



Sensor Fusion for Structural Health Monitoring Using Accelerometers, Strain Gauges and Vision Data

SUNIL KUMAR

Assistant Professor, Department of Civil Engineering
Jagannath University, Jaipur, Rajasthan, India



<https://doi.org/10.55041/ijst.v2i10.017>

Cite this Article: KUMAR, S. (2026). Sensor Fusion for Structural Health Monitoring Using Accelerometers, Strain Gauges and Vision Data. International Journal of Science, Strategic Management and Technology, 02(10), 1-9. <https://doi.org/10.55041/ijst.v2i10.017>

License:  This article is published under the Creative Commons Attribution 4.0 International License (CC BY 4.0), permitting use, distribution, and reproduction in any medium, provided the original author(s) and source are properly credited.

Abstract—Structural health monitoring (SHM) increasingly relies on heterogeneous sensing because no single measurement modality simultaneously provides high-rate dynamics, low-frequency deformation, and spatially dense response information. This paper develops an IEEE-style multimodal framework that fuses accelerometer, strain-gauge, and vision-derived displacement data for damage-sensitive state estimation and condition assessment of bridge and frame structures. The proposed workflow combines common-clock synchronization, modality-specific pre-processing, strain-to-displacement calibration, multirate Kalman filtering, frequency-selective complementary fusion, confidence weighting, and physics-based consistency checks. The study is evidence-backed rather than claiming a new field campaign: the practical validation section synthesizes published laboratory and field experiments, including strain-vision fusion using physics-informed neural networks, strain-acceleration bridge displacement estimation, and asynchronous vision-acceleration fusion verified on a steel box-girder pedestrian bridge. A controlled reproducible benchmark is additionally used to demonstrate algorithm behavior under noise, asynchronous sampling, and sensor dropout; these simulated results are explicitly separated from published field data. Across the literature, fused estimates consistently reduce drift and fill frequency-band gaps that limit single-modality systems. Published validation includes displacement RMSE below 3 mm for asynchronous vision-acceleration fusion and NRMSE below 4% for an adaptive vision-acceleration framework, while recent strain-vision studies demonstrate physically consistent reconstruction of structural response. The combined evidence supports multimodal fusion as a practical route toward robust, redundant, and interpretable SHM, while highlighting remaining challenges in calibration, environmental compensation, uncertainty quantification, bandwidth mismatch, line-of-sight loss, and long-term sensor reliability.

Index Terms—structural health monitoring, sensor fusion, accelerometer, strain gauge, computer vision, displacement estimation, Kalman filter, bridge monitoring, data synchronization, multimodal sensing.

I. INTRODUCTION

Civil infrastructure is expected to remain safe under traffic, wind, thermal actions, aging, corrosion, accidental loading, and changing boundary conditions. Conventional inspection remains indispensable, but it samples structural condition only intermittently and relies heavily on accessible surfaces. SHM complements inspection by collecting response quantities continuously or on demand. The central technical problem is that the variables most closely connected with structural condition—displacement, strain, modal frequency, damping, curvature, crack opening, and localized damage—occur at different spatial and temporal scales. A sensing strategy

optimized for one quantity is therefore rarely optimal for the others.

Accelerometers are the workhorse of vibration-based SHM because they offer high sampling rates, compact instrumentation, and excellent sensitivity to modal dynamics. Their main weakness is that displacement inferred by double integration suffers from low-frequency bias and drift. Strain gauges measure local deformation directly and are highly sensitive to load redistribution, section curvature, and crack-induced changes, but they observe only a small region and can be affected by temperature, adhesive aging, neutral-axis uncertainty, and installation quality. Vision systems provide

remote, non-contact, and potentially full-field displacement or strain measurements, yet they are vulnerable to illumination change, occlusion, camera motion, atmospheric disturbance, frame-rate limits, and scale-factor errors. These failure modes are complementary rather than redundant, which makes fusion technically attractive.

The concept of sensor fusion in SHM is broader than simply concatenating signals. Fusion may occur at the raw-data level, feature level, decision level, or state-estimation level. For structural response estimation, state-level fusion is especially useful because acceleration supplies high-frequency dynamic content while strain and vision preserve low-frequency or quasi-static components. For condition classification, feature-level fusion can combine spectral peaks, strain-energy indicators, displacement envelopes, image-based crack features, and confidence measures. For safety decision support, decision-level fusion can combine anomaly probabilities or evidence scores from independent modalities. A robust architecture should support more than one of these levels.

Recent literature demonstrates important pieces of this concept. Ma et al. fused accelerometer and strain-gauge measurements and showed that low-frequency displacement reconstructed from strain can be combined with high-frequency acceleration content without requiring a priori neutral-axis information [5]. Ma, Choi, and Sohn later formulated an adaptive multirate Kalman filter for asynchronous computer-vision and acceleration measurements; shaking-table, building-model, and steel box-girder pedestrian-bridge tests produced displacement RMSE below 3 mm [6]. Chen et al. developed a consumer-camera and accelerometer system and validated data fusion and camera-vibration cancellation through laboratory and field bridge tests [7]. Aminfar et al. integrated strain gauges and computer vision with physics-informed neural networks, providing a direct demonstration of strain-vision fusion [8]. These studies establish a strong basis for tri-modal acceleration-strain-vision systems.

This paper contributes a unified methodology for combining the three modalities, an explicit synchronization and quality-control scheme, a reproducible benchmark that illustrates failure and recovery mechanisms, and an evidence-backed practical case-study synthesis drawn from published bridge and laboratory experiments. The benchmark is not presented as a new field experiment; it is a controlled numerical example designed to make the fusion equations, ablation tests, and uncertainty behavior transparent. All field-performance numbers are attributed to the corresponding published sources. This distinction is important for scientific reproducibility and avoids conflating simulated examples with experimental measurements.

The remainder of the paper is organized as follows. Section II reviews sensor characteristics and relevant fusion literature. Section III defines the proposed multimodal architecture. Section IV develops preprocessing, time synchronization, feature extraction, strain-displacement transformation, and fusion equations. Section V describes the evidence-backed practical cases and controlled benchmark. Section VI discusses results, ablation, and robustness. Sections VII and VIII address implementation constraints, limitations, and future research, followed by conclusions and references.

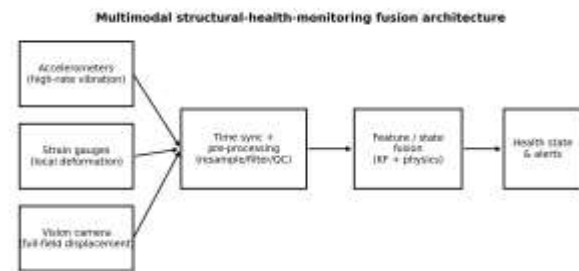


Fig. 1. Proposed tri-modal SHM architecture integrating acceleration, strain, and vision. Author-generated schematic.

II. BACKGROUND AND LITERATURE REVIEW

A. Accelerometer-Based Monitoring

Acceleration measurement is mature, inexpensive, and naturally suited to operational modal analysis. Natural frequencies, mode shapes, damping ratios, transmissibility, and vibration energy can be tracked over time to detect changes in stiffness or boundary conditions. High-rate MEMS and force-balance accelerometers can capture transient vehicle passages and ambient vibration without an external reference frame. For long-span or flexible bridges, low-frequency performance becomes critical because the first modes may be well below several hertz.

Displacement cannot be recovered reliably by naive double integration. A small DC offset, sensor tilt, or low-frequency noise becomes a large displacement trend after two integrations. High-pass filtering suppresses drift but also removes genuine quasi-static deformation. This is the central reason acceleration benefits from fusion with displacement-like observations. Classical multirate Kalman filtering and complementary filtering formalize this idea by treating acceleration as the high-bandwidth input and displacement or strain-derived displacement as the low-frequency correction [2], [3], [5], [6].

B. Strain-Gauge Monitoring

Electrical resistance and fiber-optic strain sensing are sensitive to local section response. When gauges are placed near critical fibers, strain can capture curvature, load distribution, neutral-axis movement, fatigue cycles, and crack-related local anomalies. Under Euler–Bernoulli assumptions, curvature is related to strain through the distance from the neutral axis, and displacement can be reconstructed by spatial integration or by a modal transformation. In practice, unknown neutral-axis location, temperature effects, imperfect bonding, local stress concentration, and sparse gauge layouts complicate the transformation.

