

Bringing a Sense of Smell to AI Development

Originally published at <https://www.unite.ai/bringing-a-sense-of-smell-to-ai-development/>



A new AI dataset teaches machines to smell by associating smell data with images, letting models match odors to objects, scenes, and materials.

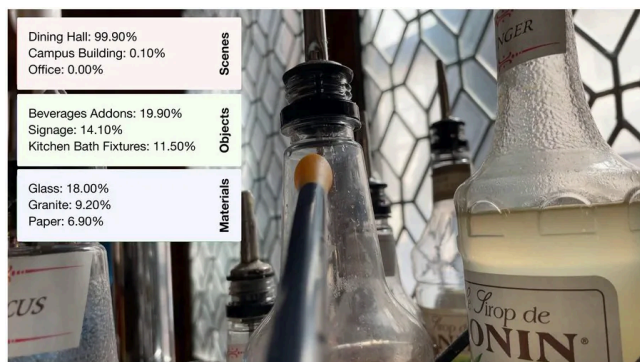
Perhaps because smell-output machines have [such a checkered history](#), olfaction is a fairly neglected sense in AI research literature. Unless you were planning on producing another entry in the long-running (more than a century, to date) [smell-o-vision saga](#), use cases have always seemed rather 'niche' in comparison to the potential exploitation of image, audio and video datasets, and the AI models that are trained from them.

In fact, the possibility of automating, industrializing and popularizing the kind of detection facilities offered by [bomb dogs](#), [cadaver dogs](#), [illness sniffer dogs](#), and diverse other types of canine sniffer units, would be a notable benefit in municipal and security services. Despite high demand, far [in excess of supply](#), training and maintaining detection dogs is an [expensive business](#) that does [not always offer](#) good value for money.

To date, most of the research that encroaches on this area of study has been confined to a lab, with curated collections typically comprised of examples with hand-crafted [features](#) – a profile inclined more towards bespoke cottage-industry solutions than industrialized applications.

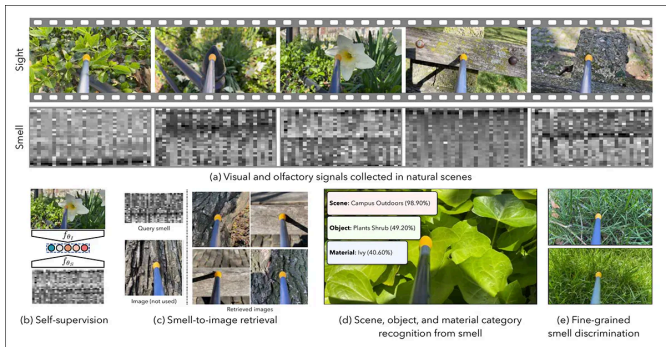
Ahead by a Nose

Into this rather fusty climate comes an interesting new academic/industry collaboration from the US, wherein a team of researchers spent several months cataloguing diverse smells in indoor and outdoor environments in New York City – and, for the first time, gathering images associated with the captured odors:



Note the central sensor, the 'nose' of the olfactory device. Trained only on smell, the model guesses whether it's sniffing granite, plastic, or leather – and even identifies the room it's in, without seeing a single pixel. [Source](#)

This research has led the authors of the new work to devise a spin on the hugely popular Contrastive Language-Image Pretraining ([CLIP](#)) framework, which connects text and images, in the form of *Contrastive Olfaction-Image Pretraining* (COIP) – which connects *smells* and images.



Top: synchronized video and olfactory sensor data are captured in natural settings using a camera-e-nose rig. Bottom left (b): a joint embedding is learned through cross-modal self-supervision. (c): the system retrieves visual matches based solely on a query smell. (d): individual smell samples are used to classify environment, object, and material categories. (e): highly similar odors, such as two types of grass, are distinguished without visual input. [Source](#)

The new dataset, titled *New York Smells*, contains 7,000 smell-image pairings featuring 3,500 different objects. When trained in tests, the new data was found to outperform the popular hand-crafted features in the relatively small number of similar prior datasets.

The authors hope that their initial outing will open the way for later and follow-on work towards olfactory detection systems designed to operate in the wild, in much the same way sniffer dogs do*:

'We see this dataset as a step toward in-the-wild, multimodal olfactory perception, as well as a step toward linking sight with smell. While olfaction has traditionally been approached in constrained settings, such as quality assurance, there are many applications in natural settings.

