

# Optimizing deep video representation to match brain activity

Hugo Richard, Ana Pinho, Bertrand Thirion, Guillaume Charpiat

▶ **To cite this version:**

Hugo Richard, Ana Pinho, Bertrand Thirion, Guillaume Charpiat. Optimizing deep video representation to match brain activity. CCN 2018 - Conference on Cognitive Computational Neuroscience, Sep 2018, Philadelphia, United States. hal-01868735

**HAL Id: hal-01868735**
**https://hal.science/hal-01868735**

Submitted on 7 Sep 2018

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Optimizing deep video representation to match brain activity

**Hugo Richard (hugo.richard@inria.fr)**
PARIETAL Team, INRIA, 1 Rue Honor d'Estienne d'Orves, 91120 Palaiseau, France

**Ana Luísa Pinho (ana-luisa.grilo-pinho@inria.fr)**
PARIETAL Team, INRIA, 1 Rue Honor d'Estienne d'Orves, 91120 Palaiseau, France

**Bertrand Thirion (bertrand.thirion@inria.fr)**
PARIETAL Team, INRIA, 1 Rue Honor d'Estienne d'Orves, 91120 Palaiseau, France

**Guillaume Charpiat (guillaume.charpiat@inria.fr)**
TAU team, INRIA, LRI, Paris-Sud University, France

## Abstract

The comparison of observed brain activity with the statistics generated by artificial intelligence systems is useful to probe brain functional organization under ecological conditions. Here we study fMRI activity in ten subjects watching color natural movies and compute deep representations of these movies with an architecture that relies on optical flow and image content. The association of activity in visual areas with the different layers of the deep architecture displays complexity-related contrasts across visual areas and reveals a striking foveal/peripheral dichotomy.

**Keywords:** deep learning; video encoding; brain mapping;

## Introduction

The understanding of brain functional architecture has long been driven by subtractive reasoning approaches, in which the activation patterns associated with different experimental conditions presented in event-related or block designs are contrasted in order to yield condition-specific maps (Poline & Brett, 2012). A more ecological way of stimulating subjects consists in presenting complex continuous stimuli that are much more similar to every-day cognitive experiences.

The analysis of the ensuing complex stimulation streams proceeds by extracting relevant features from the stimuli and correlating the occurrence of these features with brain activity recorded simultaneously with the presentation of the stimuli. The analysis of video streams has been carried in (Eickenberg, Gramfort, Varoquaux, & Thirion, 2017) or (Güçlü & van Gerven, 2015) using a deep convolutional network trained for image classification. More recently, (Güçlü & van Gerven, 2017) has used a deep neural network trained for action recognition to analyze video streams.

Like (Güçlü & van Gerven, 2017), we use a deep neural network trained for action recognition to extract video features and train a linear model to predict brain activity from these features. In contrast, our study is not restricted to dorsal stream visual areas but involves the whole brain, and the deep neural network we use is pretrained on the largest action recognition dataset available (Kay et al., 2017).

From the different layers of the deep neural networks, we build video representations that allow us to segregate (1) occipital and lateral areas of the visual cortex (reproducing the results of (Güçlü & van Gerven, 2015)) and (2) foveal and peripheric areas of the visual cortex. We also introduce an efficient spatial compression scheme for deep video features that allows us to speed up the training of our predictive algorithm. We show that our compression scheme outperforms PCA by a large margin.

## Methods

### Deep video representation

We use a deep neural network trained for action recognition to build deep representations of the Berkeley Video stimuli (Nishimoto et al., 2011). This material consists of more than four hours of color natural movies built by mixing video blocks of 5-15 seconds in a random fashion.

The deep network we use is called Temporal Segment Network (TSN) (Wang et al., 2016). Following an idea introduced in 2014 (Simonyan & Zisserman, 2014) it was intended to mimic the dorsal and ventral stream by separately processing raw frames and optical flow fields. We chose TSN for our experiments because it uses a much larger number of layers than the original network (which results in higher accuracy in action recognition) and that a version of TSN pretrained on Kinetics – a massive video dataset (300 000 unique video clips) with 400 different classes all describing a human and at least 400 videos per class – is publicly available. The network is trained to recognize human actions such as slack-lining, skateboarding, massaging feet, dancing zumba and dining.

