Behavior Recognition via Sparse Spatio- Temporal Features

Piotr Dollár, Vincent Rabaud, Garrison Cottrell, Serge Belongie

{ pdollar, vrabaud, gary, sjb }@cs.ucsd.edu

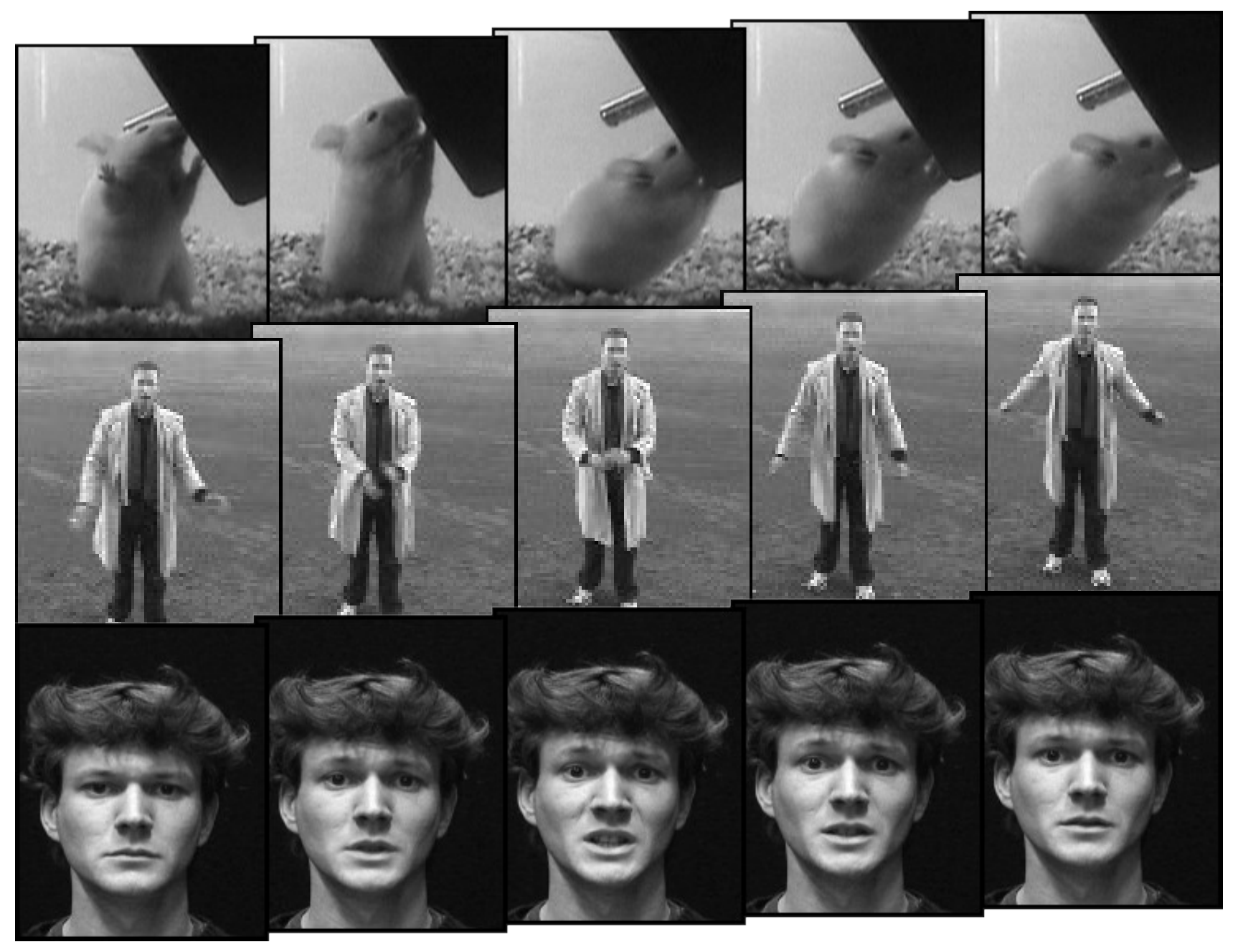
http://smartvivarium.calit2.net/

Abstract

A common trend in object recognition is to detect and leverage the use of sparse, informative feature points. The use of such features makes the problem more manageable while providing increased robustness to noise and pose variation. In this work we develop an extension of these ideas to the spatio-temporal case. For this purpose, we show that the direct 3D counterparts to commonly used 2D interest point detectors are inadequate, and we propose an alternative. Anchoring off of these interest points, we devise a recognition algorithm based on spatio-temporally windowed data. We present recognition results on a variety of datasets including both human and rodent behavior.

