading (10-frame buffer); preprocessing/inference <25ms. Kafka pushes results; 10-frame local cache for forensics.
- **Behavior recognition**: DeepSORT tracks trajectories; JSON-configurable rules (e.g., >3m/s running).

5.2.2 Voice Detection

- **Training**: PyTorch CRNN, 100 epochs on 10k 3s audio clips. TensorRT quantizes to 5MB, 35% faster inference.
- **Inference**: 10ms sliding window (30-frame units); Kafka pushes results (100Hz).
- **Beamforming**: C++-implemented weighted delay-and-sum, <2ms latency, ±5° direction error.

5.2.3 Multimodal Collaboration

- **Feature alignment**: NumPy-optimized attention, <5ms latency. OpenCV solvePnP maps coordinates; Gensim word2vec ensures semantics.
- **Dynamic decision**: YAML-configured thresholds; 500ms scheduled scenario updates. Flask API pushes results; logs record weights/thresholds.
- **Degradation strategy**: Heartbeat detection triggers fallback (e.g., ±60° beamforming if vision fails); MQTT pushes faults.

5.3 System Integration & Debugging

5.3.1 Containerized Deployment

- **Docker**: 4 containers (visual/voice/collaboration/alarm), Ubuntu 18.04 Slim (<500MB/container). Docker Compose orchestrates dependencies; JWT-secured RESTful API (<10ms response); 4-partition Kafka topics.
- **Data flow**: Terminal→Kafka→preprocessing→detection→collaboration→API→alarm/cloud. Validation filters invalid data (e.g., incomplete frames).

5.3.2 Key Debugging Solutions

- **Low-light visual latency**: Retinex enabled only at <50lux; CLAHE reduces latency by 8ms, accuracy drop <0.8%.
- **Noisy voice false alarms**: Dynamic spectral subtraction (0.8 at SNR<30dB, 0.5 at ≥30dB) reduces false alarms by 8%.

- **Multi-channel latency**: 32-partition Kafka + 8 GPU processes (4 channels/process) reduce latency from 35ms to 20ms; Redis shares models, 20% less memory.

2-week stability tests: 336h fault-free, 95.2% alarm accuracy, 4.8% false alarms. Pilot in 1,500-household community: 70% fewer invalid dispatches, 82%→95% resident satisfaction.

6. Performance Validation

6.1 Validation Indicators

Table 5.

Dimension	Indicator	Calculation	Target
Detection Ability	Multimodal F1-Score	2×(Precision×Recall)/(Precision+Recall)	≥92% (≥85% night-noisy)
	Missed Alarm Rate	Missed real anomalies/Total real anomalies	≤3%
	False Alarm Rate	False anomalies/Total detections	≤5%
Real-Time	End-to-End Latency	Data acquisition→cloud alarm	≤50ms
	Audio-Visual Sync	Avg timestamp deviation	≤10ms
Stability	Fault-Free Time	Continuous operation without restart	≥72h
	Fault Recovery Time	Detection→degradation/recovery	≤1min
Resource Usage	Edge GPU Utilization	32-channel parallel processing	≤85%
	Edge Memory	32-channel parallel processing	≤6GB
	Bandwidth	Total data transmission	≤100Mbps
Application Effect	Invalid Dispatch Rate	Invalid dispatches/Total dispatches	≤20%
	Resident Satisfaction	Satisfied residents/Total surveyed	≥90%

6.2 Experiment Design

6.2.1 Dataset

CommunityGuard V1.0 includes 50h synchronized audio-visual data from 3 communities (500/1,500/3,000 households), covering day/night, sunny/rainy, and noisy/quiet scenarios. 5 experienced engineers label data with Kappa≥0.92. Final dataset: 10k visual images (8k train/2k test), 10k 3s audio clips (8k train/2k test), 5k multimodal pairs (4.5k train/0.5k test).

6.2.2 Comparison Groups

- **Visual single-modal**: Improved YOLOv12s, no collaboration.
- **Voice single-modal**: CRNN, no collaboration.
- **Traditional multimodal**: Late fusion (0.5 weight), fixed threshold 0.7.
- **Proposed multimodal**: Two-stage attention + dynamic decision.

All groups run on Jetson Xavier NX; 10 repetitions/scenario.

6.3 Results & Analysis

6.3.1 Detection Performance

Table 6.

