Computer Vision-Based Progress Monitoring and Deviation Detection in Robotic Construction Operations

Ravi Dekker
Department of Computer Science, University of North Texas, Denton, TX, USA.
contactravi@unt.edu

Horst-Dieter Ziegert
Department of Computer Science, Colorado State University, Fort Collins, CO, USA.
contacthorstdieter@colostate.edu

Abstract

Robotic construction operations require continuous awareness of what has been built and how far the as-built state departs from the intended design. Project-level monitoring systems based on periodic imagery, laser scans, or manual inspection are poorly matched to task-level control because they do not resolve element placement, occlusion, registration uncertainty, or tolerance violations at the pace of a robotic work cycle. This paper develops a methodological framework for computer vision-based progress monitoring and deviation detection in robotic construction operations. The framework couples robot pose, multi-view images, depth or point-cloud observations, and building information model element graphs within a shared coordinate representation. Element-level progress is inferred as a probabilistic state from visual and geometric evidence, while geometric deviation is treated separately as a registered residual against element-specific tolerances. A robust scan-to-model registration objective, a progress-evidence model, a tolerance-exceedance measure, a Bayesian temporal fusion rule, a cost-sensitive decision rule, and an active-view selection criterion are formulated. The central argument is that progress and conformity should not be conflated: absence of evidence for completion differs from evidence of geometric error, and temporal fusion plus viewpoint planning can reduce false alarms under occlusion and partial observability. The framework provides a technically testable basis for supervisory control, corrective robot action, and inspection planning in robotic construction.

Keywords

robotic construction; computer vision; scan-to-BIM registration; progress monitoring; deviation detection; uncertainty-aware perception

1. Introduction

Automated progress monitoring remains difficult in construction because the observed state of a site changes rapidly, is frequently occluded, and is only weakly aligned with schedule-level abstractions. Vision-based sensing has therefore been investigated as a means of tracking temporary resources, activity, and as-built conditions without relying solely on manual surveys [1]. Early augmented reality and four-dimensional model approaches showed that site imagery can be registered to planned geometry and used to communicate progress deviations [2]. Laser scanning extended this logic from image evidence to metric as-built comparison, enabling recognition of model objects and dimensional compliance checks [3].

Appearance-based material classification further demonstrated that operation-level progress can be inferred from site photographs when contextual information from a four-dimensional building information model is available [4]. Reviews of point-cloud applications in construction document both the maturity of geometric capture and persistent difficulties in segmentation, registration, and semantic interpretation [5]. Camera-equipped unmanned aerial systems add flexible viewpoints for infrastructure inspection and monitoring [6], while articulated construction robots require pose-aware perception because the sensor and the effector move with the machine rather than with a fixed site camera [7]. Real-time image localization and registration against building information models provide one route to relating such moving observations to planned elements [8].

These developments establish that computer vision can supply useful evidence, but they do not by themselves resolve the control problem faced by a robotic construction operation. A robot placing a panel, laying a masonry unit, tying reinforcement, or welding a connection must decide whether the intended element is present, whether it is sufficiently complete to proceed, and whether its geometry lies within tolerance. Conventional progress monitoring tends to aggregate evidence at the level of activities or project zones. It may report that a floor is partially complete without resolving whether a particular element is misplaced by a few millimetres or simply occluded from the current camera view. Conversely, dimensional quality assessment may measure geometric conformity for elements already assumed to exist, without providing a temporally coherent account of completion. Dimensional quality assessment using terrestrial laser scanning has shown the value of metric comparison for precast components [9], and scan-to-building-information-model reconstruction has been widely studied as a route from point clouds to as-built representations [10]. Semantic enrichment of scanned building models further indicates that geometric data become more useful when they are attached to object identities and relationships [11]. The gap addressed here is the absence of a unified, uncertainty-aware formulation that treats progress and deviation as related but non-identical inference targets during robotic operations.

The central contribution of this paper is a methodological framework that couples vision-based element-state inference with metric deviation detection under robot motion, occlusion, and registration uncertainty. The framework does not assume that a single image classifier or a single point-cloud comparison is sufficient. It separates observable evidence into presence evidence, placement evidence, and geometric residual evidence; it fuses these over time through a Bayesian state model; and it selects viewpoints when the expected information gain justifies additional sensing. This structure is appropriate for robotic construction because the robot can act on the inferred state only if the state is both temporally stable and explicitly associated with a tolerance decision. The remainder of the paper defines the operational model, formulates the inference and decision steps, and assesses the principal error mechanisms and design trade-offs.