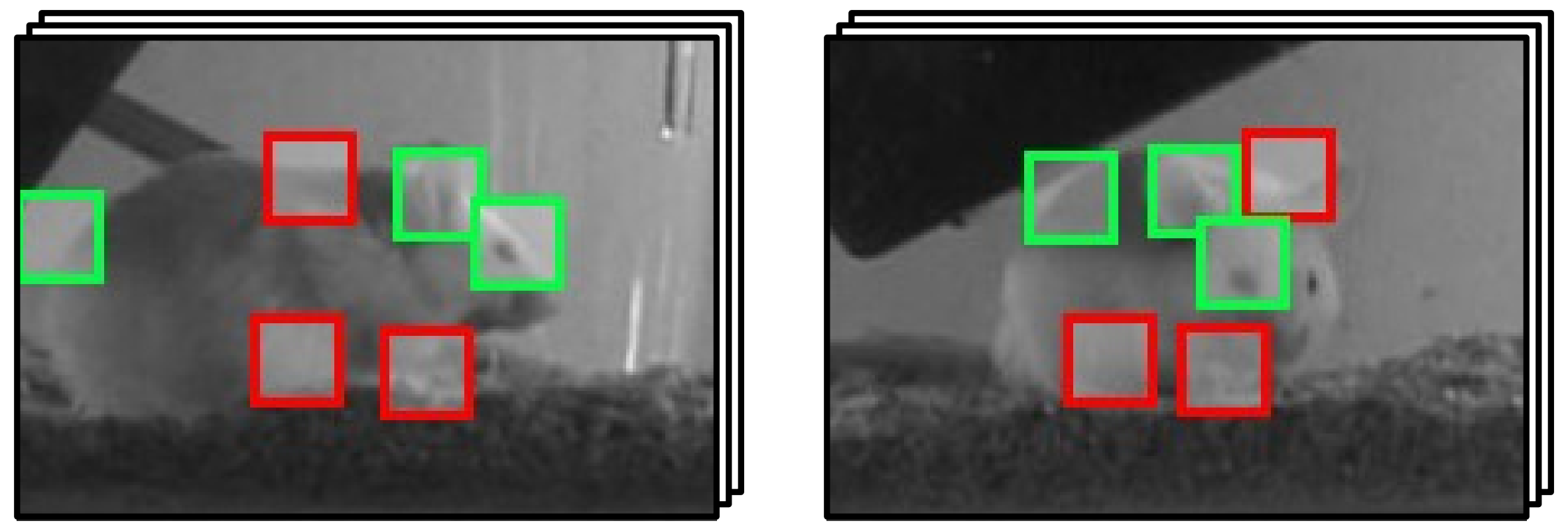
Goal

Our goal is to identify behavior from video, and more generally to create a simple yet effective representation of activity.



Domains: mouse behavior, human activity, facial expressions

Motivation



Posture, appearance, size and background can vary and parts may be occluded. However, there are often local informative regions of motion that are similar if subject is engaged in same behavior.

Feature Detection

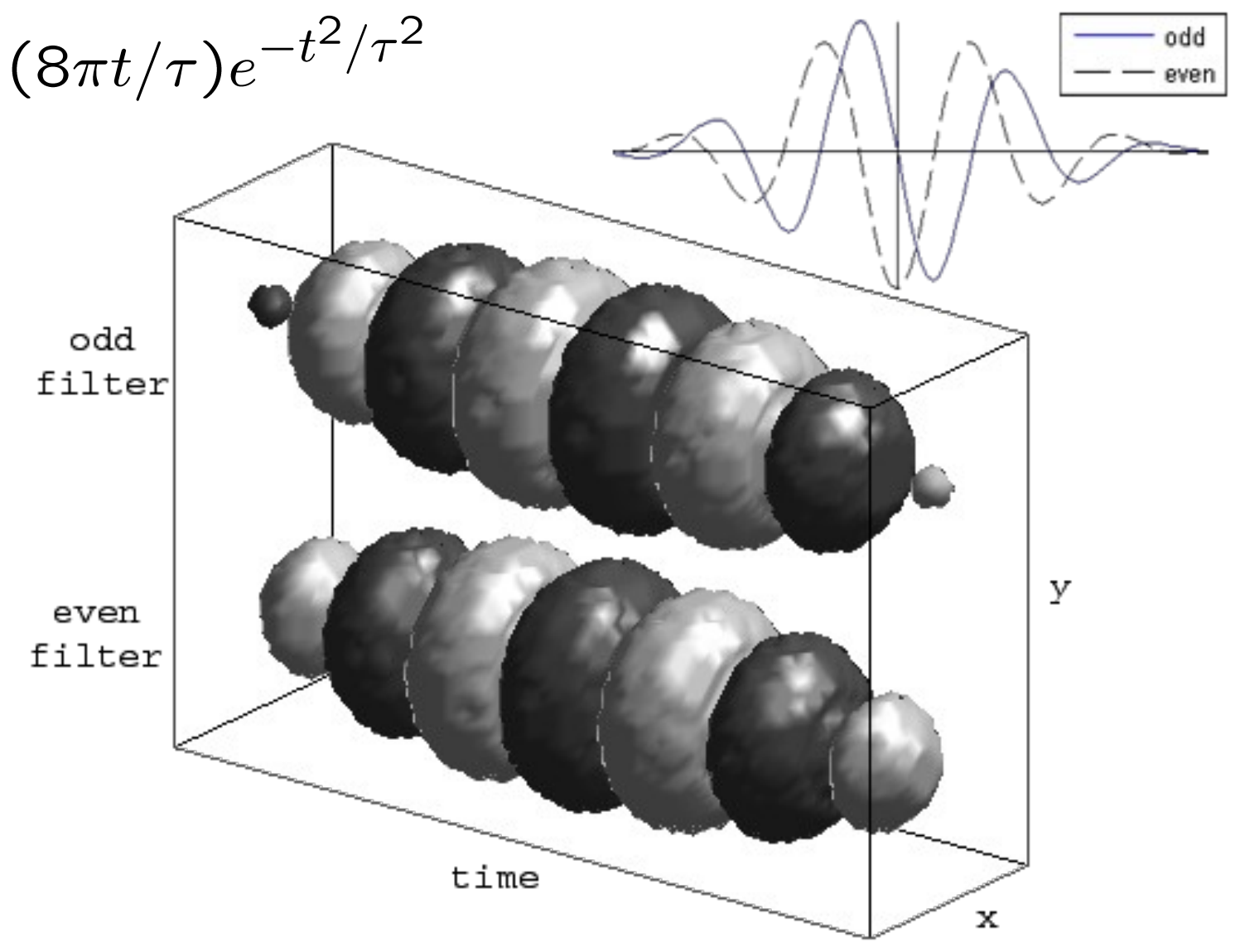
We calculate the response of the separable linear filters applied at every location in the video.