The version of TSN we use in our experiments is based on Inception v3 (Szegedy, Vanhoucke, Ioffe, Shlens, & Wojna, 2016) for both streams where small networks are used as building blocks of the main large network (Lin, Chen, & Yan, 2013). Each stream in the TSN Network is composed of more than 40 convolution layers and a fully connected layer. The activities after the last layer represent the probability of belonging to each action class.

### Feature extraction

The raw frames encode information about pixels, and flow fields encode information about pixels displacements. Although flow fields and raw frames streams do not precisely disentangle spatial content and motion information in videos,

we may expect that the raw frames stream better represent local spatial features while the flow fields stream more efficiently convey dynamic information. Following (Eickenberg et al., 2017) we consider that the activation statistics in the first layers (the ones closer to those of the input) have a low level of abstraction, whereas the last layers (closer to the labels) represent high-level information. Therefore each activity in both streams can be considered as specific features or representations of the video.

If we were to extract all network activities of the Berkeley Video Dataset we would need to store more than 6 millions floats per frame in the dataset. Such a representation would be highly redundant. In order to keep the volume of data reasonable, in each stream we only focus on four convolutional layers $L_1, L_2, L_3, L_4$ ranked by complexity. We further compress the data using spatial smoothing, and use temporal smoothing so that we get one representation every two seconds of video, which allows us to match the acquisition rate of fMRI scanners.

## Regression

10 subjects were scanned while watching the color natural movies of the Berkeley Video Dataset. The fMRI images were acquired at high spatial resolution (1.5mm), from a Prisma Scanner, using Multi-band and IPAT accelerations (mb factor=3, ipat=2). These data are part of a large-scale mapping project on a limited number of participants, called Human Brain Charting. Data acquisition procedures and initial experiments run in this project are described in (Pinho et al., 2018). In order to link extracted deep video features to the internal representation of videos in each subject we use a simple linear model to fit their brain activity in each voxel.

The use of a very simple model allows us to posit that the performance of the predictive model from a particular video representation is mostly linked to the suitability of the video representation. Hence the performance of the algorithm can be seen as a measure of the biological suitability of the video representation.

We use a kernel ridge regression with an hyper-parameter setting the magnitude of the l2-penalization on the weights. The resulting prediction is obtained using a cross validation procedure (11 sessions are used for train, 1 for test and at least 5 different splits are considered). To set the value of the hyper-parameter, we use a 5-fold cross validation on the train set and consider 20 different values. During hyper parameter selection, we only focus on the visual cortex to make this computation efficient.

The chosen measure of performance of our prediction algorithm is the coefficient of determination $m_{cv}$. Let $\mathbf{y}_{pred}$ and $\mathbf{y}_{real}$ be the respectively the prediction of a voxel activity and the real voxel activity. Then

$$m_{cv}(\mathbf{y}_{pred}, \mathbf{y}_{real}) \;=\; 1 - \frac{\sum_{t=1}^{n_b}(\mathbf{y}_{pred}[t] - \mathbf{y}_{real}[t])^2}{\sum_{t=1}^{n_b}(\mathbf{y}_{real}[t] - \overline{\mathbf{y}_{real}})^2}$$

The metric used to select the best parameter is the number of voxels having a coefficient of determination $m_{cv}$ greater than $0.1$. This procedure leads to different parameter values depending on the chosen layer activities.