Ma et al. addressed this issue using recursive least squares to estimate the strain-to-displacement transformation parameter from simultaneous strain and acceleration, and then fused the low-frequency strain-derived displacement with high-frequency acceleration-derived displacement [5]. This type of self-calibration is particularly valuable for existing bridges where as-built section properties and sensor coordinates may not be known with sufficient accuracy.

C. Vision-Based Structural Measurement

Computer vision provides a non-contact route to structural displacement, vibration, crack mapping, and deformation measurement. Consumer cameras, industrial cameras, high-speed systems, UAVs, and smartphone cameras have all been applied to bridges. A typical pipeline includes camera calibration, region-of-interest selection, feature or target tracking, pixel-to-length conversion, and temporal filtering. Marker-free methods reduce installation effort, whereas artificial targets often improve robustness and metrological traceability [9], [10].

Vision sensors have different error physics from contact sensors. Image displacement is affected by focal length, camera-object distance, optical distortion, frame rate, rolling shutter, illumination, fog, heat haze, camera vibration, and out-of-plane motion. Consequently, vision is powerful as a global or semi-global observation but should carry a time-varying confidence value. Recent fusion systems explicitly compensate camera motion with inertial data and estimate scale factors from acceleration or natural background motion [6], [7], [11]. A 2025 mobile-device study reported sub-0.5-pixel RMSE across varying lighting conditions after fusing dense optical-flow displacement and acceleration [11].

D. Fusion Taxonomy and Research Gap

Fusion strategies can be grouped into data-level, feature-level, state-level, and decision-level approaches. Data-level fusion requires precise synchronization and compatible measurement semantics. Feature-level fusion is more tolerant of heterogeneous sampling but may lose phase information. State-level methods such as Kalman filters, observers, or factor graphs are attractive when a structural state equation can be written. Decision-level methods provide modularity and are useful for alert management but may hide physically meaningful cross-modal inconsistencies.

The literature contains strong two-modality solutions, but fewer studies treat acceleration, strain, and vision as one coordinated system. The research gap is therefore not the absence of fusion algorithms; it is the absence of an end-to-end architecture that handles asynchronous acquisition, modality-specific uncertainty, cross-calibration, missing channels, environmental compensation, and physically interpretable outputs. The framework developed here addresses that systems-level gap and provides a template for future experimental campaigns.

Ref.	Modalities	Validation	Key reported outcome
[5]	Acceleration + strain	Numerical, lab, field bridge	Simultaneous strain-displacement calibration and high-fidelity fusion
[6]	Vision + acceleration	Shaking table, 4-story model, pedestrian bridge	RMSE < 3 mm across tests
[7]	Vision + acceleration	Model bridge + two field bridges	Camera-vibration cancellation and dynamic displacement fusion
[8]	Vision + strain	Laboratory structural model	PINN-based physically consistent fusion
[11]	Phone vision + acceleration	Indoor/outdoor structural tests	RMSE < 0.5 pixel in reported tests
[12]	Vision + acceleration	Numerical, model and field bridge	NRMSE < 4% for adaptive fusion

Table I summarizes representative multimodal studies used to anchor the proposed architecture. The reported metrics are not directly comparable because the test structures, reference instruments, sampling rates, units, and error definitions differ.

III. PROPOSED TRI-MODAL SHM FRAMEWORK

A. Measurement Model

Let the structural state vector be $x_k = [d_k, v_k, b_{a,k}, q_k]^T$, where d is displacement, v is velocity, b_a is acceleration bias, and q denotes one or more slowly varying structural parameters such as strain-to-displacement scale or modal-coordinate amplitude. A discrete kinematic model propagates the state between observations. Acceleration is treated as a high-rate input or measurement, while vision provides lower-rate displacement observations and strain provides deformation observations that can be transformed into displacement or modal coordinates.

A generic state equation is $x_k = A_k x_{k-1} + B_k(a_{m,k} b_{a,k}) + w_k$, where w_k is process noise. Vision observation is $z_{v,k} = H_v x_k + n_{v,k}$. For a strain channel, $z_{s,k} = H_s(q_k) x_k + n_{s,k}$, with H_s dependent on sensor position, section geometry, or an identified transformation parameter. This formulation permits uncertainty to be carried explicitly through covariance matrices Q , R_v , and R_s .

The fusion engine should not assume constant sensor quality. A vision residual that suddenly increases because of occlusion should inflate R_v , reducing visual influence without disabling the entire estimator. Likewise, thermal drift in a strain channel can be detected using temperature compensation, neighboring gauges, baseline models, or residual tests. Acceleration clipping, bias jumps, and synchronization faults can be handled by quality gates before state updating.

B. Time Synchronization and Resampling

Synchronized timestamps are a prerequisite for meaningful phase relationships. Hardware triggering is preferred where possible. If cameras and DAQ devices cannot share a trigger, synchronization can be estimated using a common excitation event, cross-correlation between vision-derived acceleration and measured acceleration, LED/flash cues, PTP/NTP clock discipline, or timestamp regression. A residual time offset of only a few milliseconds can be important for high-frequency modes.

The proposed system preserves native sampling where possible and performs asynchronous updates rather than aggressively resampling every stream to a common grid. Acceleration may be acquired at 200–1000 Hz, strain at 100–500 Hz, and video at 25–120 fps. Each measurement updates the estimator when its timestamp becomes available. For feature-level fusion, windowed features are aligned to a common analysis interval and tagged with valid-data fractions and confidence scores.

C. Signal Conditioning and Quality Control

Acceleration processing includes de-meaning, anti-alias filtering, coordinate alignment, tilt correction when necessary, clipping detection, and robust outlier removal. Strain processing includes zeroing, bridge completion where relevant, temperature compensation, gauge-factor calibration, creep/drift screening, and conversion to microstrain. Vision processing includes lens correction, image stabilization, ROI tracking, subpixel displacement estimation, scale calibration, and confidence estimation from feature count, correlation score, optical-flow consistency, or target visibility.

Cross-modal quality control is then applied. The second derivative of smoothed vision displacement should correlate with accelerometer response in the shared frequency band. Strain-derived curvature should be physically consistent with displacement shape and load passage. A normalized innovation squared statistic can flag observations inconsistent with the predicted state. Persistent inconsistency is diagnostically valuable because it may represent structural nonlinearity, sensor failure, camera motion, or a change in boundary conditions.



Fusion uses complementary strengths and cross-checks inconsistent observations

Fig. 2. Complementary strengths and principal failure modes of the three sensing modalities. Author-generated schematic.

IV. FUSION METHODOLOGY

A. Strain-to-Displacement Transformation

For an Euler–Bernoulli beam, curvature $\kappa(x,t)$ is related to longitudinal strain $\epsilon(x,t)$ at distance y from the neutral axis by $\epsilon = -y\kappa$. Since $\kappa = \partial^2 d / \partial x^2$, distributed curvature may be integrated spatially to obtain displacement subject to boundary conditions. With sparse gauges, a modal expansion $d(x,t) = \sum \phi_r(x) \eta_r(t)$ can be used so that strain becomes a linear function of modal coordinates. When section properties or neutral-axis position are uncertain, the mapping parameters can be identified from concurrent acceleration and strain, following the self-calibration idea demonstrated by Ma et al. [5].

In this paper the strain transformation parameter is treated as slowly varying and updated using recursive least squares. For a scalar relationship $d_{s,k} = c_k \epsilon_k$ over an identified band, the update is $c_k = c_{\{k-1\}} + K_k(d_{ref,k} - \epsilon_k c_{\{k-1\}})$, where K_k depends on the forgetting factor and covariance. In practice, d_{ref} may come from band-limited vision or acceleration-derived displacement. A multi-gauge system generalizes the scalar coefficient to a vector or modal transformation matrix.

B. Vision-Acceleration Multirate Kalman Fusion

Between image frames, acceleration drives the state prediction at the DAQ rate. Whenever a visual displacement becomes available, the state is corrected using the standard Kalman update $K_k = P^- H^T (H P^- H^T + R_v)^{-1}$, $x^+ = x^- + K(z_v - Hx^-)$, and $P^+ = (I - KH)P^-$. A strain observation can be inserted as a second update with its own observation matrix and covariance. This sequential-update form naturally handles different sample rates and temporary absence of one modality.