$$\mathsf{R} = (\mathbf{I} * g_x * g_y * h_{ev})^2 + (\mathbf{I} * g_x * g_y * h_{od})^2$$

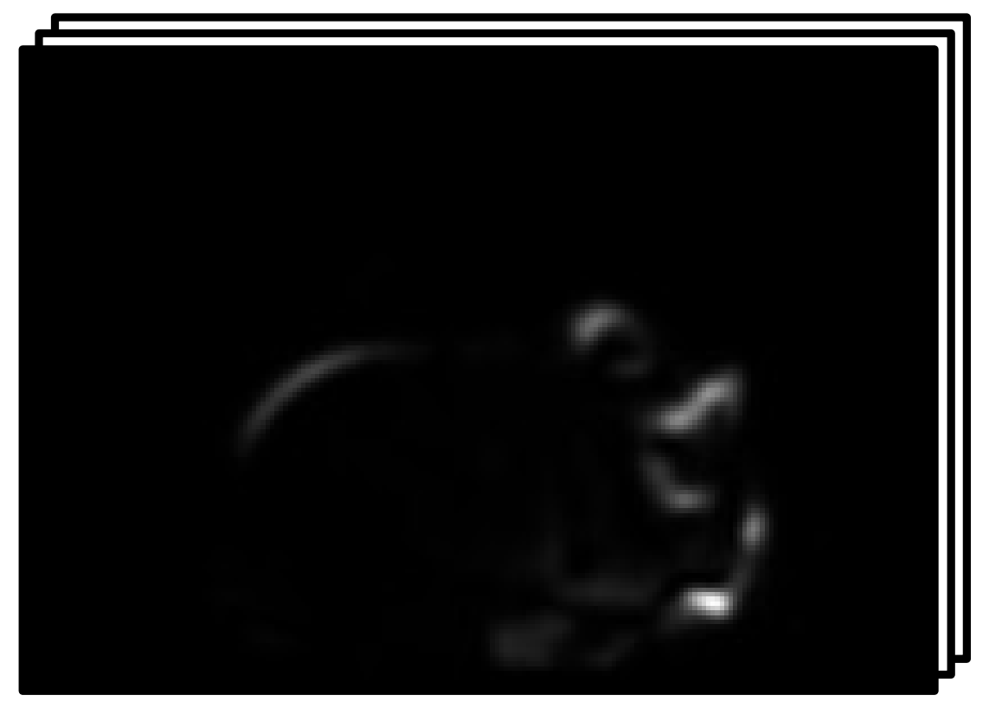
$$g(x;\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{(-x^2/2\sigma^2)}$$