2. Operational and Sensing Model for Robotic Construction

The unit of analysis is a construction element or connection represented in a building information model and linked to an operation in the robot task plan. Each element k has planned geometry, an intended pose, material or appearance attributes, and a tolerance specification. The robot task plan defines an ordered set of operations, including approach, placement, fastening, and verification. The sensing state at time t includes robot pose, camera pose, image observations, depth or point-cloud observations, and any available proprioceptive measurements such as joint angles or end-effector position. These quantities are expressed in

a world frame, a robot frame, a camera frame, and a model frame; registration is the process of maintaining the transformations among them.

A general observation model clarifies what the perception system must estimate and what it must not confuse with ground truth. Let x_t denote the latent construction state at time t, including element completion status, placement pose, and geometric deviation. The measurement vector z_t is produced by imaging, depth sensing, or scanning. The observation relationship can be written as

$$z_t = h(x_t) + \varepsilon_t, \varepsilon_t \sim N(0, \Sigma_t)$$

where h is the sensor- and scene-dependent observation function, ε_t is a zero-mean measurement error term, and Σ_t is its covariance. The covariance is not a constant property of the sensor alone; it changes with range, surface reflectance, motion blur, occlusion, and calibration drift. The practical consequence is that progress and deviation claims must carry an uncertainty estimate rather than a binary label derived from a thresholded image or a nearest-point distance. The observation model also makes clear that missing measurements are not equivalent to negative evidence: an occluded element may be complete, incomplete, or misplaced.

Three evidence channels are relevant. Presence evidence indicates whether visible surfaces, edges, material patterns, or robot interactions are consistent with the planned element. Placement evidence indicates whether the observed element pose aligns with the planned pose within the resolution of the sensor and registration. Geometric residual evidence measures metric departure from the planned surface or boundary after registration. These channels are correlated but not interchangeable. A completed element can still be out of tolerance, and an out-of-tolerance element can be fully present. Progress monitoring therefore requires a state representation that allows completion and conformity to vary independently.

3. Methodological Framework

3.1 Multi-sensor registration and coordinate consistency

Registration is the first technical bottleneck because every subsequent inference depends on associating observed points and pixels with the correct model element. The framework uses a coarse-to-fine strategy. Robot kinematic pose and model-based camera projection provide an initial estimate. Image features or planar primitives refine the camera-to-model transform. Dense depth or point-cloud data then refine the object-level transform. Iterative closest point remains a standard basis for rigid alignment [12], while dense simultaneous localization and mapping demonstrates how depth observations can be fused into a consistent surface representation during motion [13]. In a robotic setting, the robot pose prior reduces the search space and helps reject correspondences that are geometrically plausible but semantically incorrect.

For a set of observed points q_i and corresponding model or map points p_i, the registration objective can be expressed as a robust minimization:

$$T* = arg\ min_T \sum_\{i = 1\}\^\{N\} \rho(||p_i - \Pi(T\ q_i)||^2)$$

where T is the rigid transformation from the observed frame to the model frame, Π is the correspondence or projection operator, ρ is a robust loss that limits the influence of outliers, and N is the number of retained correspondences. The robust loss is important in construction scenes because temporary objects, workers, tools, and partial surfaces generate correspondences that do not belong to the target element. The transformation T should be

constrained by robot kinematics and by the known mounting relationship between the camera and the end effector. Otherwise, registration can compensate for robot positioning error by distorting the model alignment, which would corrupt both progress and deviation estimates.

3.2 Element-level progress inference

Progress inference assigns a probabilistic completion state to each element. A suitable formulation is a logistic evidence model whose inputs combine image features, point-cloud features, model context, and robot state:

$$P_{k,t} = \sigma\left(\beta_0 + \beta^{T}\varphi(I_t, B_k)\right)$$

Here P_{k,t} is the probability that element k is sufficiently complete for the operation to proceed at time t, σ is the logistic function, β₀ is an intercept, β is a parameter vector, φ is a feature mapping, I_t is the observation set at time t, and B_k is the model context for element k. The feature mapping can include learned convolutional descriptors, detected object proposals, surface coverage, edge alignment, material appearance, and the relation between the observed element and the robot end effector. Deep learning provides the representational capacity for such features [14]; residual networks support robust image recognition at depth [15], real-time object detectors support localisation of relevant components [16], and point-set networks support classification and segmentation directly from unordered three-dimensional data [17]. The expression defines an inference target and a feature interface; it does not imply that any particular parameter set has been estimated or validated.

Progress inference must also encode operation-specific sufficiency. In bricklaying, a unit may be present but not yet levelled; in panel installation, the panel may be placed but not fastened; in welding, the joint may be geometrically correct but not completed. The model context B_k should therefore include the operation type and the required predecessor state. A high presence probability alone should not trigger a transition to the next task if the predecessor condition is unmet. This is a design requirement rather than a statistical detail, because progress labels that ignore operation dependencies produce a state machine that advances on appearance rather than on construction logic.

