

Embedding models and the future of algorithmic video surveillance

Jorrit van Assen,¹ and Luc Rocher¹

¹Oxford Internet Institute, University of Oxford

Police forces worldwide are adopting artificial intelligence (AI) tools to watch, sort, and search the public at a scale no human workforce could match. Once limited to pilot projects, facial recognition is now a routine infrastructure. The Metropolitan Police scanned over 1.7 million faces in London in 2026 so far, up 87 per cent year on year [1]. France was found to have run an illegal facial-recognition program for years, with the software BriefCam deployed across more than a hundred municipalities to analyse footage [2]. Yet, for all the controversy it attracts, facial recognition does not appear to be enough. The UK’s head of a new £115 million centre for AI in policing, Alex Murray, has described his ambition to interact with AI in natural language, to ask it to ‘find the man with the red hat’ and, within the investigative golden hour, to ‘go through his door and get him with blood on his shoes’ [3].

For most of the history of computer vision, machines could only recognise what they had been taught to recognise. They could read a number plate, detect a face, or spot a pedestrian, but only after being trained on thousands of human-annotated examples. To ‘find the man with the red hat’, Murray would first need to gather and label footage of men in red hats, in daylight and at night, before training a model to pick them out. Vision-language models (VLMs) changed this. Trained on massive datasets of image–text pairs from the web, models such as CLIP learned a shared *embedding* space for pictures and words so that the phrase “red hat” and a photograph of a red hat are assigned nearby numerical vectors [4].

Multimodal foundation models such as GPT-4o went further. Building on large language models, they can encode complex physical contexts, answer open-ended questions about a scene, or generate images from scratch. These two strands of AI research have since converged. Generative VLMs are now repurposed as highly effective embedding models capable of *retrieval*, finding a record in a database from its description [5]. Because these models grasp subtle semantic nuances, such as ‘man in a red hat’ versus a ‘red man in a hat’, adapting them yields vastly better retrieval accuracy than CLIP [6]. This eliminates the need for manual labelling, allowing investigators to simply type a prompt to ‘find the man with the red hat’.

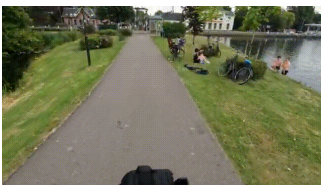


Figure 1: First result for the query ‘people chilling next to water’. Embedding models allows search with abstract words (‘chilling’) and contextual clues (‘next to water’).

Embedding-based semantic search To understand the potential of embedding-based search and its failure modes, and without access to surveillance cameras or commercial products, we worked as bricoleur programmers [7]. We reimplemented a state-of-the-art search system based on *SentrySearch*, an open-source software package that splits video into overlapping segments, embeds each into vectors, and lets users search the resulting database with free text. We modified it to use Qwen3-VL-Embedding, one of the most accurate, open-weights embedding model (*SentrySearch* did not support Qwen3-VL-Embedding at the time) [8]. The resulting system fits in the memory of a laptop (<5GB), while a well-equipped server can answer a query over 140 hours of embedded video in under 100 ms.

We first used a public outdoor video, filmed on a cycling journey in the Netherlands. Searching for ‘jogging’ and ‘people chilling next to water’ returned plausible video segments (see Figure 1). Searching for ‘traffic violation’, the second-best result showed an unsafe turn violating right-of-way. We then recorded a 20-second video clip of the first author (JvA) and two other people, with their consent, and derived embeddings from it. The query ‘white Caucasian male’ correctly retrieved the only segments where JvA appeared, as did a query consisting of a photograph of JvA’s face embedded into the same vector space. A database built for semantic search, running on conventional hardware, thus already functioned, unmodified, as a tool for biometrics surveillance. Not everything in a scene is encoded equally. In footage of a busy public street, we could retrieve a large, slow-moving target such as a bus but not a small, fast-moving target (JvA on a bicycle).

Adoption of embedding-based systems There is significant political pressure to adopt AI systems for video surveillance in public spaces [9], and attempts to slow or regulate facial recognition may displace that demand into other technologies such as embedding-based systems. Industry is already moving this way: BriefCam lets operators track people by attributes such as clothing colour [10]; Avigilon’s Visual Alerts detects events such as a fallen person from natural-language descriptions [11]; IvedaAI advertises forensic queries such as ‘a person wearing a bandana’ as going ‘beyond pre-trained object models’ [12].

Embedding models are a noteworthy fit for policing and surveillance in Europe. They allow retrospective rather than live surveillance, placing them outside the AI Act’s prohibition on real-time biometric identification (Art. 5(1)). They are also general-purpose, and a system for ‘video search’ may be easier to procure than one for ‘facial recognition’. Once expensive to run, open-weights models now match proprietary performance and, once embedded, videos can be searched indefinitely at very little cost [13].