$$h_{ev}(t;\tau) = -\cos(8\pi t/\tau) e^{-t^2/\tau^2}$$

$$h_{od}(t;\tau) = -\sin(8\pi t/\tau) e^{-t^2/\tau^2}$$

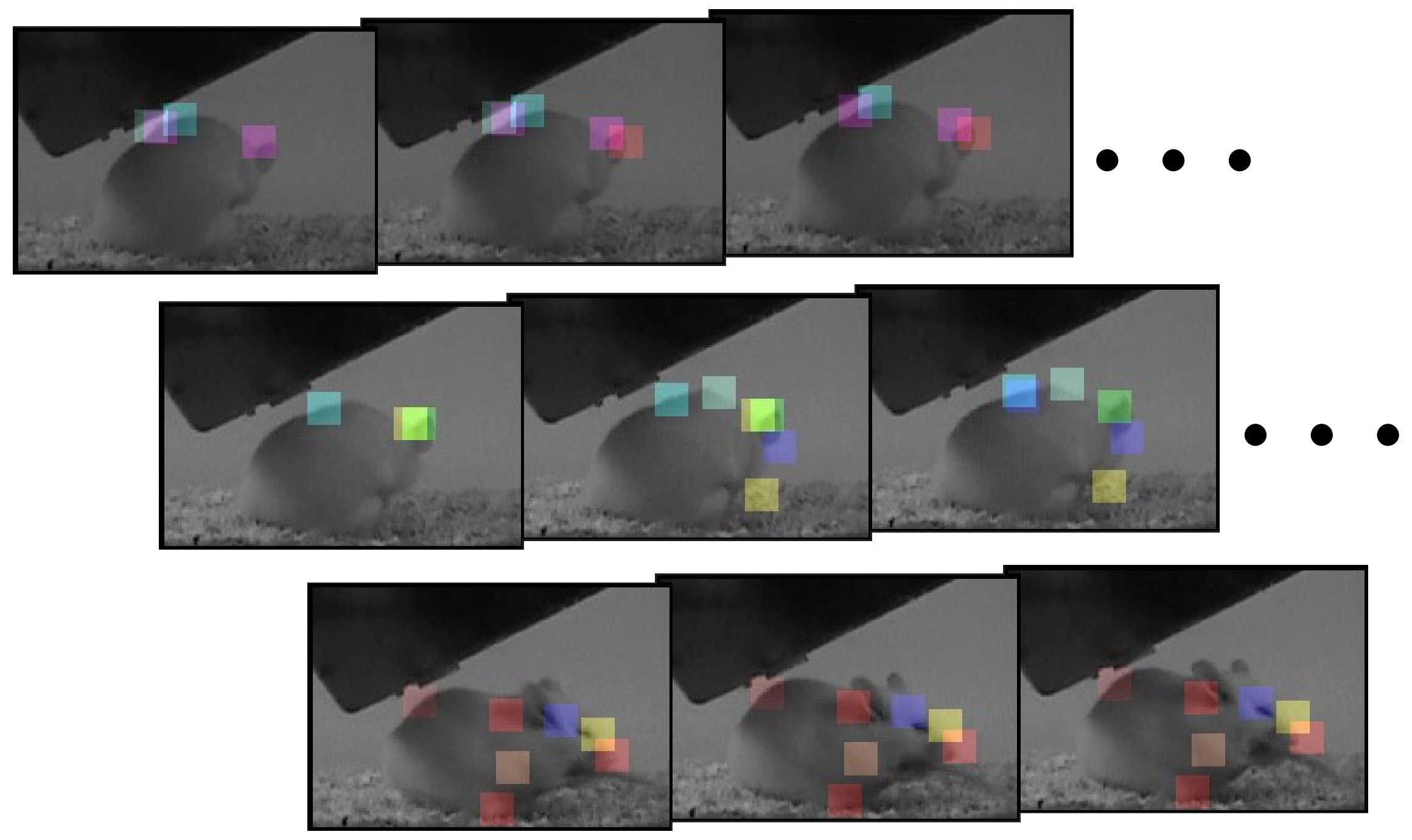


original video - $\mathbf{I}$



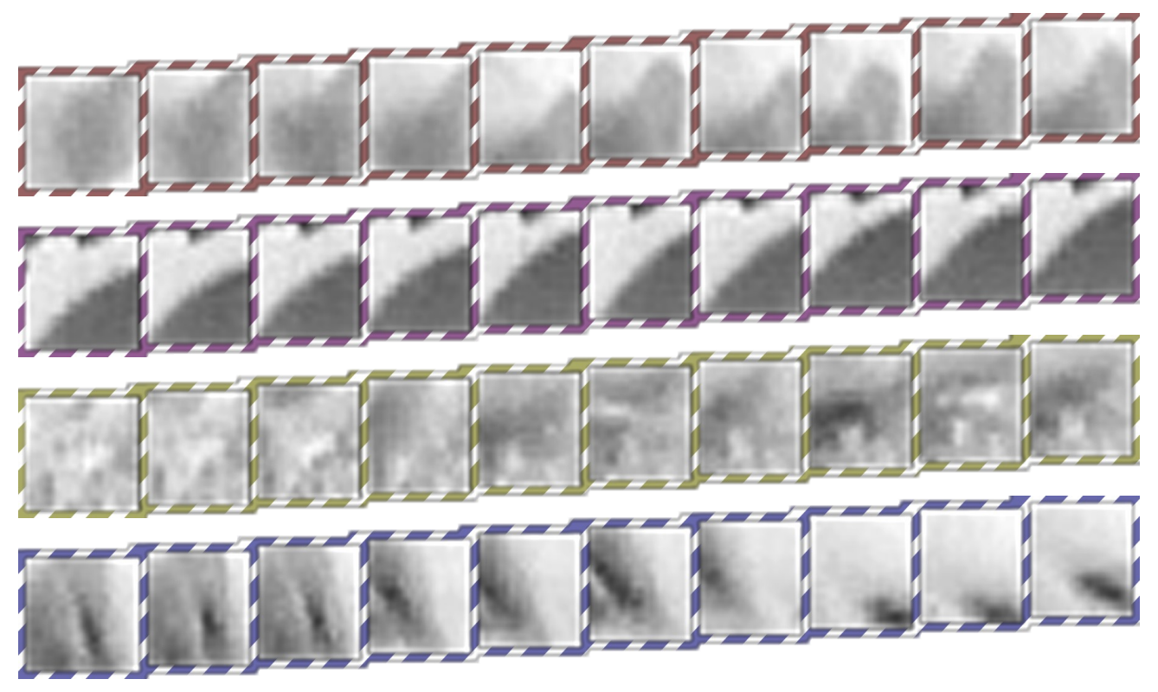
response - R

Cuboids



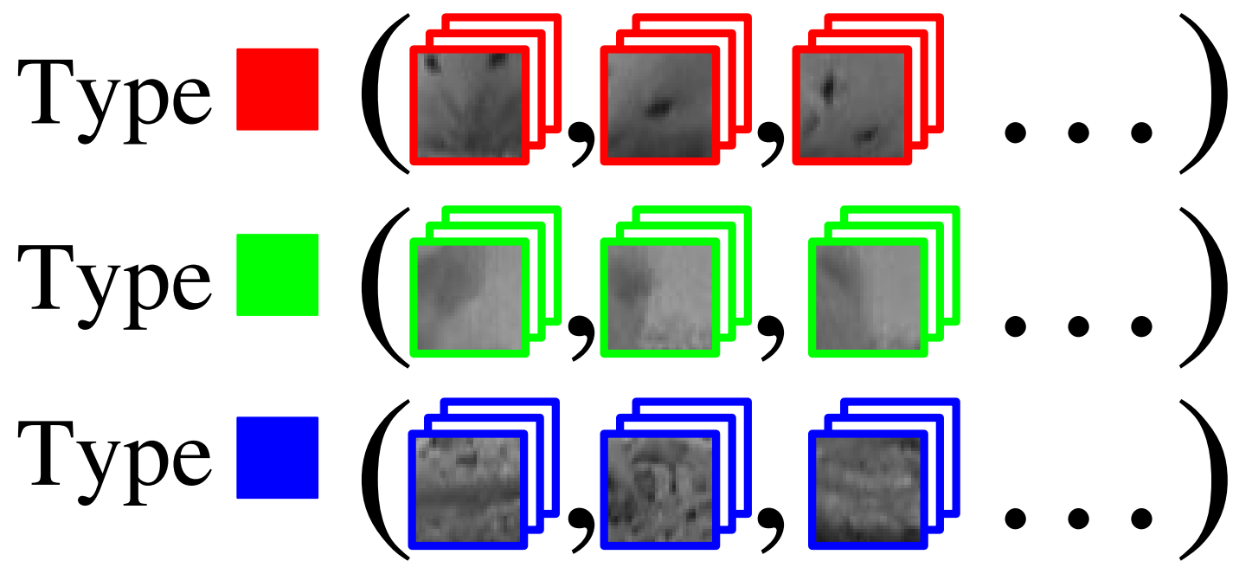