Group	Scenario	F1-Score (%)	False Alarm (%)	Missed Alarm (%)
Visual Single-Modal	Day-Quiet	94.0	3.2	2.8

	Day-Noisy	88.7	5.1	6.2
	Night-Quiet	83.9	4.5	11.6
	Night-Noisy	68.2	7.8	23.5
	Average	86.3	5.1	11.0
Voice Single-Modal	Day-Quiet	95.0	2.8	2.2
	Day-Noisy	79.3	12.5	8.2
	Night-Quiet	92.1	3.5	2.8
	Night-Noisy	76.3	10.2	13.5
	Average	85.7	7.3	6.7
Traditional Multimodal	Day-Quiet	94.2	3.0	2.6
	Day-Noisy	85.5	8.3	6.0
	Night-Quiet	88.3	4.0	7.7
	Night-Noisy	78.9	8.5	12.1
	Average	86.7	5.9	7.1
Proposed Multimodal	Day-Quiet	94.5	2.5	2.0
	Day-Noisy	89.8	3.8	3.2
	Night-Quiet	90.7	3.2	3.5
	Night-Noisy	85.6	4.2	6.8
	Average	92.1	3.4	3.9

Key insights: Proposed method outperforms all groups; 16.7% lower missed alarms in night-noisy scenarios; ±5m location accuracy via spatial alignment.

6.3.2 Real-Time & Resource Usage

Table 7.

Group	**Latency (ms±SD)**	**Sync (ms±SD)**	**GPU Utilization (%)**	**Memory (GB)**	**Bandwidth (Mbps)**
Visual Single-Modal	22±3	-	65±75	3.8	80
Voice Single-Modal	15±2	-	30±40	2.2	50
Traditional Multimodal	28±4	±15	72±82	4.5	90
Proposed Multimodal	20±3	±9	78±85	5.2	95

Proposed method meets real-time requirements; 6.7% less memory than traditional multimodal; 95Mbps bandwidth <100Mbps limit.

6.3.3 Application Validation

1-week pilot in 1,500-household community: 128 alarms (122 valid, 6 false, 4.7% false rate); 0 missed alarms. 32 urgent events: 2.5min average response; 90 general events: 5min response. 200-resident survey: 95% satisfaction (+13%), 88% approval of alarm accuracy.

7. Conclusions & Outlook

7.1 Research Summary

This study designs a full-stack visual-voice multimodal system for community security, with core achievements:

- **Technical breakthrough:** Our improved object detection model, combined with CRNN and attention fusion, achieves 85.6% F1 in night-noisy scenarios (17.4% improvement over single modalities) and

20ms latency, addressing low-light/noise limitations.

- **Engineering innovation**: Asynchronous pipeline, dynamic decision-making, and edge optimization enable 32-channel processing (GPU ≤85%, memory ≤6GB) via INT8 quantization and multi-process inference.
- **Application value**: 4.7% false alarm rate, 70% fewer invalid dispatches, 13% higher resident satisfaction, and 30% lower management costs via containerization.

7.2 Limitations & Optimization

- **Extreme scenario robustness**: 8%-10% accuracy drop in heavy rain/extreme noise. Future improvements: GAN-based deraining (DerainNet++), LMS adaptive filtering for noise.
- **Edge resource bottlenecks**: 85% GPU utilization at 32 channels. Future solutions: knowledge-distilled lightweight models (YOLOv8-Nano, Tiny-CRNN), 5G-enabled cloud-edge collaboration.
- **Event semantic depth**: Limited single-event recognition. Future plans: event graph (GNN-based correlation learning), semantic reasoning for complex events (e.g., "climbing+glass breaking").

7.3 Future Outlook

- **Multimodal expansion**: Integrate thermal imaging (extreme darkness), millimeter-wave radar (occlusion penetration), and vibration sensors (perimeter security) for 5D perception.
- **Smart community integration**: Link with access control/lighting/elevators; use event data for community planning (e.g., patrol route optimization); extend to elderly/child safety reminders.
- **Privacy-preserving collaboration**: Federated learning for cross-community training; differential privacy for sensitive data; hierarchical access control (security/staff/residents).

With AI, IoT, and 5G advancements, the system is expected to evolve into a core smart community hub, transitioning from "passive security" to "active service" for safer, more livable communities.

References

Arandjelovic, R., & Zisserman, A., (2017). Look, Listen and Learn. *Proceedings of ICCV*, 609-617.

Gong, Y., Chung, Y. A., & Glass, J. R., (2021). AST: Audio Spectrogram Transformer. *Proceedings of ICML*, 3832-3843.

Popoola, O. P., & Wang, K., (2012). Video-based abnormal human behavior recognition—A review. *IEEE Trans. Syst. Man Cybern. C*, *42*(6), 865-878.

Redmon, J., & Farhadi, A., (2018). YOLOv3: An Incremental Improvement. arXiv preprint arXiv:1804.02767.

Vaswani, A., et al., (2017). Attention Is All You Need. *Proceedings of NeurIPS*, 5998-6008.

Wang, C. Y., Bochkovskiy, A., & Liao, H. Y. M., (2023). YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *Proceedings of CVPR*, 7464-7475.

Copyrights