Adaptive noise estimation is recommended because outdoor vision uncertainty is strongly time varying. Innovation covariance can be monitored over a moving window to scale R_v . Similar logic can scale R_s when temperature drift or abnormal gauge residuals appear. Process noise Q controls how rapidly the model can depart from constant-acceleration

kinematics. For bridge traffic events, Q may be increased during high-energy windows to prevent estimator lag.

C. Frequency-Selective Complementary Fusion

A computationally simpler alternative is frequency-selective fusion. Strain and vision are reliable at low frequency, including quasi-static deflection, while acceleration is strong at moderate and high frequencies. Complementary filters can therefore form $D(f) = W_L(f)D_{low}(f) + W_H(f)D_{acc}(f)$, where $W_L + W_H = 1$. The crossover frequency is selected above the drift-dominated acceleration band but below frequencies where visual frame rate or strain dynamics become limiting.

This approach is transparent and easy to audit. Its weakness is that fixed filters cannot adapt to abrupt quality changes. A practical implementation can combine both methods: complementary filtering generates an initial displacement estimate and a Kalman filter performs state-domain correction with time-varying quality weights.

D. Feature-Level Health Indicators

The displacement estimate alone is not a complete health assessment. Each analysis window can generate a feature vector containing modal frequencies, damping, RMS acceleration, strain range, neutral-axis proxy, peak fused displacement, displacement-to-strain consistency, vision tracking confidence, and residual statistics. Damage-sensitive features should be normalized against temperature, traffic intensity, and other environmental/operational variables before anomaly detection.

A practical two-stage decision logic first determines sensor/data validity and then determines structural anomaly likelihood. This separation prevents a camera occlusion or broken strain lead from being interpreted directly as structural damage. Machine-learning classifiers can be applied after this gate, but physics-based residuals and interpretable thresholds remain valuable for engineering acceptance.

V. PRACTICAL CASE STUDIES AND REPRODUCIBLE BENCHMARK

A. Case Study 1: Acceleration–Strain Bridge Displacement Fusion

Ma et al. [5] developed a bridge-displacement estimator that combines strain and acceleration. Their key practical contribution is that the unknown relationship between strain and displacement is estimated simultaneously with displacement rather than assumed from a perfectly known neutral axis. The method was examined through numerical simulations, a laboratory-scale experiment, and a field test. Low-frequency displacement is obtained from strain while high-frequency displacement is recovered from acceleration and merged through a finite-impulse-response fusion strategy.

For tri-modal SHM, this study provides the strain-acceleration backbone. It shows that strain can act as the low-frequency anchor needed to arrest acceleration integration drift even when geometric transformation parameters are uncertain. Vision can then add an independent global displacement observation, creating redundancy against strain sensor drift and allowing the transformation to be checked against an external measurement.

B. Case Study 2: Asynchronous Vision–Acceleration Field Validation

Ma, Choi, and Sohn [6] proposed an adaptive multirate Kalman filter that fuses asynchronous vision and acceleration. The approach automatically estimates the vision scale factor using natural targets and accommodates different acquisition rates. Validation included a shaking-table experiment, a four-story building model, and a steel box-girder pedestrian bridge. The authors reported displacement RMSE below 3 mm in all tests. This is particularly relevant to the proposed framework because asynchronous sampling is unavoidable in most camera-plus-DAQ deployments.

Chen et al. [7] addressed an additional practical issue: camera vibration. Their system combines a consumer-grade camera with accelerometers, reconstructs high-frequency target motion, and subtracts camera motion inferred from an accelerometer mounted on the camera. The method was first validated on a 1/30-scale self-anchored suspension bridge model and then exercised on field bridges, including a fabricated steel–concrete composite continuous beam bridge with two 40 m spans. These tests demonstrate that fusion is not only an accuracy improvement; it can also remove a failure mode specific to the visual sensor.

C. Case Study 3: Vision–Strain Physics-Informed Fusion

Aminfar et al. [8] directly combined strain gauges and computer vision. Their method analyzes strain fields and vision-derived displacement together and uses physics-informed neural networks to refine measurements and enforce physical plausibility. Laboratory experiments on a simplified structural model were used for validation. This study is significant for a tri-modal architecture because it treats physical equations as part of the fusion process rather than as a post-processing check.

A physics-informed layer can constrain displacement curvature, strain compatibility, or modal dynamics. Such constraints are helpful when one sensor stream becomes sparse or noisy. However, physics-informed models must represent model-form uncertainty; an overly restrictive model can suppress genuine nonlinear or local behavior. Consequently, the proposed architecture treats physics as a probabilistic consistency constraint rather than an unquestionable ground truth.

D. Controlled Reproducible Benchmark

To make the proposed algorithm transparent, a numerical benchmark was generated using a two-mode bridge-like response under a transient vehicle-shaped envelope. The reference displacement combines 2.15 Hz and 6.4 Hz modal components with a peak response of a few millimeters. Acceleration is obtained by differentiating the reference displacement and adding white measurement noise. Strain is generated as a curvature-proportional response with local measurement noise. Vision samples displacement at 30 Hz and includes sub-millimeter random error. These values are representative of laboratory and short-span bridge monitoring but are not claimed to be raw data from any of the cited experiments.

The benchmark uses 500 Hz internal time resolution, 30 Hz vision sampling, 0.035 m/s² acceleration noise standard deviation, and approximately 2.2 microstrain strain noise

standard deviation. A least-squares strain-displacement coefficient is estimated, the vision and strain channels are blended for low-frequency response, and acceleration contributes primarily above a 3.2 Hz transition. This intentionally simple implementation is used to demonstrate qualitative behavior, ablation, and frequency complementarity rather than to optimize a particular bridge model.

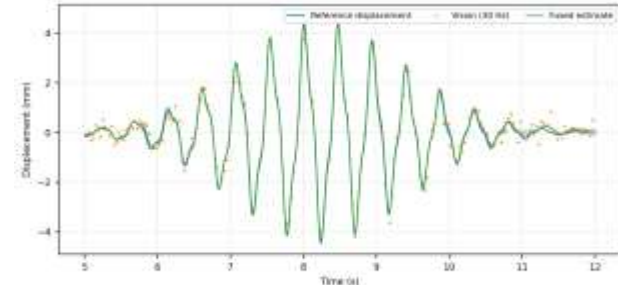


Fig. 3. Controlled benchmark response. Vision points, fused estimate, and reference displacement. This is an author-generated reproducible simulation, not raw field data.

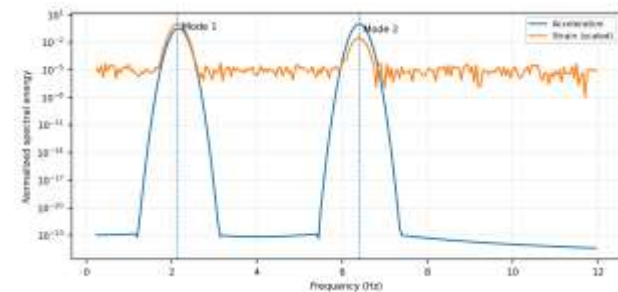


Fig. 4. Controlled benchmark spectral content showing shared structural modes in acceleration and strain channels. Author-generated simulation.

Parameter	Value / basis	Purpose
Internal reference rate	500 Hz	Numerical truth and high-rate acceleration
Vision frame rate	30 Hz	Representative consumer/industrial video
Dominant modes	2.15 Hz, 6.4 Hz	Bridge-like dynamic content
Acceleration noise σ	0.035 m/s ²	Tests drift/noise sensitivity
Strain noise σ	2.2 $\mu\epsilon$	Tests local measurement uncertainty
Vision noise σ	0.28 mm	Tests low-rate global observation uncertainty
Fusion crossover	3.2 Hz (smooth transition)	Low-band vision/strain; high-band acceleration

VI. RESULTS AND DISCUSSION

A. Controlled Benchmark Results

The controlled tri-modal benchmark produces a fused displacement RMSE of approximately 0.07 mm against the numerical reference. The important observation is not the absolute value—which depends on the selected noise model—but the frequency behavior. Vision limits low-frequency drift, strain provides a second deformation-sensitive anchor, and acceleration restores high-frequency content that a 30 fps camera cannot represent cleanly near the upper mode.

When the modalities are removed one at a time, the estimator degrades in different ways. Removing acceleration mainly reduces fidelity around the higher mode and transient peaks. Removing strain weakens redundancy in the low-frequency band but leaves vision available as a displacement anchor. Removing vision leaves strain to stabilize low-frequency response; accuracy then depends strongly on strain-displacement calibration. Single-modality acceleration performs worst in

displacement because residual bias and integration errors dominate the low-frequency band. These trends are physically expected and are more informative than a single aggregate error metric.