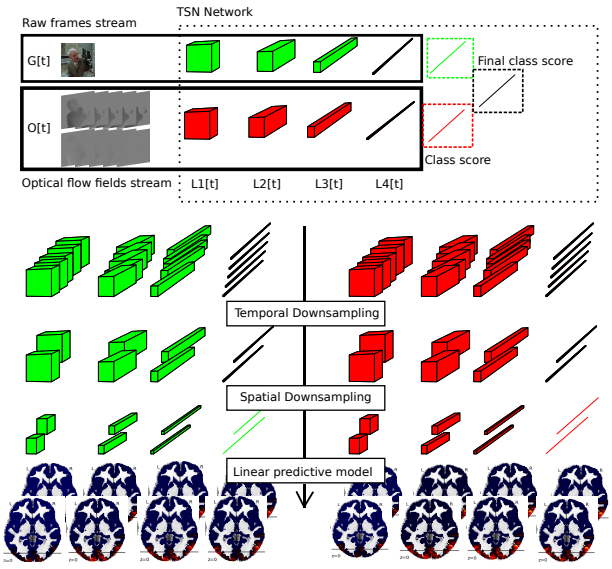



Figure 1: Feature extraction and regression scheme: at each time frame we compute and extract the activities of four layers $L_1, \cdots, L_4$ of the temporal segment network on a single frame and on a stack of 5 consecutive optical flow fields. The extracted activities are spatially and temporally down-sampled and then used to predict brain activity of subjects exposed to the video stimuli.

Figure 1 gives an overview of the pipeline used to extract and process deep video features to estimate the brain activity of subjects.

# Results

The extracted deep network features lead to different prediction performance depending on the down-sampling procedure, the stream used and the localization of target voxels.

## An efficient spatial compression scheme

We show that preserving the channel structure of the network during spatial compression procedure is key for developing an efficient compression scheme.

We compare three spatial compression schemes for network activities: (1) Standard principal component analysis (PCA) with $2000$ components; the transformation is learned on training sessions before it is applied to all sessions. (2) Average pooling inside channels (APIC) which computes local means of activities located in the same channel. (3) Average pooling inside and between convolution layers (APBIC) which is used to get the same number of output features for all layers while minimizing the number of convolutions between channels. It allows us to check that the performance of the predictive algorithm is not merely driven by the number of features.

The procedure for activities extraction, temporal down-

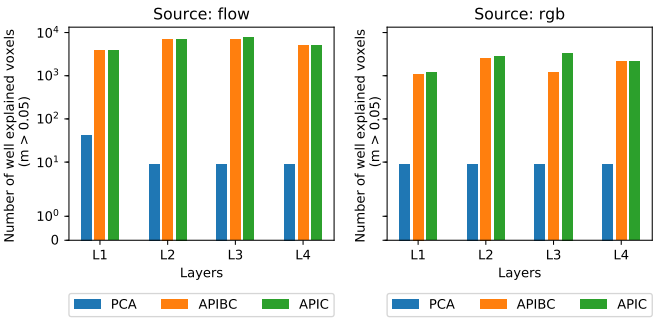



Figure 2: Comparison of the different neural network compression streams. The APIC approach slightly outperforms APIBC, and both strongly outperform PCA. When using APIC or APIBC we predict correctly up to 850 times more voxels than when using PCA.

sampling and brain activity prediction is not changed while the spatial compression scheme varies. The benchmark is performed using a leave-one-out cross-validation procedure with two splits in three subjects.

Figure 2 shows that both approaches preserving channel organization structure outperform PCA by a large margin.

These results suggest that data stored in the same channel are similar and that mixing data between channels tends to destroy valuable information. In our pipeline, we average only inside same channels (APIC) because it yields the best performance. Choosing APBIC would be trading performance for computation speed since its high compression rate enables a much faster training of the prediction algorithm.

### Data based parcellation of the brain using deep video representation

Depending on the considered region of the brain, the best fitting representation varies. We show that the compressed activities of different layers show contrasts between low-level (retinotopic) versus high-level (object-responsive) areas, but also between foveal and peripheral areas.

The difference between the prediction score from high level layer activity and low level layer activity of both streams ($L_4^{flow} - L_2^{flow}$ and $L_4^{rgb} - L_2^{rgb}$) yields a clear contrast between occipital (low-level) and lateral (high-level) areas (see Fig 3). This highlights a gradient of complexity in neural representation along the ventral stream which was also found in (Güçlü & van Gerven, 2015).

