Remembering What Matters

Semantic Compression and Structured Memory for Video at the Edge

Prakhar Singh[1] Abdulhamid Hassan[1]
[1]Vatar, London, United Kingdom
p@vatar.com a@vatar.com

Technical White Paper, Version 0.1, August 2026

Abstract

Video is the largest and least useful data stream most organisations produce. Cameras record continuously, but the information density of that recording is extremely low: the same corridor, the same road, the same wall, for hours at a time. Conventional video infrastructure treats every frame as equally worth keeping, moving and searching. This paper argues that the binding constraint on video systems is no longer the cost of storing bits but the cost of understanding them, and that the correct response is to change what is preserved rather than how it is encoded. We introduce the distinction between *video compression*, which reduces the bits needed to reproduce a signal, and *information compression*, which reduces the information that needs to exist at all. We describe an architecture in which a camera feed is interpreted at the point of capture into a hierarchy of objects, actions, events and context; in which retention fidelity is allocated according to significance rather than time; and in which retrieval operates over a structured memory of what happened rather than over raw footage. We describe how this architecture runs on constrained edge hardware without connectivity, how it exposes a natural-language query, alerting and decision interface, and how it should be evaluated. The specific models and mechanisms Vatar uses are proprietary and are not described here.

1. The problem

The world generates more video than it can store, transmit, process or understand. Organisations deploy hundreds or thousands of cameras across offices, factories, warehouses, estates, banks, retail sites and public space. Autonomous systems observe the physical world continuously. Phones add hours of footage per person per day.

The abstraction used to handle this video has barely changed in twenty years: capture it, compress it, store it, and hope a human can find what they need later. Video management systems are, at their core, a recorder and a scrub bar.

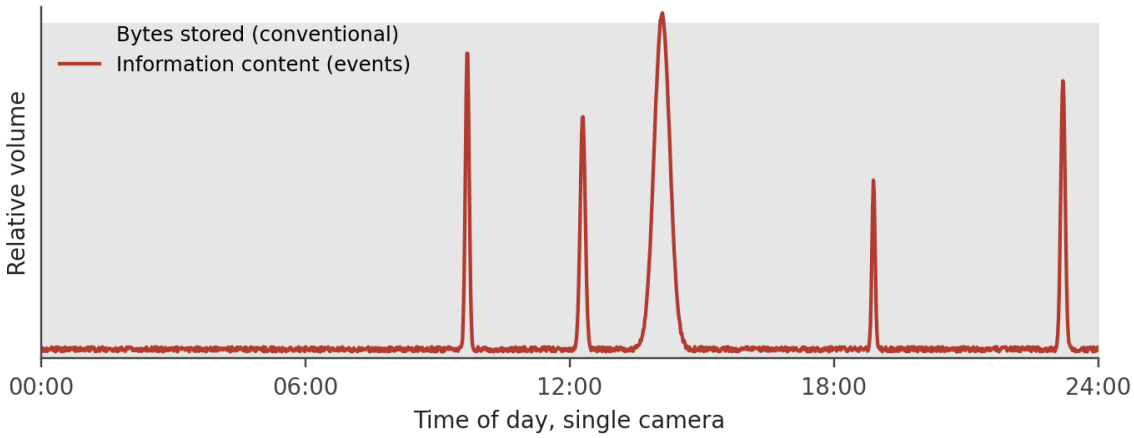

Figure 1. Bytes stored versus information content over one camera-day. Conventional systems store a constant volume regardless of activity. Almost all information of any value is concentrated in a handful of short intervals.

This model has three structural weaknesses.

Uniform value assumption. Every second of footage is stored at the same fidelity, whether it contains an intrusion or an empty car park. A camera at a building entrance records 86,400 seconds per day. The number of those seconds that will ever matter to anyone is usually in the tens.

Search cost scales with footage, not with events. Answering “what happened before the alarm” requires a human, or a machine, to inspect footage linearly. The cost of a query grows with the amount recorded, not with the amount that happened.

Bandwidth and compute are spent in the wrong place. Moving every frame from every camera to a central store, and then running analysis there, means the most expensive resources (transport and central compute) are spent on the least valuable data (routine frames). In environments with unreliable or costly connectivity, which describes most critical infrastructure and every forward-deployed setting, this model fails outright.