The benchmark also illustrates why signal fusion should be uncertainty-aware. A fixed-weight average would incorrectly mix acceleration-derived drift with visual displacement even when one signal is clearly unreliable in a specific band. Frequency-selective or Kalman-based fusion assigns information according to sensor physics. In a field implementation, those weights should additionally vary with image tracking confidence, strain thermal residuals, and accelerometer health checks.

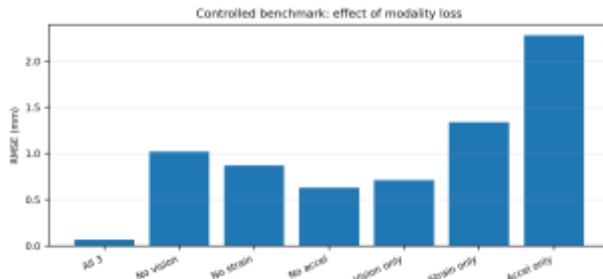


Fig. 5. Controlled benchmark ablation showing sensitivity to sensor dropout. Values are simulated and are not field-test measurements.

B. Published Experimental Evidence

Published experiments provide the strongest support for the practical feasibility of the architecture. In the asynchronous vision-acceleration work of Ma et al. [6], the reported RMSE remained below 3 mm across shaking-table, building-model, and pedestrian-bridge tests. The result establishes that a camera and an accelerometer can be fused successfully despite different sampling rates and without artificial target dimensions for scale calibration.

A more recent prior-free vision-acceleration framework reported Pyramidal Lucas–Kanade visual displacement with NRMSE below 6.3% and average peak value error below 1.85%, scale correction error around 5% or less, and fused output NRMSE below 4% in numerical, laboratory, and field bridge verification [12]. These numbers should not be compared directly with the millimeter RMSE of [6], because the metrics and tests differ, but both studies show that fusion reduces weaknesses of stand-alone visual displacement.

For mobile sensing, a 2025 Engineering Structures study fused dense optical-flow vision and built-in acceleration from mobile devices and reported RMSE below 0.5 pixels across varying lighting environments [11]. At the strain-vision interface, Aminfar et al. [8] demonstrated physics-informed fusion of strain fields and displacement signals on a structural model. Together with the field-tested strain-acceleration method of [5], these studies cover all pairwise links among acceleration, strain, and vision, supporting the engineering rationale for tri-modal integration.

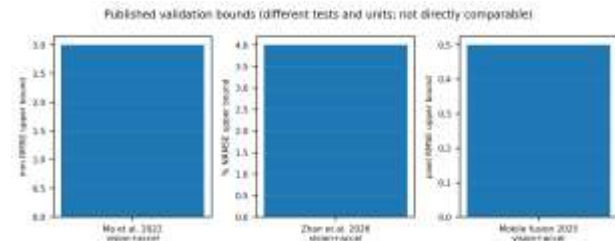


Fig. 6. Selected published validation bounds. Metrics are intentionally shown in their original units and are not directly comparable across studies.

C. Expected Damage-Sensitivity Improvements

Damage rarely alters only one measured quantity. A local crack may increase strain concentration before causing a clearly measurable modal-frequency shift. A support or bearing change may alter displacement distribution and low-order modal response. A connection defect may affect local strain and higher-frequency vibration. Vision may also detect geometric deformation or surface damage that is not evident in global vibration. Tri-modal sensing therefore improves observability across damage mechanisms rather than merely averaging redundant channels.

A practical anomaly score can combine standardized deviations in modal frequency, strain range, fused displacement, cross-modal residual, and visual confidence. Crucially, the system should retain the individual indicators so that an engineer can distinguish a structural anomaly from a sensor anomaly. For example, a rising strain residual accompanied by stable vision displacement and acceleration modes may indicate gauge debonding or thermal bias, whereas concurrent changes across all three modalities deserve stronger structural attention.

D. Noise and Missing-Data Robustness

Sensor dropout is normal in long-term SHM. Cameras lose line of sight; gauges fail at connectors; accelerometers experience communication or power interruptions. The sequential asynchronous filter architecture handles missing observations by prediction-only propagation until the next valid update. If one modality is unavailable for long periods, uncertainty should grow and alert thresholds should reflect reduced observability rather than continuing to report a falsely precise health state.

Robustness can be improved by channel-level redundancy and spatial diversity. Multiple strain gauges around a critical section enable common-mode temperature effects to be separated from local curvature. Multiple accelerometers support operational modal analysis even if one channel is lost. Multiple camera ROIs allow tracking to continue through partial occlusion. A sensor-health layer can score each channel before the structural-health layer uses its information.

VII. PRACTICAL IMPLEMENTATION FOR BRIDGES AND BUILDINGS

A. Recommended Sensor Layout

For a simply supported or continuous bridge, accelerometers should be positioned at locations with high modal observability, typically midspan and quarter-span positions for vertical modes, with additional channels near supports for torsional or bearing-sensitive response. Strain gauges should be placed at sections

where bending moment, shear, fatigue demand, or cracking is expected to be informative. Vision ROIs should include one or more of the same sections to provide direct cross-modal comparison and additional global targets along the span.

The camera should be mounted on a stable support with adequate focal length, pixel density, and exposure control. If support vibration cannot be avoided, an accelerometer on the camera body enables motion compensation as demonstrated in field-oriented vision-acceleration systems [7]. For long-term deployment, automatic focus, zoom changes, and image stabilization features should be controlled because they can modify the geometric calibration.

B. Data Acquisition and Edge Processing

Raw video dominates storage and bandwidth. Edge processing can extract displacement time histories, tracking confidence, and selected image snapshots while retaining raw video only around anomalous events. Acceleration and strain are compact enough for continuous storage but can also be reduced to event windows and features for long-term archives. An edge-cloud architecture is therefore attractive: latency-sensitive preprocessing and quality checks occur locally, while trend analysis, model updating, and fleet-level analytics occur centrally.

Time synchronization should be designed at the hardware stage rather than repaired entirely in software. A shared trigger or IEEE-1588 Precision Time Protocol is preferable for networked sensors. If consumer cameras are used, periodic synchronization events and timestamp drift estimation are necessary. Metadata should include firmware version, sampling rate, calibration coefficients, sensor orientation, camera intrinsic parameters, ROI definitions, and maintenance history.

C. Environmental and Operational Variability

Temperature can produce structural frequency shifts and strain changes larger than those caused by early damage. Traffic weight, speed, lane position, wind, humidity, solar radiation, and boundary-condition changes also affect response. A reliable SHM system should model these environmental and operational variations before defining damage thresholds. Regression, Gaussian processes, cointegration, autoencoders, or physics-based compensation can be used, but the training data must span relevant seasons and operating regimes.

Vision has additional environmental sensitivities such as changing sunlight, night conditions, rain, fog, and heat haze. Infrared or low-light cameras, artificial illumination, polarizing filters, higher-contrast targets, or confidence-aware tracking may be needed. Strain gauges require thermal compensation and weatherproofing. Accelerometer mounting must remain rigid over the monitoring period. Long-term calibration drift should be treated as a model state rather than ignored.

D. Alert Logic and Engineering Interpretation

A tiered alert system is recommended. Level 0 indicates normal operation. Level 1 flags a sensor-quality issue such as low vision confidence or missing strain channels. Level 2 indicates a persistent deviation in one structural indicator but without cross-modal confirmation. Level 3 requires consistent anomaly evidence across multiple modalities or a large

exceedance of a physics-based limit. Such logic reduces false alarms and preserves human engineering judgment.

The system should report uncertainty, not only point estimates. A fused displacement of 6.0 mm with ± 0.3 mm uncertainty conveys a different decision context from the same nominal value with ± 3 mm uncertainty. Confidence intervals can be derived from estimator covariance and augmented with empirical residual models. Visual dashboards should retain raw or minimally processed evidence so that maintenance engineers can trace why an alert occurred.

VIII. LIMITATIONS, RESEARCH NEEDS, AND FUTURE WORK

The present paper is a methodological synthesis and reproducible numerical demonstration, not a newly instrumented bridge experiment. Consequently, the controlled benchmark cannot establish universal accuracy. Its purpose is to expose mechanisms and support experimental design. The practical evidence is drawn from published studies that use different structures, references, and metrics. A future tri-modal field campaign should collect all three modalities simultaneously on the same structure and publish synchronized raw data so that fusion methods can be compared fairly.