The difference between predictions score from low-level layers activity of flow fields stream and high level layers activity of raw frames stream ($L_1^{flow} - L_4^{rgb}$) yields a contrast that does not match boundaries between visual areas; instead, it does coincide with the retinotopic map displaying preferred eccentricity (see Figure 4). Intuitively this means that regions where brain activity is better predicted from the highest layer of optical flow fields than from the lowest layer of raw frames stream are involved in peripheric vision whereas regions where activity is better predicted from the lowest layer of raw frames

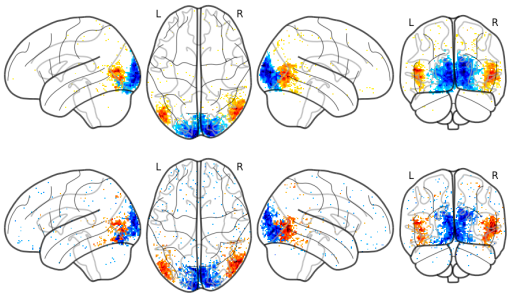


Figure 3: High level and low level areas contrasts: Difference between predictions score from high level layer activity and low level activity of the raw frames stream $L_4^{rgb} - L_2^{rgb}$ (top) and flow fields stream $L_4^{flow} - L_2^{flow}$ (bottom). The results show a clear contrast between occipital areas better predicted by lower level layers (blue) and lateral areas better predicted from highest level layers (red), illustrating a gradient of complexity across areas.

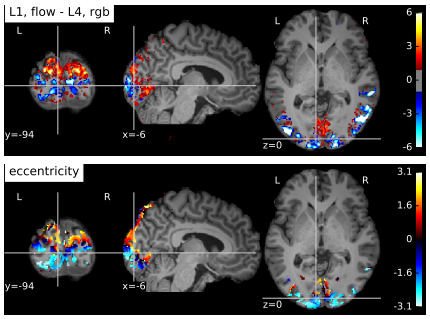



Figure 4: The difference between predictions score from low-level layers activity of flow fields and high-level layers activity of raw frames stream $L_1^{flow} - L_4^{rgb}$ (top) resembles the preferred eccentricity map of the same subject (bottom). Areas that are better predicted from low level flow fields streams are mostly involved in peripheric vision whereas areas better predicted from high level raw frames stream are mainly foveal.

stream than from the highest layer of optical flow fields are mainly foveal.

We use the contrasts between high level layers and low level layers, and the eccentricity related contrast to construct a parcellation of the brain based on these contrasts (see Figure 5). From the 8 possible resulting profiles, three major clusters stand out allowing us to successfully depict a clustering of the voxels using contrasts from deep representation of the stimuli.

## Discussion

Reproducing the results of (Güçlü & van Gerven, 2015) we have shown that lateral areas are best predicted by the last layers of both streams whereas occipital areas are best predicted by first layers of both streams. We have also shown that foveal areas are best predicted by last layers of the raw