The result is that most recorded video is never watched, most of what is watched is irrelevant, and the questions organisations actually want answered remain expensive or impossible.

2. Two kinds of redundancy

Conventional codecs (H.264, H.265, AV1) [1, 2] achieve remarkable compression by exploiting *signal redundancy*: correlation between neighbouring pixels within a frame and between successive frames. A static scene compresses well because little changes at the pixel level.

There is a second form of redundancy that codecs do not touch: *semantic redundancy*. Two hours of footage of an empty corridor and one second of that footage carry the same information about the world. A codec will still encode two hours. The redundancy is not in the pixels. It is in the meaning.

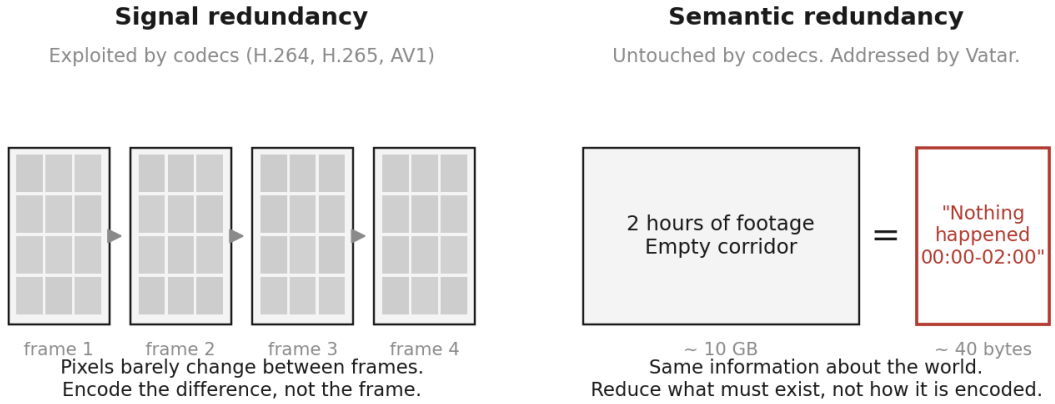

Figure 2. Signal redundancy (left) is handled by codecs, which encode differences between frames. Semantic redundancy (right) is not: two hours of an empty corridor and a forty-byte statement carry the same information about the world.

Consider a single camera pointed at a building entrance for 24 hours. The conventional question is: how do we store this video more efficiently? The more useful question is: what information from this video will matter tomorrow? The honest answer is usually short.

- A person entered at 09:41.
- A vehicle arrived at 12:18.
- Someone remained in a restricted area for seven minutes.
- A door was opened outside operating hours.
- An object moved from one location to another.

Almost none of the original pixels are needed to answer those questions. This leads to the central claim of this paper.

> *The ultimate form of compression is not making the file smaller. It is reducing the amount of information that needs to be preserved in the first place.*

We call this *information compression*, to distinguish it from video compression. Video compression asks how to represent the same signal in fewer bits. Information compression asks what the minimum information is to preserve what we actually care about. The framing follows Shannon's original distinction between the signal and the message it carries [3]. The first is a mature engineering discipline. The second is still an open frontier, and it has become tractable only because models can now understand images and video with sufficient fidelity: open-vocabulary recognition [4], general segmentation [5], and vision-language models that answer questions about images [6].

3. Video as a stream of events

We treat a camera not as a device that produces video but as a sensor observing reality. Video is the raw signal from that sensor. The object of interest is not the signal. It is what changed in the world.

This reframing produces a hierarchy of representations. Each layer contains less raw data than the one before it and, in general, more useful information.