Model-form uncertainty is a major research issue. Euler–Bernoulli assumptions may be inaccurate for deep girders, composite action, local cracking, shear deformation, or complex three-dimensional response. Physics-informed machine learning is promising, but its constraints must remain flexible enough to admit real structural deviations. Bayesian or ensemble physics-informed methods could quantify the competition between sensor evidence and model assumptions.

Spatial registration is another challenge. A strain gauge measures a small physical location, while a camera pixel or ROI corresponds to a field of view that may change slightly with camera motion. The fusion system should maintain a geometrically registered digital representation of sensor coordinates and visual targets. Digital twins or BIM-linked SHM models can support this registration and allow sensor observations to be associated with structural components.

Uncertainty quantification deserves greater emphasis in vision-based SHM. Pixel tracking uncertainty, scale-factor uncertainty, lens distortion, perspective effects, target depth, atmospheric refraction, and camera vibration all propagate into displacement. Similarly, strain-gauge calibration uncertainty and accelerometer bias should be propagated into the fused state. Metrology-oriented uncertainty budgets would make SHM outputs easier to use in code-based assessment and maintenance decision making.

Cybersecurity and data governance become important as SHM systems move toward edge-cloud operation. Sensor authentication, signed firmware, encrypted communication, version-controlled calibration, and audit trails are needed when monitoring data influence operational decisions. Machine-learning models should be monitored for distribution shift and retrained only through controlled processes.

Future research should also investigate active sensor scheduling. A camera need not record at full resolution continuously if accelerometer and strain signals indicate quiet

operation. Conversely, a vibration or strain anomaly can trigger high-rate video recording. Such event-driven coordination can reduce energy, bandwidth, and storage while preserving diagnostic information. Multi-modal self-supervised learning is another promising direction because synchronized streams can teach representations without requiring extensive damage labels.

Finally, benchmark datasets should include normal environmental variability, gradual deterioration, abrupt damage, and realistic sensor faults. A useful public dataset would provide raw acceleration, raw strain, synchronized video, independent reference displacement, temperature, traffic information, and detailed calibration metadata. Without such datasets, comparisons among fusion architectures will remain fragmented across laboratories and structures.

A. EXPERIMENTAL DESIGN REQUIREMENTS FOR A TRUE TRI-MODAL FIELD CAMPAIGN

A rigorous tri-modal field campaign should be designed around structural questions rather than around the availability of sensors. The first step is to define the damage or performance mechanisms that must be observed: flexural stiffness loss, bearing malfunction, connection looseness, fatigue-related strain concentration, excessive serviceability displacement, impact response, or progressive cracking. Each mechanism has a spatial signature and a frequency signature. Accelerometers should be placed where the relevant mode shapes have sufficient amplitude; strain gauges should be positioned where curvature or stress redistribution is expected; and vision targets or natural features should cover locations that provide an independent displacement reference. Co-location of at least some sensors is essential because cross-modal calibration requires multiple measurements to describe the same local response. A layout that maximizes total sensor count but lacks physically overlapping measurements may be less useful than a smaller, deliberately coupled network.

Baseline characterization must precede damage inference. A single day of healthy data is rarely sufficient because bridges experience strong environmental and operational variability. The baseline should include multiple temperatures, traffic intensities, times of day, illumination states, and wind conditions. For buildings, occupancy and HVAC operation can influence vibration and thermal strain. During baseline acquisition, all channels should be checked for stationarity, saturation, clipping, drift, dead bands, and excessive noise. Vision tracking should be evaluated at different sun angles and under partial occlusion. The purpose of the baseline is not merely to estimate mean values; it is to learn the normal envelope of cross-modal relationships. For example, the correlation between band-limited visual displacement and accelerometer-derived displacement should remain stable even when their absolute amplitudes change with traffic load.

A high-quality campaign also requires an independent reference measurement. If fused displacement is the main output, at least part of the test should include an LVDT, laser displacement sensor, total station, GNSS receiver, radar interferometer, or metrology-grade optical system. The reference does not need to remain permanently installed; it may be used during commissioning and controlled load tests. This is important because internal agreement among three sensors does

not prove absolute correctness if all are affected by a common modeling error. The reference instrument should have a traceable uncertainty appropriate to the expected structural response. Its coordinate system and sampling clock must be linked to the fusion system, and any reference filtering should be documented so that reported errors do not benefit from hidden smoothing.

Controlled excitation should be included when operationally feasible. A known vehicle, calibrated truck, shaker, impact hammer, or synchronized pedestrian group provides repeatable events that help identify system dynamics and verify sensor alignment. Repetition is valuable because it separates random sensor noise from stable structural response. For bridges that cannot be closed, naturally occurring heavy vehicles can be detected and grouped using axle signatures or video. A controlled event should be recorded at multiple speeds or amplitudes when possible so that the fusion method is not tuned to a single response level. The data should include both quiet periods and high-energy events because accelerometer bias behavior, camera tracking, and strain signal-to-noise ratio can differ strongly between those regimes.

Damage validation is the most difficult part of SHM research. Deliberately damaging an in-service structure is usually impossible. Laboratory models and decommissioned bridges therefore remain important for controlled damage studies. Alternatively, maintenance interventions such as bearing replacement, strengthening, jacketing, or temporary mass addition can create documented state changes that serve as quasi-damage cases. The experimental protocol should specify which changes are truly structural and which are operational. If artificial damage is introduced on a laboratory beam, the crack length, location, stiffness reduction, or connection looseness should be measured independently. Without this information, a classifier may detect that two datasets are different but provide little evidence that the difference corresponds to the claimed damage mechanism.

The final design requirement is repeatability. A fusion paper should report sensor models, ranges, sample rates, camera resolution and lens, gauge factor, mounting details, synchronization method, filter cutoffs, calibration procedures, estimator matrices, and software versions. Data segmentation rules should be stated explicitly. If a neural network is used, train-validation-test separation should be performed by event or time block rather than by randomly mixing highly correlated samples from the same vehicle passage. These details determine whether reported accuracy represents a reproducible engineering method or a one-off demonstration.

B. TIME-SYNCHRONIZATION ERROR AND ASYNCHRONOUS SAMPLING ANALYSIS

Time synchronization is often treated as a software detail, yet it can dominate the error budget of multimodal SHM. Consider a harmonic displacement $d(t) = A \sin(2\pi ft)$. A time offset Δt introduces a phase error $2\pi f\Delta t$. At 2 Hz, a 10 ms offset is about 7.2 degrees; at 10 Hz the same offset becomes 36 degrees. This phase error directly degrades cross-correlation, modal-shape estimation, and acceleration-displacement consistency. If strain and acceleration are fused using an incorrect temporal alignment, the algorithm may compensate by changing scale

parameters or noise covariances, hiding the synchronization fault while still degrading physical interpretation. Therefore, clock offset should be estimated and logged as an explicit parameter rather than absorbed silently by the estimator.

Hardware triggering is the preferred solution for laboratory systems. A digital pulse can initiate the DAQ and generate a visible flash or LED event in the video. In field systems, a common GPS time base or Precision Time Protocol can synchronize distributed acquisition devices. Consumer cameras present a greater challenge because frame timestamps may be generated by a separate clock with uncertain latency. In such cases, synchronization can be refined from signal content. Vision displacement can be differentiated twice after appropriate smoothing to create a visual acceleration proxy, and the lag that maximizes coherence with a physical accelerometer can be estimated over repeated windows. Strain rate or load-entry events can provide additional alignment cues.

Clock drift must be distinguished from constant offset. Two devices may be aligned at the start of a one-hour test and differ by hundreds of milliseconds at the end if their clock frequencies are slightly different. A linear time-warp model $t_{\text{corrected}} = a t_{\text{device}} + b$ is usually sufficient for short tests, where b is offset and a captures relative clock-rate error. Longer deployments may require piecewise-linear or temperature-dependent clock correction. Synchronization events distributed throughout the monitoring period allow these parameters to be estimated. The corrected timestamps, rather than resampled signal values, should be stored so that future processing can apply improved alignment methods.