3.3 Geometric deviation and tolerance testing

Geometric deviation is computed only after registration and element association. For an element with a planned reference plane or surface, let n_k be the outward unit normal from the model, c_k a point on the planned surface, S_k the set of observed points assigned to the element, and τ_k the element-specific tolerance. A signed or absolute exceedance measure can be written as

$$d_{k,t} = \max_{i \in S_k} |n_k \cdot (x_i - c_k)| - \tau_k$$

where x_i is an observed point and d_{k,t} is the positive exceedance beyond tolerance. A positive value indicates that at least one assigned point lies outside the allowed band, while a non-positive value indicates that the observed maximum residual is within tolerance. The measure is deliberately simple and interpretable, but it is only valid if S_k is correctly segmented and if registration error is small relative to τ_k. If S_k is empty, d_{k,t} is undefined, and the system should record absence of geometric evidence rather than a conformity pass.

The relationship between registration uncertainty and tolerance is a central design constraint. An angular registration error produces a tangential displacement that grows with

distance from the rotation centre, so the same angular uncertainty can be negligible for a compact element and unacceptable for a long structural member. Similarly, a small translation bias can shift all residuals in one direction and mimic a systematic placement error. Tolerance testing should therefore use an uncertainty-adjusted residual, or at minimum a guard band, in which residuals near τ_k are not treated as decisive. The framework can use the covariance Σ_t from the observation model to inflate the effective tolerance when registration quality is poor, and it can suppress deviation decisions when the observed point count or spatial coverage is insufficient.

3.4 Temporal fusion and decision logic

Single-frame progress and deviation decisions are vulnerable to occlusion, motion blur, and transient objects. A temporal state model reduces this vulnerability by treating completion and conformity as persistent states observed through noisy evidence. Let s_t denote the latent element state at time t, and let z_{1:t} denote all observations up to time t. The posterior can be updated recursively:

$$p(s_t|z_{1:t}) \propto p(z_t|s_t) \sum_{s_{t-1}} p(s_t|s_{t-1}) p(s_{t-1}|z_{1:t-1})$$

The transition term p(s_t | s_{t-1}) encodes construction logic and robot task progression, while the likelihood p(z_t | s_t) encodes the compatibility of current visual and geometric evidence with each state. This formulation allows a temporarily occluded element to retain a plausible posterior instead of being reset to “not started,” and it allows a single outlier scan to be down-weighted when prior observations support conformity. The transition model should be operation-specific: a completed weld cannot revert to uncompleted without a documented removal or rework event, whereas a loosely placed panel may legitimately change pose before fastening.

The decision rule converts the posterior into an action. Let F_t be a binary flag indicating that corrective action or engineering review is required, and let γ be a threshold selected from the relative cost of false acceptance and false rejection. The rule can be written as

$$F_t = 1(\mathrm{P}(s_t = 1|z_{1:t}) \geq \gamma)$$

where 1(·) is the indicator function. The threshold γ should not be a universal constant. A deviation that would create a safety hazard or block subsequent robot motion warrants a lower threshold than a cosmetic surface irregularity. The posterior also provides a principled way to separate “not yet verified” from “verified acceptable”: if the posterior remains diffuse because of occlusion, the appropriate output is a request for another viewpoint, not a declaration of progress or compliance.

Active sensing follows from the same logic. When the posterior is uncertain, the robot or an external camera can move to a viewpoint that maximizes expected information gain. Let v denote a candidate viewpoint or sensing pose, V the admissible set, H the entropy of the element state, and z a hypothetical observation from viewpoint v. A next-best-view criterion is

$$v* = argmax_\{v \in V\}\, E_\{z|v\}[H(s_t) - H(s_t \mid z)]$$

where v is the selected viewpoint. The admissible set V must respect robot reachability, collision constraints, cycle-time limits, and sensor field of view. This criterion is preferable to a fixed inspection pose when element geometry or temporary obstructions make some surfaces unobservable from the nominal camera position. The practical effect is to spend

sensing effort where it changes a progress or deviation decision, rather than collecting redundant views of already resolved elements.

4. Analytical Assessment and Technical Discussion

4.1 Error mechanisms and false decisions

The framework's credibility depends on how it handles error propagation. Registration error is the dominant metric risk. A rigid transform error has both translational and rotational components; the rotational component produces residuals that increase with lever arm, so long elements and elements far from the registration centre require tighter angular accuracy. Segmentation error is a second risk. If points from an adjacent element or temporary object are assigned to S_k, the maximum residual in the deviation equation is driven by the outlier rather than by the target surface. Robust loss in registration reduces this effect during alignment, but it does not replace element-level outlier rejection after alignment. Spatial coverage checks are therefore necessary: a small cluster of points on one edge should not be treated as evidence of full-surface conformity.