Whether the AI Act captures these systems is unclear. The Commission’s draft guidelines on high-risk classification, published in May 2026, suggest it might not. The draft states that systems searching for objects ‘based on number plates, specific clothes, weapons, background, etc.’ are not high-risk under Annex III [14]. It notes that ‘an AI system following a person with a red scarf or a blue jacket is not based on biometric data.’ Yet, as our experiments show, a database that returns ‘person with a red scarf’ also returns ‘white Caucasian male’ and even matches a photograph of a specific face. Should we accept these object-search systems as non-biometric, when an operator can query for biometric features at any point?

Open Questions Embedding systems outsource interpretation to models. A query for a ‘red car’ has a verifiable answer, but a query for an ‘unsafe overtake’ or ‘loitering teenager’ does not. We believe there are important, open questions about how we should conceptualise, measure, and contest machine interpretation of judgment-laden queries, and what meaningful contestability look like. Retrospective search also threatens to deepen incidental privacy costs of dashcams, doorbell cameras, and municipal video surveillance, by finally assembling them together [15]. Chilling effects on assembly are already documented for video surveillance with, e.g., younger Londoners more likely to avoid events monitored by facial recognition [16]. We hope the SUP community finds this emerging topic of interest and joins us in identifying what can be done.

References

1. Murray, J. & Booth, R. *AI facial recognition oversight lagging far behind technology, watchdogs warn* 3rd May 2026. <https://www.theguardian.com/technology/ng-interactive/2026/may/03/ai-facial-recognition-oversight-lagging-far-behind-technology-watchdogs-warn> (2026).
2. Disclose. *La reconnaissance faciale déployée à grande échelle sur les téléphones des forces de l'ordre* <https://disclose.ngo/fr/article/la-reconnaissance-faciale-deployee-a-grande-echelle-sur-les-telephones-des-forces-de-lordre> (2026).
3. Wright, R. Police in England and Wales Told to Halt AI Use in Court Statements. *Financial Times*. *UK police*. <https://www.ft.com/content/229e5949-3ebc-4151-8a86-a01b5e259241> (2026) (6th June 2026).
4. Radford, A. et al. *Learning Transferable Visual Models From Natural Language Supervision* in *Proceedings of the 38th International Conference on Machine Learning* (eds Meila, M. & Zhang, T.) 139 (PMLR, July 2021), 8748–8763. <https://proceedings.mlr.press/v139/radford21a.html>.
5. Shanbhogue, M. et al. *Gemini Embedding 2: A Native Multimodal Embedding Model from Gemini* arXiv: 2605.27295 [cs.CV]. <http://arxiv.org/abs/2605.27295>.
6. Yuksekgonul, M., Bianchi, F., Kalluri, P., Jurafsky, D. & Zou, J. *When and why vision-language models behave like bags-of-words, and what to do about it?* arXiv: 2210.01936 [cs.CV]. <http://arxiv.org/abs/2210.01936>.
7. Turkle, S. & Papert, S. Epistemological pluralism and the revaluation of the concrete. *J. Math. Behav.* **11**, 3–33 (1 1992).
8. Li, M. et al. *Qwen3-VL-embedding and Qwen3-VL-Reranker: A unified framework for state-of-the-art multimodal retrieval and ranking* arXiv: 2601.04720 [cs.CL]. <http://arxiv.org/abs/2601.04720>.
9. Yesilgöz-Zegerius, D. & Bontenbal, H. 28 684 *Naar een veiliger samenleving* <https://www.tweedekamer.nl/kamerstukken/kamervragen/detail> (2026).
10. Milestone Systems A/S. *BriefCam* <https://www.milestonesys.com/products/software/briefcam/>. (2026).
11. Avigilon. *Visual Alerts: Turn everyday language into powerful, custom alerts* https://www.avigilon.com/fs/documents/Fact-sheet_Visual-Alerts.pdf.
12. *IvedaAI™ Breakthrough in Natural Language Opens New Markets and Revenue Opportunities* <https://ivedaai.com/ivedaai-breakthrough-in-natural-language-opens-new-markets-and-revenue-opportunities/> (2026).
13. Golden, A., Wu, C.-J., Wei, G.-Y. & Brooks, D. *The xPU-athalon: Quantifying the Competition of AI Acceleration* arXiv: 2604.10852 [cs]. <http://arxiv.org/abs/2604.10852> (2026).
14. European Commission. *Draft Commission guidelines on the classification of high-risk AI systems under Article 6 of Regulation (EU) 2024/1689 (AI Act) for stakeholder consultation* <https://digital-strategy.ec.europa.eu/en/library/draft-commission-guidelines-classification-high-risk-ai-systems>.
15. Haggerty, K. D. & Ericson, R. V. The surveillant assemblage. *Br. J. Sociol.* **51**, 605–622 (4 2000).
16. London Policing Ethics Panel. *Final Report on Live Facial Recognition* http://www.policingethicspanel.london/uploads/4/4/0/7/44076193/live_facial_recognition_final_report_may_2019.pdf.