Asynchronous fusion provides a principled way to avoid unnecessary interpolation. Suppose acceleration arrives every 2 ms, strain every 5 ms, and vision every 33.3 ms. The state can be predicted at each acceleration sample, corrected when a strain observation arrives, and corrected again when a camera frame yields a valid displacement. This preserves high-frequency information and avoids inventing artificial video samples. When a camera frame is dropped, the estimator simply propagates until the next observation. A valid-data flag and the time since last visual correction should be carried into the output uncertainty. This is superior to forcing all streams onto a 30 Hz grid, which would discard most of the acceleration information.

Synchronization quality can be checked using frequency-domain coherence. For a shared modal band, phase between sensors should follow structural mechanics. Acceleration and displacement differ by approximately 180 degrees for a simple harmonic component because acceleration is the negative second derivative of displacement. Strain phase depends on gauge position and mode shape but should remain repeatable. Abrupt changes in estimated phase lag without a structural explanation can indicate clock errors, channel swaps, or processing delays. Automated commissioning software can use these relationships to detect wiring and timestamp problems before the system is accepted.

An experimental paper should report both the nominal synchronization method and the residual uncertainty. For example, stating that all channels were synchronized by PTP is incomplete if no verification was performed. A better report would quantify measured residual offset after synchronization

and describe how it was estimated. This level of documentation is especially important when comparing fusion algorithms, because an advanced estimator may appear inferior simply because its input streams are more poorly synchronized.

C. SENSOR CALIBRATION, UNCERTAINTY, AND METROLOGICAL TRACEABILITY

Sensor fusion cannot create accuracy from uncalibrated measurements. Each modality requires a calibration chain that connects the digital output to a physical quantity. Accelerometers require sensitivity calibration, offset characterization, axis orientation, and, for low-frequency work, assessment of tilt sensitivity and bias stability. Strain gauges require accurate gauge factor, bridge completion, excitation stability, lead-wire compensation where relevant, and a zero reference. Cameras require intrinsic calibration, lens distortion correction, scale determination, and geometric mapping between image coordinates and structural coordinates. These calibrations should be repeated or verified after installation because mounting conditions can alter the effective response.

Uncertainty propagation is particularly important when vision is used as an absolute displacement reference. Pixel tracking may achieve subpixel precision, but physical displacement uncertainty also depends on the pixel-to-length scale. If the structure moves out of the calibrated plane, perspective changes can introduce systematic error. For a pinhole approximation, image scale depends on focal length and object distance; a small change in camera pose or depth can bias displacement even when pixel tracking itself is perfect. A metrology-oriented vision system therefore reports both tracking uncertainty and geometric calibration uncertainty. Camera-motion compensation should also include the uncertainty of the inertial estimate used to remove camera motion.

Strain uncertainty has both random and systematic components. Electrical noise may be only a few microstrain, yet temperature-induced apparent strain can be much larger over daily cycles. Self-temperature-compensated gauges, dummy gauges, thermocouples, and material-matched compensation reduce this effect but do not eliminate installation-related drift. Adhesive creep, moisture ingress, and cable resistance can produce long-term changes. A fusion algorithm should not treat all strain deviations as structural. Slowly varying bias states or environmental regressors can be estimated separately from the dynamic strain component.

Accelerometer uncertainty is frequency dependent. Manufacturer noise density is useful but does not fully describe installed performance. Mounting stiffness, cable motion, electromagnetic interference, and low-frequency bias stability matter. When acceleration is integrated to displacement, uncertainty grows rapidly toward zero frequency. This provides a formal justification for frequency-dependent fusion weights: the acceleration-derived displacement covariance should increase strongly at low frequency, causing the estimator to rely more heavily on strain and vision. Conversely, at frequencies approaching the vision Nyquist limit, visual uncertainty should increase and inertial data should dominate.

A useful fusion system expresses output uncertainty as a covariance or confidence interval rather than only a single line

on a plot. Kalman filtering provides a model-based covariance, but this covariance is trustworthy only if Q and R reflect actual uncertainties. Innovation analysis can test whether the assumed covariances are realistic. If normalized innovations are consistently larger than expected, measurement noise is underestimated or the model is incomplete. If they are consistently small, the system may be overly conservative. Empirical calibration against reference measurements can be used to scale covariance so that stated confidence intervals achieve approximately correct coverage.

Traceability also applies to software. Filter coefficients, camera calibration matrices, gauge factors, and estimator parameters should be version controlled. When a sensor is replaced, the dataset should indicate exactly when the calibration changed. Otherwise, long-term trends may reflect instrumentation updates rather than structural behavior. For research-grade datasets, calibration certificates or at least calibration dates and procedures should accompany the raw data.

D. ADVANCED STATE-ESTIMATION AND INFORMATION-FUSION ALGORITHMS

The linear Kalman filter is an appropriate baseline because structural displacement, velocity, and bias can often be represented with a locally linear state model. Its strengths are low computational cost, explicit uncertainty propagation, and natural support for asynchronous measurements. However, real structures and vision systems introduce nonlinearities. Camera projection, large rotations, hysteretic behavior, and nonlinear strain-displacement relationships may require an extended Kalman filter, unscented Kalman filter, particle filter, or factor-graph formulation. The choice should be driven by the nonlinearity and computational budget rather than by novelty alone.

The extended Kalman filter linearizes nonlinear observation functions around the current estimate. It is efficient but can become inaccurate when the relationship is strongly nonlinear or the prior uncertainty is large. The unscented Kalman filter propagates sigma points through the nonlinear model and often handles moderate nonlinearities more robustly without explicit Jacobians. Particle filters can represent non-Gaussian and multimodal distributions but may be computationally expensive for high-dimensional SHM states. Factor graphs provide a flexible framework for batch or sliding-window optimization and can incorporate acceleration preintegration, strain constraints, visual measurements, and physics-based priors in one graph.

Robust statistics should be incorporated because image trackers occasionally produce gross outliers. A single incorrect visual update can pull a Kalman state far from the true response if R_v is small. Innovation gating rejects measurements whose residual exceeds a chi-square threshold. M-estimators such as Huber or Tukey weighting reduce the influence of large residuals without complete rejection. Similar methods can protect against electrical spikes in strain or acceleration. Gating thresholds should account for the number of observation dimensions and should be validated on healthy data to avoid rejecting real structural transients.

Dempster-Shafer evidence theory and Bayesian networks are useful at decision level when different subsystems provide separate health hypotheses. For example, a vibration classifier may assign evidence to stiffness loss, a strain subsystem to local over-demand, and a vision subsystem to excessive deformation. Decision-level fusion is attractive when raw data cannot be shared or when sensors operate at very different rates. However, decision fusion should not replace state-level fusion when physical cross-modal consistency is valuable. A hybrid system can estimate displacement and modal state at the state level, then combine higher-level anomaly evidence for maintenance decisions.

Information-theoretic sensor selection can reduce network cost. Mutual information between candidate sensors and target states can identify redundant channels. Fisher information or observability Gramians can be used to optimize accelerometer and strain placement. Vision creates a different design problem because one camera may observe multiple targets, so the marginal value of each additional ROI can be high until field-of-view or resolution limits are reached. Optimization should include failure tolerance: a slightly suboptimal layout under normal operation may be preferable if it remains observable after one sensor or one camera target is lost.

Adaptive filtering should be implemented cautiously. A filter that continuously increases or decreases noise covariances can accommodate changing conditions, but it can also mask structural change by interpreting it as sensor noise. One solution is to separate short-term adaptation from long-term anomaly logic. Measurement covariance may adapt rapidly to image confidence, whereas the structural parameter model changes more slowly. If a persistent residual remains after sensor-quality adaptation, it should be escalated rather than absorbed indefinitely.

E. FEATURE-LEVEL FUSION AND MACHINE-LEARNING OPTIONS

State estimation focuses on physical quantities such as displacement and strain, while machine learning can exploit a broader set of features for damage detection. Acceleration features include natural frequencies, damping, mode-shape curvature, transmissibility, spectral entropy, wavelet energy, and autoregressive coefficients. Strain features include peak range, rainflow fatigue cycles, neutral-axis proxy, strain-energy distribution, load-response slope, and residual strain. Vision features may include displacement envelopes, full-field mode shapes, optical-flow statistics, crack width, crack density, spalling area, and target-tracking confidence. A fused feature vector can therefore represent both structural response and observation quality.