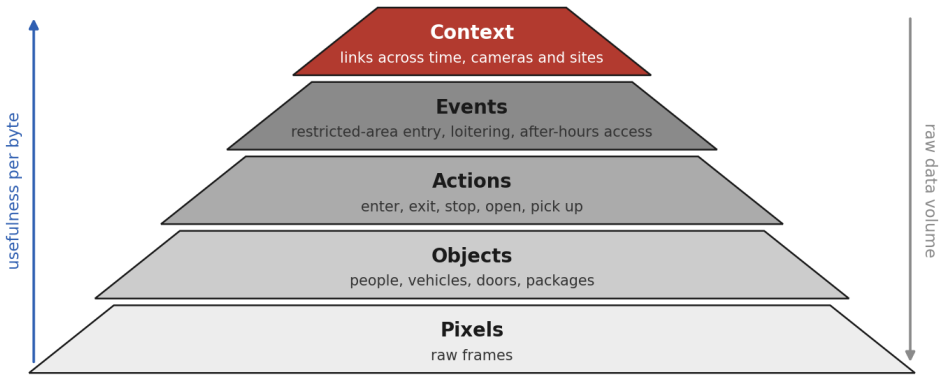

Figure 3. The representation hierarchy. Raw data volume falls and usefulness per byte rises as the system moves from pixels toward context.

- **Pixels** are the raw frames.
- **Objects** are the entities present: people, vehicles, packages, doors, tools.
- **Actions** are what objects do over short intervals: enter, exit, stop, pick up, open.
- **Events** are actions with significance: a person entered a restricted area; a vehicle loitered; a door was opened after hours.
- **Context** links events to each other and to the world: the same person seen at three sites; the vehicle that arrived before the alarm; the shift on duty when it happened.

The purpose of the architecture is not to discard the pixels. It is to demote them. Raw video becomes the lowest-level representation, retained where it is needed as evidence, rather than the primary representation that everything else is derived from on demand.

This inverts the conventional pipeline. Today, organisations store first and extract meaning later, if ever. We propose the opposite order: **understand first, preserve intelligently, retrieve when necessary.**

4. Architecture

The Vatar system sits between the physical world and the volume of raw visual data that cameras generate. It has four stages, all of which run on the edge device.

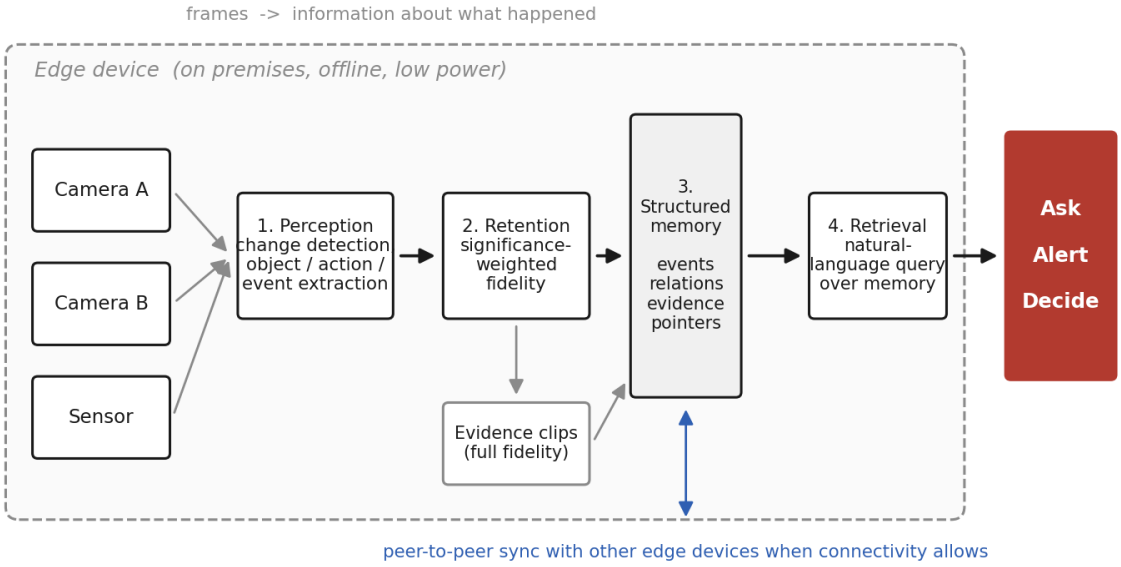

Figure 4. System architecture. Perception, retention, memory and retrieval all run on a single low-power device co-located with the cameras. Only information about what happened, and preserved evidence where required, ever leaves it.