A cuboid (or right prism) of data is extracted around each feature point (local maximum of the response function). Each cuboid has spatial and temporal extend, represented here by a colored square that appears in a number of consecutive frames.

Cuboid Types



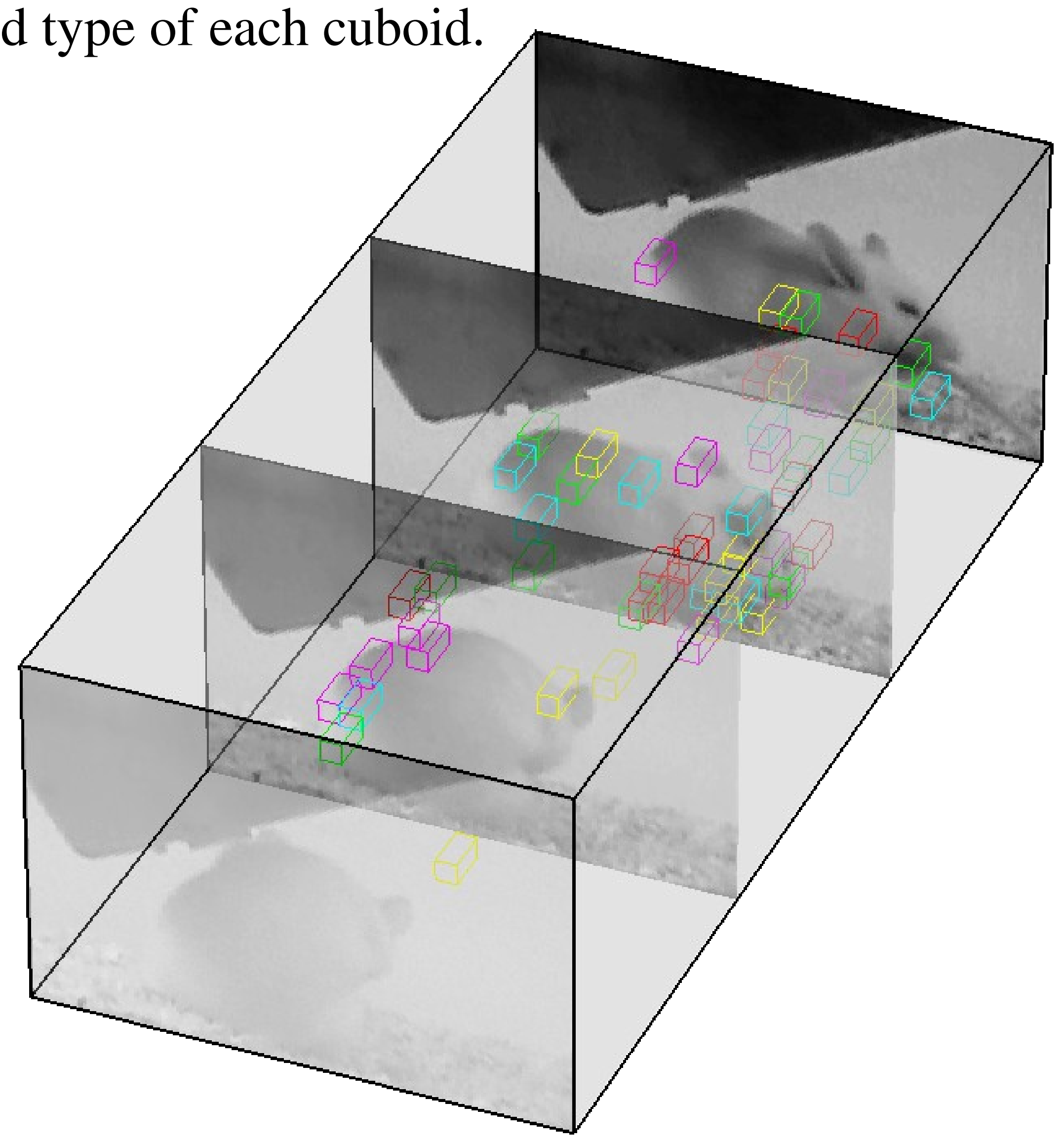
Extracted Cuboids (from top to bottom: ear, back, paw, eye)