Appearance-based progress inference has complementary failure modes. Repeated textures, similar materials, and partial views can cause a completed element to be classified as incomplete, while temporary formwork or stored materials can create false presence evidence. Temporal fusion mitigates these errors only if the transition model is correctly specified. If the model permits frequent backward transitions, occlusion can produce oscillating progress estimates; if it forbids them too strongly, a genuine removal or rework event will be missed. The appropriate design is to make transitions directional but evidence-sensitive, with explicit rework states rather than silent reversals.

The separation of progress from deviation prevents a common logical error. A system that treats "no detected deviation" as "complete" will mark occluded elements as acceptable. A system that treats "complete" as "conforming" will ignore placement error once presence is detected. The proposed framework avoids both by maintaining a completion posterior and a geometric exceedance measure, then combining them only at the decision stage. This distinction is especially important for robotic operations because the next action depends on both: a robot should not proceed over an element that is present but out of tolerance, and it should not repeat placement simply because the element is temporarily hidden.

4.2 Calibration, observability, and operating conditions

Calibration is not a one-time laboratory procedure in robotic construction. Camera-to-robot transforms can drift because of vibration, impact, thermal variation, or tool changes. The observation covariance Σ_t should therefore reflect calibration state, and registration should be rejected when the estimated transform is inconsistent with the robot kinematic prior beyond a plausible bound. Dimensional and surface quality assessment using building information models and laser scanning illustrates how model-based comparison depends on consistent reference geometry [19]. In the present framework, that dependence is made explicit: the model is not a neutral backdrop but a source of element identity, reference surfaces, and tolerances.

Observability varies with task stage. Before placement, the target region may be visible but the element absent. During placement, the end effector and the element occlude one another. After fastening, the relevant surfaces may be partly covered by subsequent work. The active-view criterion addresses this by selecting views that discriminate among competing

states. The cost of additional sensing is not zero, especially for a robot with a constrained cycle time, so viewpoint selection should be integrated with task scheduling rather than treated as a separate inspection step.

The design also has implications for robot platform architecture. On-site industrial robots require sensor placement, protective enclosures, and calibration procedures that differ from factory cells [18]. A robot-mounted camera offers a consistent view of the effector but a narrow and moving view of the workpiece. An external camera offers a stable global view but is vulnerable to occlusion by the robot and by temporary structures. The framework is compatible with either configuration, but it requires that the chosen configuration provide enough spatial coverage to estimate the relevant residuals. If the sensor cannot observe the tolerance-critical surface, the correct output is an observability warning, not a numerical deviation.

4.3 Trade-offs and implementation logic

The framework involves deliberate trade-offs. Dense point-cloud registration provides metric evidence but increases computation and payload requirements. Sparse image-based inference is lighter but less directly tied to tolerance. A practical system can use image evidence for progress hypotheses and reserve dense geometric verification for elements whose tolerance or safety criticality warrants it. Similarly, temporal fusion improves stability but adds latency; a fast robotic loop may require a short-horizon update, while a supervisory layer maintains the longer-horizon posterior.

The choice of threshold γ and tolerance τ_k should be derived from the operation's consequence model, not from a generic accuracy target. False acceptance of an out-of-tolerance element can propagate error into subsequent placements; false rejection can trigger unnecessary rework and reduce productivity. The decision rule should therefore be cost-sensitive and should expose its uncertainty to the robot planner. A planner can then choose among proceed, re-scan, request human verification, or execute a corrective motion.

Compared with manual inspection, the framework offers repeatability and temporal continuity, but it does not remove the need for validated calibration and defined tolerances. Compared with purely data-driven progress classification, it provides a metric route to deviation detection and an explicit treatment of missing evidence. Compared with purely geometric scan-to-model comparison, it incorporates operation dependencies and temporal state, which are necessary when the same geometric configuration can correspond to different task stages. The contribution is therefore not a single algorithm but a coherent inference architecture in which progress, deviation, uncertainty, and sensing action are linked.