4.1 Perception at the point of capture

Interpretation happens on a device co-located with the camera or camera cluster, not in a central store. Frames are read from any existing camera or sensor feed, including legacy analogue-to-IP converters, ONVIF streams, and non-visual sensors. Models on the device extract the object, action and event layers in near real time, with identity maintained across frames and cameras using tracking methods in the tradition of [7]. The specific models, their training, and the change-detection mechanisms that gate them are proprietary and outside the scope of this paper.

Doing this at the edge has two consequences. First, what leaves the device is information about what happened, which is orders of magnitude smaller than the footage. Second, the system works with no connectivity at all. Nothing about perception depends on a data centre.

4.2 Significance-weighted retention

Retention fidelity is a function of significance, not of time. Routine intervals are represented compactly: a summary of what was present, at what density, over what period. Events retain progressively more: object crops, short clips at full resolution around the event boundary, and the complete surrounding footage where the event class warrants it.

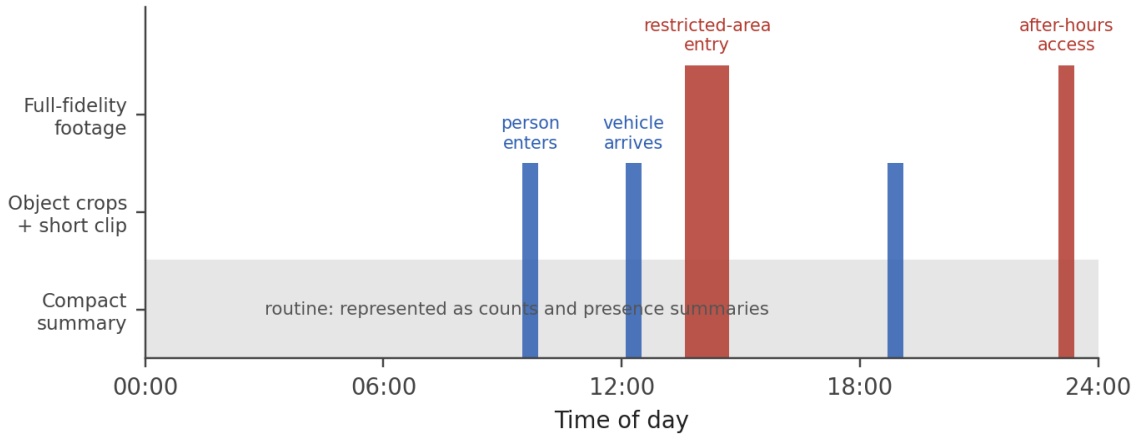

Figure 5. Significance-weighted retention across one camera-day. Routine periods are kept as compact summaries. Ordinary events retain object crops and a short clip. High-significance events retain full-fidelity footage around the event boundary.

The retention policy is configurable per deployment and can be tuned by the organisation's own definitions of significance (Section 5). The default policy preserves full-fidelity evidence for any event that could plausibly be investigated, and compact representations for everything else.

4.3 Structured memory

The output of perception and retention is a structured memory: a time-indexed store of objects, actions, events and their relationships, with pointers to whatever visual evidence was preserved. This memory is queryable in ways that raw footage is not. It supports temporal reasoning (before, after, during), spatial reasoning (in zone, crossed line, near), identity persistence (same object across time and cameras), and aggregation (how many, how often, how long).

Where several devices are deployed, memories synchronise peer to peer when connectivity allows and reconcile on reconnection. No device depends on any other, and none depends on a central server, to continue operating.

4.4 Retrieval

Retrieval operates over the structured memory, not over footage. Searching an index of events is fundamentally different from scanning hours of video. Prior systems work on video query optimisation [8, 9] reached the same conclusion from the database side: the cost of a query should be governed by what is asked, not by how much was recorded. The search space is narrowed progressively: from the memory to candidate events, from candidate events to preserved evidence, and only then, if required, to raw frames. Most questions never reach the last step.