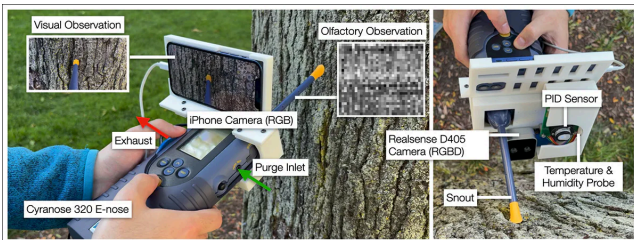
'For example, as humans, we constantly use our sense of smell to assess the quality of food, identify hazards, and detect unseen objects.

'Moreover, many animals, such as dogs, bears, and mice, show superhuman olfaction [capabilities](#), suggesting that human smell perception is far from the limit of machine abilities.'

Though the [new paper](#), titled *New York Smells: A Large Multimodal Dataset for Olfaction*, promises that data and code will be released, a 27GB data file is already available via the paper's [project site](#). The paper was produced by nine researchers across Columbia University, Cornell University, and Osmo Labs.

Method

To collect material for the new collection, the researchers used the [Cyrano 320](#) electronic nose, with an iPhone mounted above the forward intake to capture visually what smells were being registered:



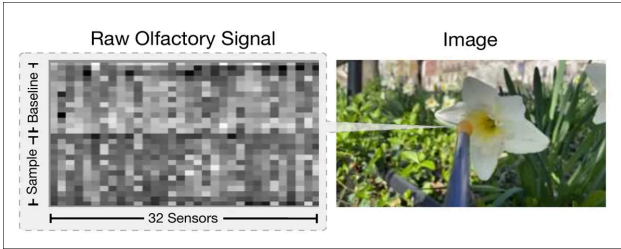
A handheld sensor rig collects paired video and smell data by mounting an iPhone camera onto a Cyranose 320 e-nose. The snout is aimed at objects while the exhaust and purge inlet manage airflow during sampling. An RGB-D camera captures depth, while Volatile organic compound (VOC) concentration, temperature, and humidity are recorded through integrated sensors including a Proportional-integral-derivative (PID) module and environmental probe.

The Cyranose device runs at 2Hz, recording 32-dimensional olfactory time-steps. Volatile Organic Compound (VOC) concentrations were recorded with a [MiniPID2 PPM WR](#) sensor.

The portable unit functioned as a nimble sensor, relaying data to a more compute-capable mobile station for processing.

To place the target smell in context, a 'baseline smell' was registered, before the more specific object was targeted directly with the 'snout' of the Cyranose. The ambient sample was then taken from a side-port in the unit, to ensure it was distant enough from the main odor source to not be contaminated.

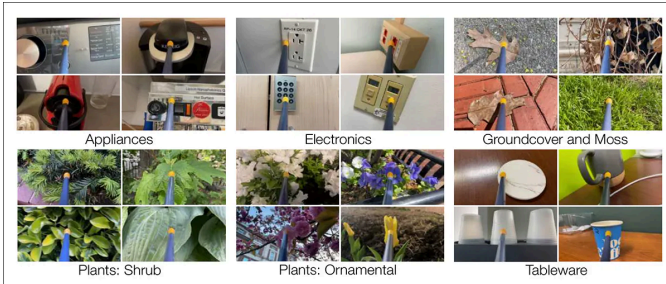
Two samples were taken through the main intake of the sensor, with each ten-second recording captured from a different position around the object, to improve data efficiency. The samples were then combined with the ambient baseline to form a 28×32 matrix, representing the full olfactory measurement:



This example shows the signal and corresponding image for a flower. The full olfactory signal consists of a 28×32 matrix, combining a 14-frame ambient baseline with two 10-second samples taken from different angles around the target object.

Data and Tests

Vision Language Models (VLMs) were used to automatically label objects and materials captured by the iPhone in the Cyranose rig, with GPT-4o utilized for the task; however, scene categories were manually labelled:



A small sample from an extensive illustration in the source paper detailing the varied smell sources and environments captured in the project.