5. Conclusion

This paper has developed a computer vision-based methodological framework for progress monitoring and deviation detection in robotic construction operations. The framework represents construction elements as model-linked entities, registers multi-view and depth observations to those entities under robot pose priors, infers element completion probabilistically, measures geometric exceedance against element-specific tolerances, fuses evidence over time, and selects additional viewpoints when uncertainty affects the decision. Its central technical claim is that progress and conformity must be estimated separately and combined only through an explicit decision rule. This avoids the twin errors of treating occlusion as absence and treating presence as conformity. The formulation also makes uncertainty operational: registration covariance, spatial coverage, and observability determine

whether a deviation claim is decisive or whether further sensing is required. For robotic construction, the practical value lies in providing a state estimate that can support proceed, re-scan, rework, or corrective-action decisions without conflating task completion with geometric acceptance. The framework is analytically grounded in established registration, point-cloud, deep-learning, and model-based quality-assessment methods, and it defines the conditions under which those methods can be coupled rather than merely juxtaposed.

References

1. Teizer, J. (2015). Status quo and open challenges in vision-based sensing and tracking of temporary resources on infrastructure construction sites. Advanced Engineering Informatics, 29(2), 225–238.

2. Golparvar-Fard, M., Peña-Mora, F., & Savarese, S. (2009). D4AR–a 4-dimensional augmented reality model for automating construction progress monitoring data collection, processing and communication. Journal of Information Technology in Construction, 14, 129–153.

3. Bosché, F. (2010). Automated recognition of 3D CAD model objects in laser scans and calculation of as-built dimensions for dimensional compliance control in construction. Advanced Engineering Informatics, 24(1), 107–118.

4. Han, K. K., & Golparvar-Fard, M. (2015). Appearance-based material classification for monitoring of operation-level construction progress using 4D BIM and site photologs. Automation in Construction, 53, 44–57.

5. Wang, Q., & Kim, M. K. (2019). Applications of 3D point cloud data in the construction industry: A fifteen-year review from 2004 to 2018. Advanced Engineering Informatics, 39, 306–319.

6. Ham, Y., Han, K. K., Lin, J. J., & Golparvar-Fard, M. (2016). Visual monitoring of civil infrastructure systems via camera-equipped Unmanned Aerial Vehicles (UAVs): a review of related works. Visualization in Engineering, 4(1), 1–8.

7. Liang, C. J., Lundeen, K. M., McGee, W., Menassa, C. C., & Kamat, V. R. (2019). A vision-based marker-less pose estimation system for articulated construction robots. Automation in Construction, 104, 80–94.

8. Asadi, K., Ramshankar, H., Noghabaei, M., & Han, K. (2019). Real-time image localization and registration with BIM using perspective alignment for indoor monitoring of construction. Journal of Computing in Civil Engineering, 33(5), 04019031.

9. Kim, M. K., Sohn, H., & Chang, C. C. (2014). Automated dimensional quality assessment of precast concrete panels using terrestrial laser scanning. Automation in Construction, 45, 163–177.

10. Tang, P., Huber, D., Akinci, B., Lipman, R., & Lytle, A. (2010). Automatic reconstruction of as-built building information models from laser-scanned point clouds: A review of related techniques. Automation in Construction, 19(7), 829–843.

11. Xiong, X., Adan, A., Akinci, B., & Huber, D. (2013). Automatic creation of semantically rich 3D building models from laser scanner data. Automation in Construction, 31, 325–337.

12. Besl, P. J., & McKay, N. D. (1992). A method for registration of 3-D shapes. IEEE Transactions on Pattern Analysis and Machine Intelligence, 14(2), 239–256.

13. Newcombe, R. A., Izadi, S., Hilliges, O., Molyneaux, D., Kim, D., Davison, A. J., Kohli, P., Shotton, J., Hodges, S., & Fitzgibbon, A. (2011). KinectFusion: Real-time dense surface mapping and tracking. In 2011 10th IEEE International Symposium on Mixed and Augmented Reality (pp. 127–136). IEEE.

14. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436–444.

15. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 770–778).

16. Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 779–788).

17. Qi, C. R., Su, H., Mo, K., & Guibas, L. J. (2017). PointNet: Deep learning on point sets for 3D classification and segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 652–660).

18. Luo, D., & Yu, L. (2021). From factory to site—designing for industrial robots used in on-site construction. In Automating cities: design, construction, operation and future impact (pp. 87-109). Singapore: Springer Singapore.

19. Kim, M. K., Cheng, J. C., Sohn, H., & Chang, C. C. (2015). A framework for dimensional and surface quality assessment of precast concrete elements using BIM and 3D laser scanning. Automation in Construction, 49, 225–238.

20. Bircher, A., Kamel, M., Alexis, K., Oleynikova, H., & Siegwart, R. (2016). Receding horizon next-best-view planner for 3D exploration. In 2016 IEEE International Conference on Robotics and Automation (ICRA) (pp. 1462–1468). IEEE.