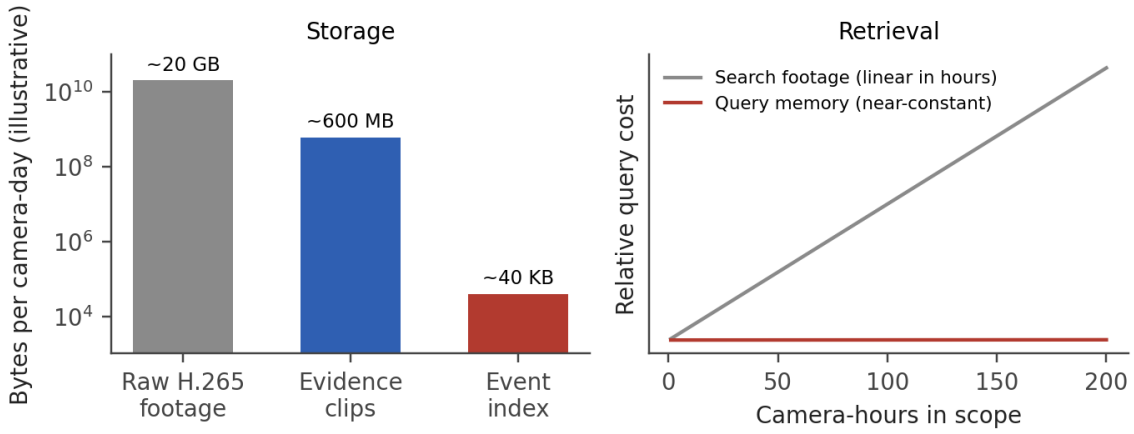

Figure 6. Illustrative economics. Left: bytes per camera-day for raw footage, preserved evidence, and the event index (log scale; figures indicate order of magnitude only and vary with scene activity). Right: query cost against footage in scope. Scanning footage is linear in camera-hours; querying memory is near-constant.

5. The interface: Ask, Alert, Decide

The architecture above is exposed to operators through three primitives.

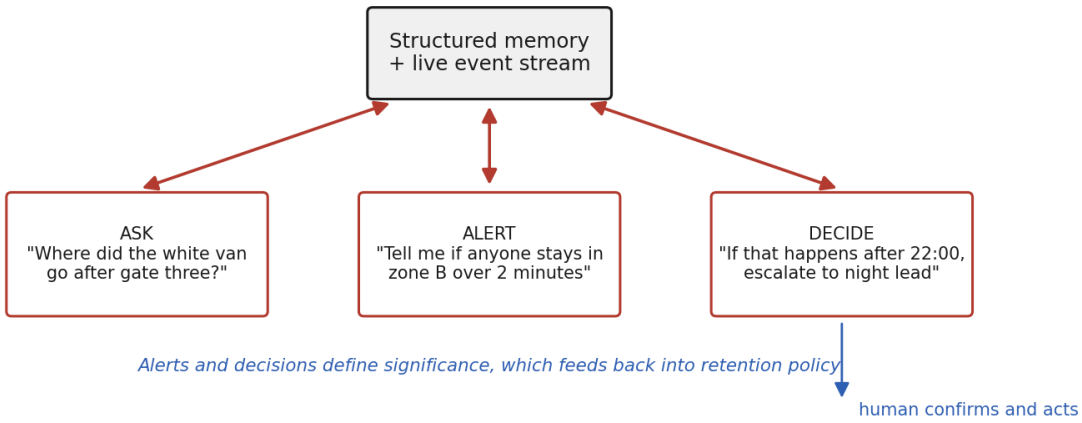

Figure 7. The three interface primitives operate over the same structured memory and live event stream. Alerts and decisions encode the organisation's definition of significance and feed back into retention policy. A human confirms and acts at the end of every Decide path.

Ask. A question in plain language is translated into a query over the structured memory. "Where did the white van go after it left gate three." "Show every vehicle that stopped on the perimeter road for more than two minutes last week." "How many people passed the east entrance between midnight and four." There are no rules to configure in advance and no fixed list of analytics classes. If it can be described, it can be found.

Alert. A standing condition is registered against the live event stream. When the condition is met, the relevant person is notified with the evidence attached. Conditions can be spatial, temporal, behavioural or compositional: a crowd forming where none should be, a vessel loitering outside an approach, a door propped open on a locked floor, two conditions occurring together.