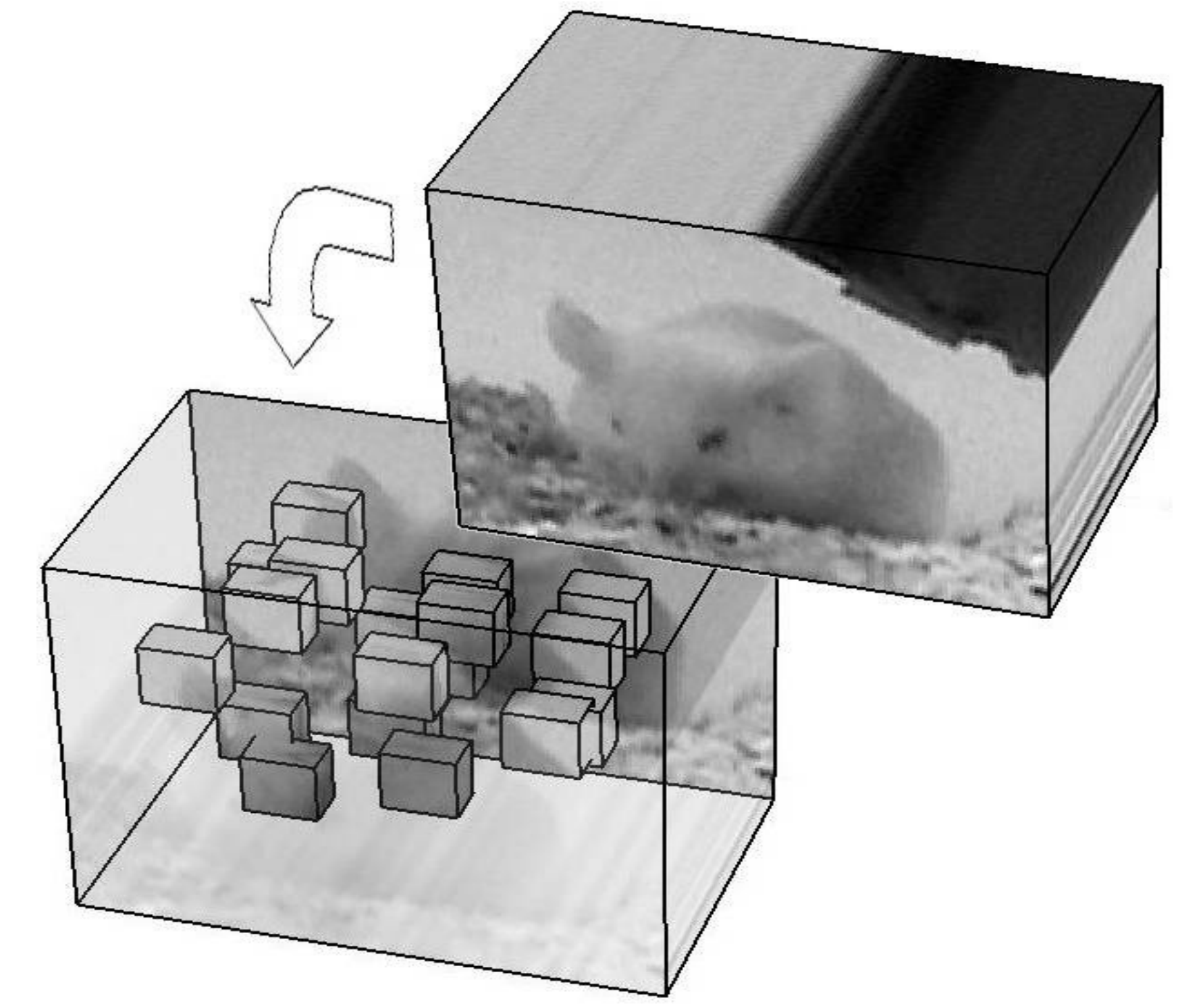
We assume that although number of possible cuboids is virtually unlimited, the number of different types of cuboids is relatively small. We generate the library of types for a given domain by clustering a large number of cuboids. Each cluster center represents a type.



Representation

After the library of cuboid types is generated, each subsequent cuboid detected is either assumed to be one of the known types or rejected as an outlier. The final representation includes only the location and type of each cuboid.