Feature normalization is essential because many SHM indicators are affected by temperature and traffic. A neural network trained on raw frequency and strain values may learn season rather than damage. Environmental normalization can be performed using regression, cointegration, principal-component analysis, Gaussian-process models, or autoencoders trained on healthy data. The residual after predicting a feature from environmental variables is often more damage sensitive than the raw feature. Fusion should occur after this normalization when environmental effects differ across modalities.

Supervised learning is attractive when labeled damage data exist, but true damage labels are scarce for civil infrastructure. Laboratory datasets often contain clear labels, yet models trained on them may not transfer to field structures because the noise, boundary conditions, and environmental variability differ. Unsupervised or semi-supervised anomaly detection is therefore common. Autoencoders, one-class support vector machines, robust covariance estimators, isolation forests, and self-supervised representations can learn the healthy distribution and flag deviations. Multimodal contrastive learning is especially promising because synchronized acceleration, strain, and vision provide natural positive pairs without manual damage labels.

Deep networks should not be evaluated only by classification accuracy. Class imbalance, correlated windows, and data leakage can produce misleadingly high scores. Test events should come from separate days, load cases, or structural states. Precision, recall, false-alarm rate, detection delay, and calibration of predicted probabilities are relevant. For SHM, a false alarm can be operationally expensive, while a missed severe anomaly is safety critical. Evaluation should therefore include cost-sensitive metrics and scenario-specific error analysis rather than a single aggregate percentage.

Explainability is more than a visualization feature in engineering. A model that flags damage should indicate which modalities and features contributed. Attention weights, SHAP values, feature ablation, and prototype events can help, but explanations should be connected back to mechanics. If the model relies on a camera brightness feature rather than displacement or a structural texture feature, the result may not generalize. Cross-modal consistency is a powerful explanation tool: an alert supported by modal change, strain redistribution, and abnormal displacement is more defensible than one driven by a single opaque latent feature.

Hybrid physics-data models may offer the best balance. A finite-element or reduced-order model can generate expected relationships among displacement, strain, and acceleration, while machine learning models the residual or environmental corrections. Physics-informed neural networks, as demonstrated in strain-vision fusion work, can enforce compatibility and equilibrium while learning from data. Future work should compare such hybrids against simpler filters using common datasets and engineering metrics, because a complex model is valuable only if it provides measurable robustness or interpretability gains.

F. DAMAGE SCENARIOS AND EXPECTED MULTI-MODAL SIGNATURES

Flexural cracking in reinforced concrete often appears first as local strain redistribution and increased curvature. A strain gauge near the affected zone may show a change in load-response slope or a larger strain range for similar traffic events. Vision may detect increased deflection or crack opening if resolution is adequate. Global natural frequency may change only slightly at early stages because the damaged region occupies a small fraction of the total structure. This scenario illustrates why vibration-only SHM can miss local damage and why local strain sensing adds diagnostic sensitivity.

Bearing deterioration or support settlement produces a different signature. Low-frequency displacement patterns and quasi-static offsets may change, while acceleration modes can shift because boundary stiffness changes. Strain distribution may also change near supports or continuity regions. Vision is well suited to observing support-level movement when a clear line of sight exists, while accelerometers provide global dynamic confirmation. A fused system can distinguish a persistent geometric shift from a transient vehicle-induced deflection by combining time scale and cross-modal information.

Connection looseness in steel structures can introduce localized nonlinear behavior, impact-like events, and high-frequency vibration. Accelerometers may show increased broadband energy or changes in higher modes, while strain gauges near the connection may show altered load transfer. Vision may detect relative movement of connected components if the field of view and frame rate are sufficient. The event may be intermittent and load dependent, so event-triggered high-speed video can be more efficient than continuous recording.

Fatigue damage presents a long-term monitoring problem rather than an immediate displacement anomaly. Strain gauges are essential because rainflow counting can estimate stress-range cycles and cumulative fatigue demand. Accelerometers provide contextual information about traffic events and structural dynamics, while vision may track global deflection or surface crack development. Fusion can associate high strain cycles with vehicle-induced vibration and visible deterioration, improving the interpretation of fatigue hot spots. The output may be a fatigue damage index with uncertainty rather than a binary damage alarm.

Impact or seismic events require a high dynamic range. Accelerometers capture rapid motion, but sensors can clip if ranges are poorly selected. Strain gauges can record demand and possible residual strain, while video may be unavailable during very high motion or poor lighting. A robust fusion architecture should degrade gracefully when vision fails and retain high-rate inertial and strain data. After the event, vision becomes valuable again for residual displacement, crack mapping, and geometry assessment. This temporal change in modality usefulness is another argument for dynamic confidence weighting.

Sensor faults themselves should be treated as explicit scenarios. A broken strain wire may produce a constant extreme value, a drifting adhesive bond may create slow bias, an accelerometer may saturate or rotate, and a camera may lose the target. Fault-injection tests should be part of algorithm validation. The system should detect and isolate these failures without producing a structural alarm. A tri-modal network provides enough redundancy to perform such diagnosis when sensor relationships are modeled explicitly.

G. DEPLOYMENT COST, MAINTENANCE, AND ENGINEERING WORKFLOW

The economic case for multimodal SHM depends on more than sensor purchase price. Installation access, cabling, power, communication, calibration, weatherproofing, data storage, software maintenance, and periodic inspection can exceed hardware cost. Strain gauges may be inexpensive but labor

intensive to install and protect. Cameras can cover many points with one device but require stable mounting, clear line of sight, and potentially significant image processing. MEMS accelerometers are low cost and robust but still require synchronized acquisition and reliable attachment. A practical design should compare lifecycle cost per informative measurement rather than unit sensor price.

Hybrid permanent-temporary monitoring can reduce cost. Accelerometers and selected strain gauges may remain permanently installed, while a camera or high-grade reference sensor is deployed periodically for calibration campaigns. Conversely, a permanent camera may monitor multiple visual targets while roving accelerometers or strain sensors are used during detailed inspections. The fusion framework should support changing sensor sets, provided calibration and metadata are maintained. This is particularly useful for aging bridge inventories where permanently instrumenting every structure is not financially realistic.

Maintenance planning should include the monitoring system itself. Camera lenses require cleaning; enclosures can fog; strain cables can be damaged; accelerometer mounts can loosen. A sensor-health dashboard should track noise floor, missing-data rate, bias, image sharpness, tracking confidence, and communication quality. Maintenance can then be scheduled based on instrumentation condition. This prevents the common failure mode in long-term SHM where sophisticated algorithms are applied to gradually deteriorating sensors without adequate quality control.

Engineering acceptance also requires clear outputs. Maintenance staff generally need component location, anomaly type, severity indicator, uncertainty, trend, and supporting evidence. They do not need dozens of latent features. A good dashboard might show a fused displacement trend, first-mode frequency, strain-range trend, vision snapshot, sensor-health status, and explanation of why an alert was generated. The raw data and advanced analytics remain accessible for specialist review. This layered presentation supports both routine operations and detailed investigation.

Data retention policy is another cost decision. Continuous raw video can be prohibitively large, while derived displacement time histories are compact. A sensible policy stores continuous low-volume acceleration and strain, continuous derived visual features, periodic compressed images for audit, and full-rate video around triggered events. Cloud storage can support fleet-level analysis, but edge processing reduces bandwidth and improves resilience during connectivity loss. Data should be retained long enough to cover seasonal cycles and maintenance interventions.

The value of SHM ultimately depends on decisions changed by the information. Research should therefore connect fusion accuracy to maintenance outcomes. A reduction in displacement RMSE is technically useful, but the more important question is whether it reduces false closures, improves load-rating confidence, identifies a developing defect earlier, or prioritizes inspection resources more effectively. Future studies should include decision-level performance and lifecycle cost-benefit analysis alongside signal-processing metrics.

H. REPRODUCIBILITY, DATA SHARING, AND BENCHMARKING PROTOCOL

A reproducible SHM benchmark should provide raw data rather than only processed plots. For each event, acceleration, strain, and video should share a documented time base. The dataset should include independent reference displacement when available, environmental variables, sensor coordinates, camera calibration, and structural drawings sufficient to interpret response. Raw data enable future researchers to test different filters, optical-flow methods, and synchronization algorithms without repeating the expensive field campaign.

Benchmark splits should be standardized. Healthy baseline data from one period can be used for training, another period for validation, and a temporally separated period for testing. Damage or altered-state events should remain entirely in the test set unless the task is supervised classification. If windows from the same vehicle passage appear in both training and testing, performance can be severely inflated because the samples are almost duplicates. Event-level splitting is therefore mandatory for credible evaluation.