Decide. A response is attached to a condition. If this is seen, notify that person. If that threshold is crossed, escalate to this team. If two conditions coincide, trigger the protocol. The organisation encodes its judgement once and the system applies it every time, at every site, at three in the morning as reliably as at noon. Humans remain in the loop by design: the system proposes and routes; people confirm and act.

These primitives also feed back into retention. An organisation's alerts and decisions are, in effect, its definition of significance. The retention policy learns from them.

6. Why the edge

Everything described above runs on small, low-power hardware, fully offline. This is not a deployment option. It is the design constraint the architecture was built around, for three reasons. Video analytics has been identified as the defining workload for edge computing precisely because of these constraints [10, 11].

Connectivity cannot be assumed. Critical infrastructure, remote sites, vessels, borders and forward positions have intermittent, expensive or absent connectivity. A system that needs the cloud to think does not work where it is most needed.

Data sovereignty cannot be negotiated. Governments and regulated enterprises do not send their footage to third-party infrastructure. The data must never leave the site. Intelligence has to live where the decision is made.

The economics only work at the edge. If every frame must be moved to a central store before it is understood, the bandwidth cost of ubiquitous cameras is unsustainable. Interpreting at capture and moving only information is the only model that scales with camera count.

The engineering consequence is that models must be small enough, and inference efficient enough, to run within the power and thermal budget of a device the size of a book. This constraint shapes every part of the system.

7. Evaluation

A system built on this thesis should be judged on measures that conventional video infrastructure cannot report. We propose the following.

Measure	What it captures
Event recall and precision	Are the events that matter detected, and are false events suppressed
Evidence sufficiency	When an event is investigated, was enough visual fidelity preserved to resolve it
Retention ratio	Bytes preserved per camera-day versus raw encoded footage
Query latency	Time to answer a natural-language question over N camera-days
Query coverage	Proportion of operator questions answerable without reverting to raw footage
Alert lead time	Time between condition onset and notification
Power and thermal envelope	Sustained inference within the device budget
Offline continuity	Operation and reconciliation across connectivity loss

Table 1. Proposed evaluation measures.

Reporting these together is important. Retention ratio without evidence sufficiency rewards deleting the wrong things. Recall without precision rewards alerting on everything.

8. Limitations and open questions

Several problems are unsolved and we state them plainly.

Significance is not known in advance. Something routine today may matter after an incident is reported next week. The architecture mitigates this by keeping compact representations of routine periods and by allowing evidence policies to be conservative, but the risk of having demoted something that later mattered

is real and must be managed by policy, not eliminated by technology.

Model error compounds. An error at the perception layer propagates upward. A missed object becomes a missed action, a missed event, and a gap in memory. Robustness under degraded imagery, occlusion, adverse weather and adversarial manipulation is an active area of work.

Evidentiary standards. Where preserved output may be used in legal or disciplinary proceedings, chain of custody, tamper evidence, and the relationship between derived events and original frames must meet standards that are still being defined for AI-interpreted footage. In the UK, deployments must also satisfy the Surveillance Camera Code of Practice [12] and data protection law.

Bias and fairness. Perception models inherit the distribution of their training data. Deployment in security contexts places a high burden on measuring and correcting differential performance across people and conditions, a failure mode documented in commercial vision systems [13].

9. Beyond video

Video is the starting point because it is the hardest data to make sense of and the most neglected. It is not the boundary of the thesis. The same principle, understand first, preserve according to significance, retrieve from structured memory, applies to any continuous data stream: audio, radio, sensor telemetry, transaction logs, documents. Each new modality is a new source feeding the same memory. The system that begins by answering questions about footage is designed to end by answering questions about everything an organisation holds.