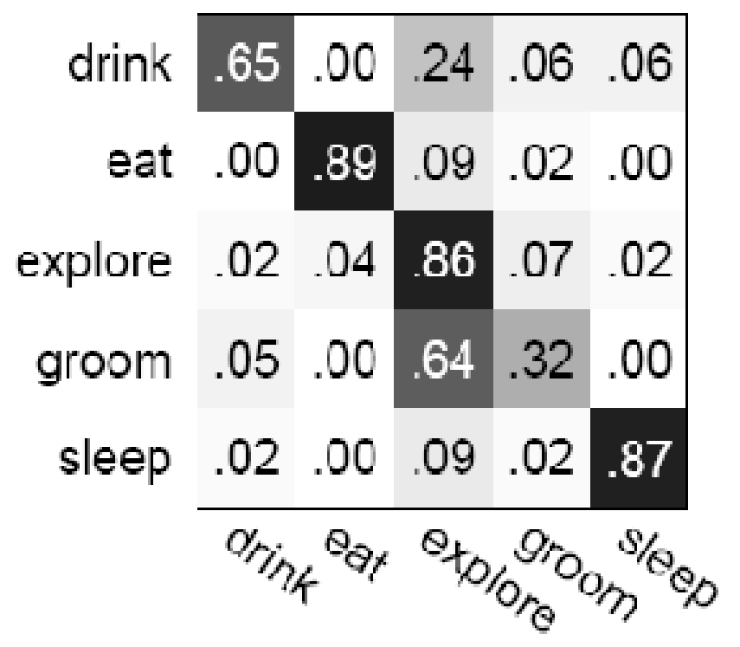


Results

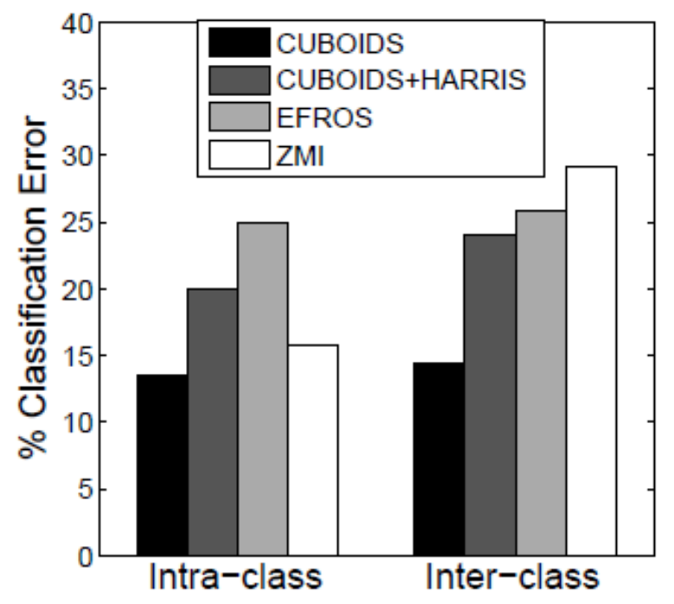
We histogram the cuboid types present in a video. The similarity of two behaviors is then calculated using the χ^2 distance between the two corresponding histograms.

We compare our method, which we refer to as CUBOIDS, to two other algorithms, referred to as EFROS and ZMI, and to a variant of our algorithm using Harris features. For details see our publication or the project website.

Mouse Behavior:

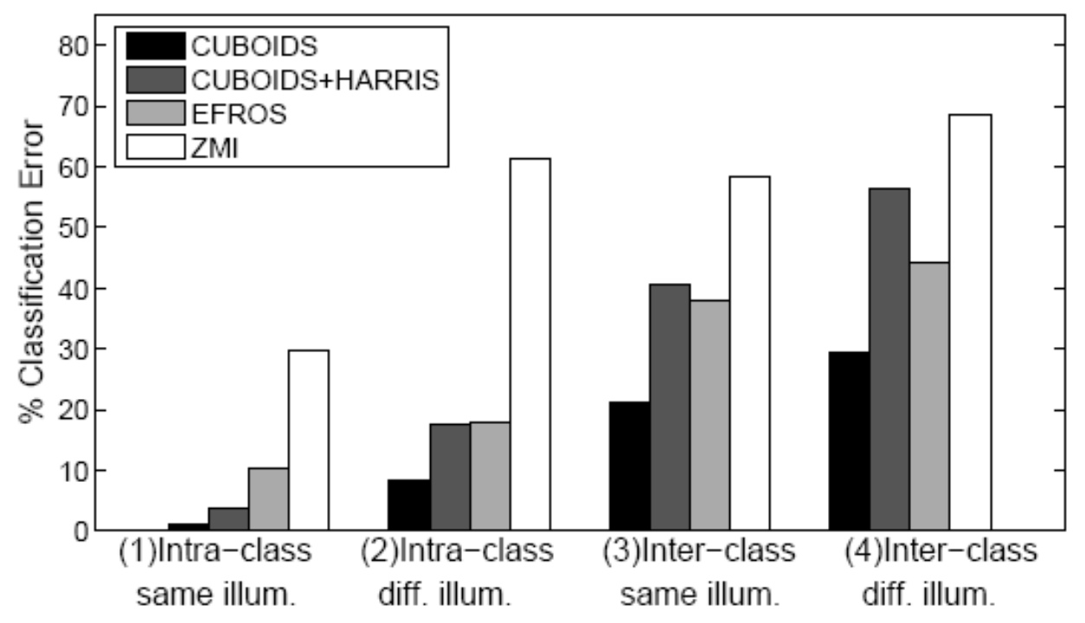


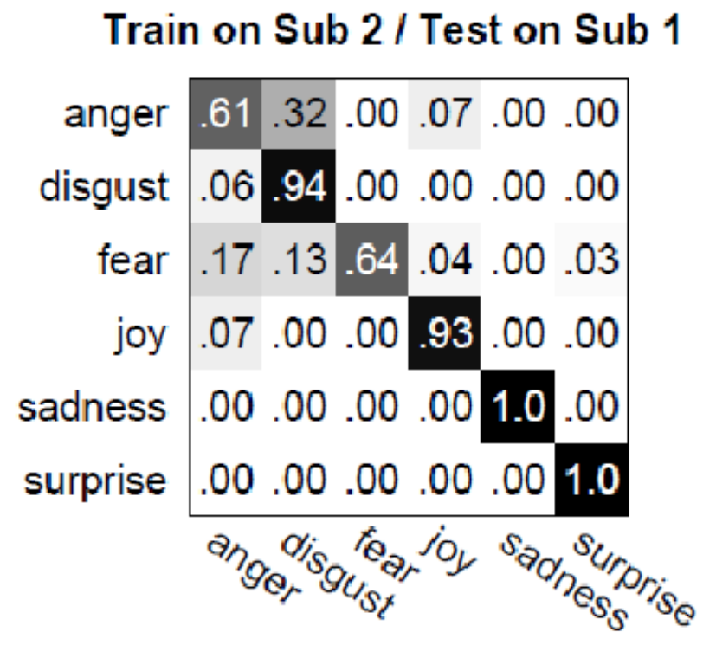
	drink	eat	explore	groom	sleep
drink	.65	.00	.24	.06	.06
eat	.00	.89	.09	.02	.00
explore	.02	.04	.86	.07	.02
groom	.05	.00	.64	.32	.00
sleep	.02	.00	.09	.02	.87



Left: Confusion matrix generated by CUBOIDS on the full mouse dataset. *Right:* Due to the form of the data, a full comparison of algorithms was not possible. Instead, we created a simple small scale experiment and ran all four algorithms on it, CUBOIDS achieved the best results.

Facial Expressions:



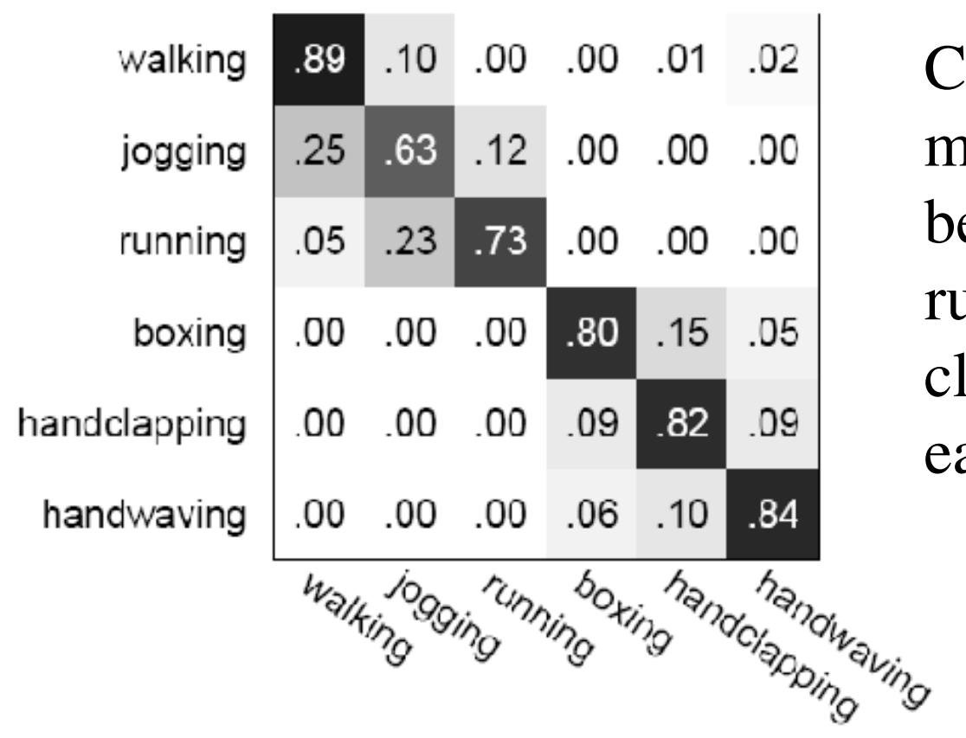


Train on Sub 2 / Test on Sub 1

	anger	disgust	fear	joy	sadness	surprise
anger	.61	.32	.00	.07	.00	.00
disgust	.06	.94	.00	.00	.00	.00
fear	.17	.13	.64	.04	.00	.03
joy	.07	.00	.00	.93	.00	.00
sadness	.00	.00	.00	.00	1.0	.00
surprise	.00	.00	.00	.00	.00	1.0

Left: Our method outperformed all competing algorithms under a number of setups. *Right:* A typical confusion matrix generated by our algorithm – the two subjects had different ways of expressing anger and fear, other expressions were similar.

Human Activity



	walking	jogging	running	boxing	handclapping	handwaving
walking	.89	.10	.00	.00	.01	.02
jogging	.25	.63	.12	.00	.00	.00
running	.05	.23	.73	.00	.00	.00
boxing	.00	.00	.00	.80	.15	.05
handclapping	.00	.00	.00	.09	.82	.09
handwaving	.00	.00	.00	.06	.10	.84

Confusion matrices generated by our method. Most of the confusion occurs between jogging and walking or running, and between boxing and clapping, most other activities are easily distinguished.