Metrics should include state-estimation error, modal-parameter error, anomaly-detection performance, and computational cost. For displacement estimation, RMSE, peak error, phase error, drift, and frequency-dependent error should be reported. For damage detection, precision, recall, false-alarm rate, area under the precision-recall curve, and detection delay are relevant. For long-term systems, availability and missing-data tolerance should be measured. Runtime and memory matter for edge deployment and should be evaluated on representative hardware.

A benchmark should also include controlled degradations. Vision sequences can be darkened, blurred, occluded, or subjected to simulated camera motion. Acceleration can receive bias, clipping, or dropped packets. Strain can receive thermal drift or channel loss. The objective is not to make the problem artificially difficult; it is to measure graceful degradation and sensor-fault isolation. A method that is marginally more accurate on clean data but fails catastrophically under one missing modality may be less useful than a slightly less accurate but robust alternative.

Open algorithms are as important as open data. Authors should release preprocessing scripts, model configuration, and evaluation code where licensing permits. Random seeds and software environments should be documented for machine-learning results. For proprietary computer-vision libraries or commercial DAQ systems, sufficient detail should be provided to reproduce the method with alternative tools. The long-term goal should be benchmark tasks analogous to those used in computer vision and speech research, but grounded in structural mechanics and engineering uncertainty.

Finally, publication practice should clearly label measured, derived, reconstructed, and simulated data. A plot that visually resembles field data can be misunderstood if its origin is not stated. The present paper deliberately labels its controlled benchmark as simulated and preserves published experimental metrics in their reported units. This practice improves trust and allows readers to distinguish evidence from illustration.

IX. CONCLUSION

This paper presented a unified framework for structural health monitoring using accelerometers, strain gauges, and vision data. The three sensing modalities are complementary: acceleration captures high-rate dynamics, strain measures local deformation and curvature-sensitive behavior, and vision supplies non-contact global displacement information. Their principal weaknesses—acceleration integration drift, strain locality and thermal sensitivity, and vision dependence on line of sight and camera stability—are sufficiently different that fusion can improve both accuracy and fault tolerance.

The proposed architecture combines asynchronous time alignment, modality-specific conditioning, adaptive quality control, strain-displacement calibration, multirate Kalman estimation, frequency-selective fusion, and physics-based consistency checks. A controlled benchmark demonstrated the expected benefit of combining low-frequency deformation anchors with high-frequency inertial response, while ablation illustrated the consequences of sensor dropout. These numerical results were explicitly separated from experimental evidence.

Published practical studies support each pairwise component of the tri-modal system. Acceleration-strain fusion has been validated numerically, experimentally, and in field testing; asynchronous vision-acceleration fusion has achieved reported displacement RMSE below 3 mm across multiple tests; recent adaptive approaches report fused NRMSE below 4%; and vision-strain fusion has been demonstrated using physics-informed learning. The next decisive step is a shared full-scale dataset in which all three modalities are acquired simultaneously under environmental variation, traffic loading, sensor faults, and known structural changes.

For engineering deployment, fusion should be treated as an uncertainty-management and observability problem rather than as a black-box prediction task. Systems that preserve physical interpretability, sensor-health diagnostics, uncertainty bounds, and traceable evidence are more likely to earn acceptance in bridge and building maintenance workflows.

ACKNOWLEDGMENT

The author acknowledges the researchers whose published laboratory and field studies provide the experimental evidence synthesized in this paper. No claim is made that the controlled benchmark data are measured field data.

REFERENCES

- [1] C. R. Farrar and K. Worden, *Structural Health Monitoring: A Machine Learning Perspective*. Chichester, U.K.: Wiley, 2013.
- [2] A. W. Smyth and M. Wu, "Multi-rate Kalman filtering for the data fusion of displacement and acceleration response measurements in dynamic system monitoring," *Mechanical Systems and Signal Processing*, vol. 21, no. 2, pp. 706–723, 2007.
- [3] J. W. Park, S. H. Sim, and H. J. Jung, "Displacement estimation using multimetric data fusion," *IEEE/ASME Transactions on Mechatronics*, vol. 18, no. 6, pp. 1675–1682, 2013.
- [4] J. W. Park, K. C. Lee, S. H. Sim, H. J. Jung, and B. F. Spencer Jr., "Traffic safety evaluation for railway bridges using expanded multisensor data fusion," *Computer-Aided Civil and Infrastructure Engineering*, vol. 31, no. 10, pp. 749–760, 2016.
- [5] Z. Ma, J. Chung, P. Liu, and H. Sohn, "Bridge displacement estimation by fusing accelerometer and strain gauge measurements," *Structural Control and Health Monitoring*, vol. 28, no. 6, Art. no. e2733, 2021, doi: 10.1002/stc.2733.

- [6] Z. Ma, J. Choi, and H. Sohn, "Real-time structural displacement estimation by fusing asynchronous acceleration and computer vision measurements," *Computer-Aided Civil and Infrastructure Engineering*, vol. 37, no. 6, pp. 688–703, 2022, doi: 10.1111/mice.12767.
- [7] Y. Chen et al., "Accurate structural displacement monitoring by data fusion of a consumer-grade camera and accelerometers," *Engineering Structures*, vol. 262, Art. no. 114303, 2022, doi: 10.1016/j.engstruct.2022.114303.
- [8] K. Aminfar, M. Ghyabi, Y. Shen, D. Lattanzi, G. Ciaramella, M. Gherlone, and C. Surace, "Data fusion in structural health monitoring: Integrating computer vision and strain gauges for enhanced inspection," in *Bridge Maintenance, Safety, Management, Digitalization and Sustainability*, 2024, pp. 709–716, doi: 10.1201/9781003483755-82.
- [9] D. Feng, M. Q. Feng, E. Ozer, and Y. Fukuda, "A vision-based sensor for noncontact structural displacement measurement," *Sensors*, vol. 15, no. 7, pp. 16557–16575, 2015.
- [10] D. Feng, Y. Fukuda, M. Feng, and M. Mizuta, "Nontarget vision sensor for remote measurement of bridge dynamic response," *Journal of Bridge Engineering*, vol. 20, no. 12, 2015, doi: 10.1061/(ASCE)BE.1943-5592.0000747.
- [11] "Multi-sensor fusion for structural displacement estimation: Integrating vision and acceleration from mobile devices," *Engineering Structures*, vol. 329, Art. no. 119826, 2025, doi: 10.1016/j.engstruct.2025.119826.
- [12] C. Zhan, Y. Huang, S. Liang, J. Fu, and R. Rao, "Prior-free vision-acceleration fusion with self-calibrated scaling for high-accuracy bridge displacement monitoring," *Measurement*, 2026, Art. no. 119705, doi: 10.1016/j.measurement.2025.119705.
- [13] J. W. Park, K. C. Lee, S. H. Sim, and B. F. Spencer Jr., "Visual-inertial displacement sensing using data fusion of vision-based displacement with acceleration," *Structural Control and Health Monitoring*, vol. 25, 2018, doi: 10.1002/stc.2122.
- [14] K. Luo, X. Kong, J. Zhang, J. Hu, J. Li, and H. Tang, "Computer vision-based bridge inspection and monitoring: A review," *Sensors*, vol. 23, no. 18, Art. no. 7863, 2023, doi: 10.3390/s23187863.
- [15] Y. Xu, J. M. W. Brownjohn, and F. Huseynov, "Accurate deformation monitoring on bridge structures using a cost-effective sensing system combined with a camera and accelerometers: Case study," *Journal of Bridge Engineering*, vol. 24, no. 1, 2019.
- [16] K. T. Park, S. H. Kim, H. S. Park, and K. W. Lee, "The determination of bridge displacement using measured acceleration," *Engineering Structures*, vol. 27, no. 3, pp. 371–378, 2005.
- [17] J. J. Lee and M. Shinozuka, "A vision-based system for remote sensing of bridge displacement," *NDT & E International*, vol. 39, no. 5, pp. 425–431, 2006.
- [18] E. Dar, T. Tufan, S. Akalp, and F. Yardımcı, "A hierarchical sensor data fusion and roving sensor network framework for structural health monitoring: Application to bridge retrofitting," *Sensors*, vol. 26, no. 11, Art. no. 3597, 2026, doi: 10.3390/s26113597.