10. Conclusion

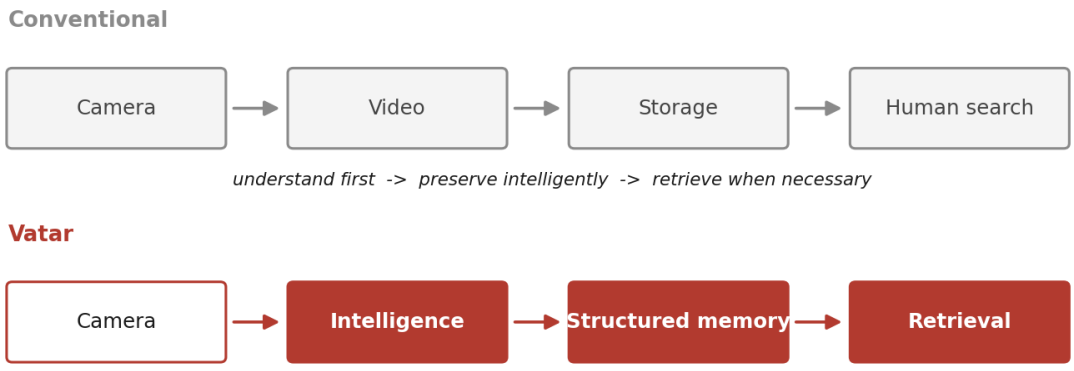

Figure 8. The video stack, conventional and proposed. Raw video does not disappear; it becomes the lowest layer of a system that understands before it stores.

Raw video does not disappear in this model. It becomes one layer of a richer system: high-value events keep their original evidence, routine activity is represented efficiently, context is indexed, and the system continuously decides what deserves to be remembered.

This is not a storage optimisation. It is a change in the economics of machine perception. If cameras are going to be ubiquitous, we cannot afford to treat everything they see as equally important. The next generation of infrastructure will not win by storing all of it. It will win by understanding it, and by remembering only what matters.

Correspondence. Prakhar Singh (p@vatar.com), Abdulhamid Hassan (a@vatar.com). Vatar, London, United Kingdom.

Vatar is a UK company building edge-deployed intelligence for critical decisions. This paper describes architecture and principles. Implementation details are proprietary. Figures 1, 5 and 6 are illustrative and do not report measured results.

References

[1] T. Wiegand, G. J. Sullivan, G. Bjøntegaard and A. Luthra. Overview of the H.264/AVC video coding standard. *IEEE Transactions on Circuits and Systems for Video Technology*, 13(7), 2003. doi:10.1109/TCSVT.2003.815165

[2] G. J. Sullivan, J.-R. Ohm, W.-J. Han and T. Wiegand. Overview of the High Efficiency Video Coding (HEVC) standard. *IEEE Transactions on Circuits and Systems for Video Technology*, 22(12), 2012. doi:10.1109/TCSVT.2012.2221191

[3] C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27, 1948. doi:10.1002/j.1538-7305.1948.tb01338.x

[4] A. Radford et al. Learning transferable visual models from natural language supervision. *ICML*, 2021. arXiv:2103.00020

[5] A. Kirillov et al. Segment Anything. *ICCV*, 2023. arXiv:2304.02643

[6] H. Liu, C. Li, Q. Wu and Y. J. Lee. Visual instruction tuning. *NeurIPS*, 2023. arXiv:2304.08485

[7] N. Wojke, A. Bewley and D. Paulus. Simple online and realtime tracking with a deep association metric. *IEEE ICIP*, 2017. arXiv:1703.07402

[8] D. Kang, J. Emmons, F. Abuzaid, P. Bailis and M. Zaharia. NoScope: Optimizing neural network queries over video at scale. *Proceedings of the VLDB Endowment*, 10(11), 2017. arXiv:1703.02529

[9] K. Hsieh et al. Focus: Querying large video datasets with low latency and low cost. *USENIX OSDI*, 2018. usenix.org/conference/osdi18/presentation/hsieh

[10] M. Satyanarayanan. The emergence of edge computing. *IEEE Computer*, 50(1), 2017. doi:10.1109/MC.2017.9

[11] G. Ananthanarayanan et al. Real-time video analytics: The killer app for edge computing. *IEEE Computer*, 50(10), 2017. doi:10.1109/MC.2017.3641638

[12] UK Home Office. Surveillance Camera Code of Practice, 2021. gov.uk/government/publications/update-to-surveillance-camera-code

[13] J. Buolamwini and T. Gebru. Gender Shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of Machine Learning Research*, 81, 2018. proceedings.mlr.press/v81/buolamwini18a.html