The dataset was divided into training and validation splits, with both samples from each object assigned to the same split to avoid cross-contamination. The final collection comprises 7,000 olfactory-vision pairs drawn from 3,500 unlabeled objects, along with 70 hours of video and 196,000 time-steps of raw olfactory data from both baseline and sample phases.

Data was collected across 60 sessions over a two-month period, spanning parks, university buildings, offices, streets, libraries, apartments, and dining halls, with multiple sessions conducted at each location. The resulting dataset contains 41% outdoor, and 59% indoor environments.

To develop general-purpose olfactory representations, the authors trained a contrastive model to associate synchronized image-smell pairs from the dataset. This approach, the aforementioned COIP, uses a loss function adapted from CLIP to align the embeddings of co-occurring visual and olfactory signals.

Training used both a visual encoder and a smell encoder, with the goal of teaching the model to bring matching smells and images closer together in a shared representation space. The resulting representations support a range of downstream tasks, including smell-to-image retrieval, scene and object recognition, material classification, and fine-grained odor discrimination.

The model was trained using two types of olfactory inputs: the full raw sensor signal and a reduced hand-crafted summary known as smellprints – widely-used features in olfaction research that compress each sensor’s response into a single number by comparing the peak resistance during sampling with the average resistance during the ambient baseline.

By contrast, the raw input recorded all over NYC consists of a time-series from 32 chemical sensors inside the Cyranose device, capturing how each sensor’s electrical resistance changed over time as it reacted to the odor.

For curation of the dataset, this unprocessed signal was fed directly into a neural network, allowing for end-to-end learning with either a convolutional or transformer-based backbone. Models were trained using both smellprints and the raw input gathered from various environs in New York City, with both input types evaluated using contrastive learning.

Cross-Modal Retrieval

Cross-modal retrieval was evaluated by embedding each smell sample and its paired image into a shared representation space, and testing whether the correct image could be retrieved based solely on the olfactory input.

Ranking was determined by the proximity of each image embedding to the query smell within this space, and performance measured using mean rank, median rank, and recall at multiple thresholds:

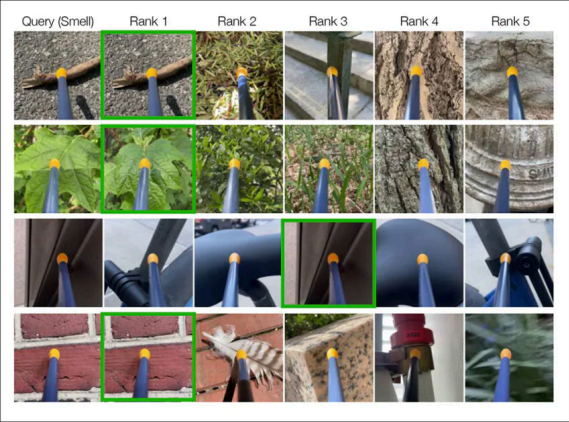
Smell Encoder	Mean Rank ↓	Median Rank ↓	Recall @ 5 (%) ↑	Recall @ 10 (%) ↑	Recall @ 20 (%) ↑
Chance	467	467	0.536	1.07	2.14
MLP (Smellprint)	375.9	329	2.04	3.43	6.22
CNN (Raw Smell)	118.4	41	12.9	21.1	32.6
MLP (Raw Smell)	159.5	56	17.3	24.2	33.8
Transformer (Raw Smell)	104.0	28	16.5	29.6	43.1

Cross-modal retrieval accuracy for different smell encoders, showing how well each model identifies the correct image from a smell query. The results compare architectures trained on raw olfactory signals with those using smellprints.

Regarding these results, the authors state:

‘Contrastive pretraining using smellprint performs better than chance in all metrics. However, training the olfactory encoder on the raw olfactory signal leads to significant improvement compared to the smellprint encoder, independent of architecture.

‘This shows the richer information present in the raw olfactory data, unlocking stronger cross-modal associations between sight and smell.’



A detail from the seventh illustration in the source paper, which is too condensed to reproduce meaningfully here. Here, cross-modal retrieval examples showing how the model links smells to matching images. Each row begins with a smell query, followed by the top ranked image predictions in the shared embedding space. The correct image for each query is outlined in green, illustrating how odors from books, plants, masonry, and other materials pull the model toward visually and semantically related scenes.