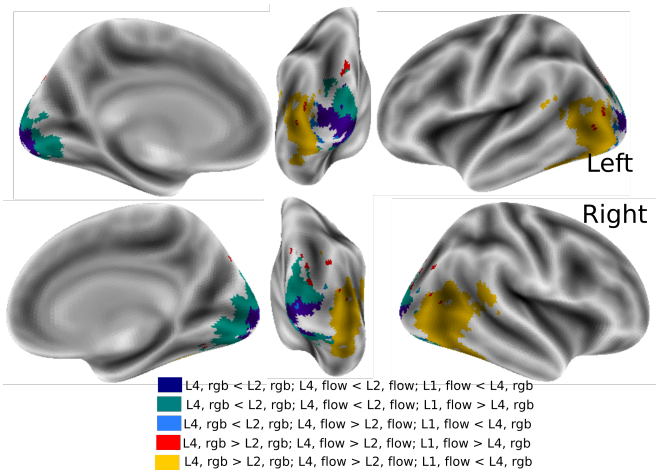



Figure 5: Parcellation summarizing artificial-biological correspondences: The set of active voxels were split into subgroups according to their differential response to three contrasts: $L2^{flow} - L4^{flow}$, $L2^{rgb} - L4^{rgb}$, and $L1^{flow} - L4^{rgb}$. From the 8 possible resulting profiles, 3 major clusters stand out: deep blue, $L2^{flow} > L4^{flow}$, $L2^{rgb} > L4^{rgb}$, and $L1^{flow} < L4^{rgb}$; it corresponds to a voxel set in primary visual areas that has low eccentricity (foveal regions); green, $L2^{flow} > L4^{flow}$, $L2^{rgb} > L4^{rgb}$, and $L1^{flow} > L4^{rgb}$, it corresponds to the same visual areas, but for voxels with higher eccentricity (peripheric voxels); yellow , $L2^{flow} < L4^{flow}$, $L2^{rgb} < L4^{rgb}$, and $L1^{flow} < L4^{rgb}$, it corresponds to lateral and lateral visual areas that encode more abstract representations of the objects.

frames stream and that peripheric areas are best predicted by the first layers of the flow fields stream. We have introduced a compression procedure for video representation that does not alter too much the channel structure of the network, yielding tremendous gains in performance compared to PCA.

The linear prediction from deep video features yields predictions scores that are far better than chance. However the TVL1 algorithm (Zach, Pock, & Bischof, 2007) used in the TSN network does not produce high quality flow fields. Using more recent algorithms to compute optical flow such as Flownet 2 (Ilg et al., 2017), our performance could be further improved. The TSN Network would have to be retrained though.

In contrast to (Güçlü & van Gerven, 2017), the data used to train the network are not the same as the data presented to the subjects. We rely in fact on transfer between computer vision datasets and the visual content used for visual stimulation. This transfer is imperfect: the Berkeley video dataset contains videos of landscapes and animated pictures that are not present in the Kinetic dataset, which introduces some noise.

In conclusion, our study provides key insights that areas have a role linked to their retinotopic representation when performing action recognition. Future studies should focus on finessing this result by using a network tuned for other tasks.

## Acknowledgments

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No. 720270 (HBP SGA1).

## References

Eickenberg, M., Gramfort, A., Varoquaux, G., & Thirion, B. (2017). Seeing it all: Convolutional network layers map the function of the human visual system. *NeuroImage*, *152*, 184–194.

Güçlü, U., & van Gerven, M. A. (2015). Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, *35*(27), 10005–10014.

Güçlü, U., & van Gerven, M. A. (2017). Increasingly complex representations of natural movies across the dorsal stream are shared between subjects. *NeuroImage*, *145*, 329–336.

Ilg, E., Mayer, N., Saikia, T., Keuper, M., Dosovitskiy, A., & Brox, T. (2017). Flownet 2.0: Evolution of optical flow estimation with deep networks. In *Ieee conference on computer vision and pattern recognition (cvpr)* (Vol. 2).

Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., ... others (2017). The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*.

Lin, M., Chen, Q., & Yan, S. (2013). Network in network. *arXiv preprint arXiv:1312.4400*.

Nishimoto, S., Vu, A., Naselaris, T., Benjamini, Y., Yu, B., & Gallant, J. L. (2011). Reconstructing visual experiences from brain activity evoked by natural movies. *Current Biology*, *21*(19), 1641–1646.

Pinho, A., Amadon, A., Ruest, T., Fabre, M., Dohmatob, I., E.and Denghien, Ginisty, C., ... Thirion, B. (2018). *Individual brain charting.* Retrieved 2018, from `https://project.inria.fr/IBC/`

Poline, J.-B., & Brett, M. (2012, Aug). The general linear model and fmri: does love last forever? *Neuroimage*, *62*(2), 871–880.

Simonyan, K., & Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos. In *Advances in neural information processing systems* (pp. 568–576).

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 2818–2826).

Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., & Van Gool, L. (2016). Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision* (pp. 20–36).

Zach, C., Pock, T., & Bischof, H. (2007). A duality based approach for realtime tv-l 1 optical flow. In *Joint pattern recognition symposium* (pp. 214–223).