The authors note also that retrieval results showed clear semantic patterns:

‘Retrievals from our model often show semantic groupings. The odor of a book retrieves images of other books, the odor of leaves retrieves images of foliage.

‘These results suggest that the learned representation captures meaningful cross-modal structure.’

Scene, Object and Material Recognition

The ability of the model to recognize smells without visual input was evaluated by training it to identify scenes, objects, and materials based solely on olfactory data; to this end, a linear probe (a simple classifier trained on [frozen](#) representations) was used to assess how much information was encoded in the learned smell embeddings.

Labels were derived from the paired images in the training set using GPT-4o – but only the olfactory signal was used during classification.

Several encoder types were tested: some initialized randomly, some trained from scratch, and others trained using contrastive learning to align smell and vision in a shared representation space, with raw data and smellprints evaluated:

Method	Input	Scenes			Materials			Objects		
		Scratch	SSL	Rand	Scratch	SSL	Rand	Scratch	SSL	Rand
Chance		12.5	12.5	12.5	1.9	1.9	1.9	2.0	2.0	2.0
MLP	Smellprint	42.2	32.5	31.4	3.84	2.0	6.0	3.3	4.96	5.9
Transformer	Raw smell	91.0	90.4	72.7	2.33	14.0	7.1	13.8	18.4	12.3
CNN	Raw smell	99.5	95.0	74.5	11.9	12.3	9.4	17.9	19.8	8.7

Classification accuracy for scenes, materials, and objects was assessed using olfactory signals alone. Raw sensor input outperformed smellprints, with CNNs trained from scratch yielding the highest results, including 99.5% for scenes. SSL pretraining helped in some cases, but was generally surpassed by supervised training. Random-weight baselines indicate that model capacity alone proves insufficient.

Significantly higher accuracy was obtained when raw olfactory data was used, especially in models trained with cross-modal supervision. The authors comment**:

‘Models trained on raw sensory inputs also achieve higher accuracy than models trained with the hand-crafted smellprint features. These results show that deep learning from raw olfaction signals is significantly better than hand-crafted features.’

Fine-Grained Discrimination

To assess whether fine-grained odor distinctions could be learned, a benchmark was built from two grass species coexisting on the same campus lawn. Alternating samples were collected over six 30-minute sessions, yielding 256 examples. A [linear classifier](#) was trained on features from olfactory-visual contrastive learning, and evaluated on a held-out set of 42 samples:

Method	Input	Accuracy
Chance		50.0
Random weights	Smellprint	66.7
Trained from scratch	Smellprint	85.7
SSL + linear probe	Smellprint	90.0
Random weights	Raw smell	47.6
Trained from scratch	Raw smell	52.4
SSL + linear probe	Raw smell	92.9

Accuracy of grass species classification from smell alone. Models were evaluated on their ability to distinguish between two visually similar grass types using only olfactory input. Performance was compared across smellprints and raw sensor data, with models either randomly initialized, trained from scratch, or trained using self-supervised learning (SSL) followed by a linear probe. The highest accuracy, 92.9%, was achieved using raw olfactory signals with SSL, indicating that fine-grained odor differences are best captured through raw input and vision-guided training.

Here the researchers state:

'Training on the raw olfactory sensor signal (instead of hand-crafted features) yields the highest accuracy – exceeding all variants based on smellprints.'

'These results suggest that olfactory-visual learning preserves more fine-grained information than learning with smellprints, and that visual supervision provides a signal for exploiting this information.'

Conclusion

Though odor synthesis seems likely to remain an unsolved problem some way into the future, an effective and affordable in-the-wild odor *analysis* system has enormous potential, not only for police, security and medical purposes, but also for quality-of-life and urban monitoring.

At the present time, the equipment involved is niche and usually quite expensive; therefore real progress in 'olfactory AI' for detection seems likely to require a visionary and affordable sensor in the Raspberry PI spirit.

* My conversion of the authors' inline citations to hyperlinks.

** Please note that further illustrations (figure 8) are available in the source paper, but are best viewed in that context.

First published Friday, November 28, 2025

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.

Portfolio site: martinanderson.ai

Contact: martin@martinanderson.ai



Discover More