Master thesis in Artificial Intelligence

Reconstructing naturalistic movies from fMRI brain responses:

Comparing motion-energy features with convolutional neural network representations

Thesis by

L.M. van den Bulk

s4238516

Supervised by

K. Seeliger[1]

and

M.A.J. van Gerven[1]

[1] Donders Institute for Brain, Cognition and Behaviour

Department of Artificial Intelligence

Radboud University Nijmegen

August 2019

Contents

Abstract

In 2011 Nishimoto et al. successfully managed to reconstruct naturalistic movies from the brain activity of the visual cortex. In this study we repeat their experiment with a bigger dataset containing 24 hours of densely sampled fMRI data of a single subject watching a television series, overcoming the problem of having to adjust for a low amount of data in multiple brains. Originally, motion-energy was used to create features for the encoding model, which is a model of how motion is processed in the early visual cortex. We compared the performance of motion-energy features to the performance of an encoding model that uses features from a trained convolutional neural network in order to cover more higher-order areas in the visual cortex. These two types of features were also combined to create an encoding model selective to both lower- and higher-order information. We showed that this combination performs significantly better than the performance of the features separately. However, it must also be concluded that the current method of reconstruction is not sufficient enough to create scenes that have the complexity that a regular television series displays, leaving room for future research in methods to create more detailed reconstructions.

1 Introduction

Making reconstructions from brain activity is currently a hot topic within neuroscience. Being able to model brain activity leads to a better understanding of the dynamic processes that make up the brain and shows us how information is represented along the cortex. Over the past few years, multiple studies have been able to successfully reconstruct brain activity using functional magnetic resonance imaging (fMRI), mainly in the visual system. The visual system is an interesting and useful area to reconstruct as the input of the system is quite clear: the light that enters our eyes forms an image which is projected on the visual cortex in the back of the brain. If one is able to successfully reconstruct the information in the visual system, this would not only lead to a better understanding of how information is represented across the different visual areas in the brain, but the output of the model could also be seen as a brain-reading device, being capable of showing what a person was seeing.

In order to make these reconstructions, an encoding and decoding model are required to make a mapping between the brain activity and the stimulus that was seen. Brain activity is measured as the blood oxygen level-dependent (BOLD) response in separate voxels in the brain using fMRI. Encoding models are used to predict the brain activity in response to the seen stimuli, while decoding models use the brain activity to predict the seen stimuli instead. It is only necessary to train the parameters of the encoding model, as the decoding model can be learned from the encoding model using Bayes' rule (Naselaris et al., 2011), which states that the decoding model is proportional to the product of the encoding model and a prior.

Training the encoding models consist of two parts: First, the stimuli are transformed into features by using a non-linear feature model. This is done non-linearly as most computations performed by the brain are nonlinear

(Naselaris et al., 2011). The second step is to train a response model that linearly transforms features to brain activity. In contrast to the feature model this is done linearly, because the mapping should represent which features stimulate activity in what voxels.

Over the last few years, multiple successful reconstruction models of perceived visual stimuli have been made using the technique described above. Reconstructions have been made of a variety of visual stimuli, for example geometric patterns (Thirion et al., 2006; Miyawaki et al., 2008), handwritten letters (Schoenmakers et al., 2013, 2015) and naturalistic images (Naselaris et al., 2009). In 2011, Nishimoto et al. even successfully reconstructed movie clips (Nishimoto et al., 2011). The research by Nishimoto et al. was the first to show that dynamic brain activity can be decoded using fMRI data. The combination of the ability to capture visual motion using motion-energy features (Adelson and Bergen, 1985; Watson and Ahumada, 1985) and being able to represent the mechanisms of the slow hemodynamic response was key for their good results. In this study we have reproduced the work by Nishimoto et al. on a new and bigger dataset, containing 24 hours of densely sampled fMRI data, and tried to improve upon their results by using an additional feature model: Convolutional neural networks.

Convolutional neural networks (CNNs) are currently the state-of-the-art when it comes to automatic object recognition and are successful in many machine learning applications. Neural networks were originally designed to act as a computational model to simulate how neurons in the brain work (McCulloch and Pitts, 1943) and convolutional neural networks specifically were inspired by the inner workings of the visual cortex (Fukushima, 1980). Just like the receptive fields that become increasingly more complex along the visual system, the earlier layers of a CNN are more sensitive to simple features while the higher layers respond to more abstract, object-like features. Recently, they have steadily increased their popularity in the neuroscience community again. Research has shown that the features created by convolutional neural networks accurately predict responses in both the ventral and dorsal stream in the visual cortex (Güçlü and van Gerven, 2015, 2017). Furthermore, Güçlü et al. (2015) showed that CNNs produce state-of-the-art results when they are used as feature models for encoding and decoding models in the domain of naturalistic images. They improved upon results of earlier research where only low-level features were used, showing that also using high-level features is crucial for optimal performance. Since motion-energy features are solely low-level features, it seems plausible that the features from CNNs can improve the reconstructions when decoding movies instead of images as well. This hypothesis gets substantiated by the the fact that convolutional neural networks are also state-of-the-art in the domain of action recognition models where videos are classified based on their contents (Tran et al., 2018).

To create the CNN features, we used a feedforward CNN that was trained on 250,000 videos to predict 400 classes of different actions. The learned filters of the last convolutional layer of the network were used as the features to train the encoding model to capture the high-level information in the data. To investigate if CNN features perform better than its motion-energy counterpart, we collected both types of features on the same

data set and trained encoding models to predict the BOLD responses collected from the entire brain. A combined model of the best predictive voxels over both feature types was also constructed to see if having a combination of both features would lead to better reconstructions. A Bayesian approach was used to combine a naturalistic movie prior with the encoding models to create reconstructions from unseen BOLD responses.

Using this approach we were able to show that a combined model that includes both features that represent lower-order and features that include higher-order information is significantly better at making reconstructions than models that use only one of these types of features. However, the quality of the reconstructions is still very crude. This is mainly caused by the fact that the stimuli set, a television series, displays very complex scenes and is not easily reconstructed, leaving room for future research on the improvement of more detailed reconstructions.

1.1 Previous work

The research towards building a so called brain-reading device has increased for the last few decades, following the emergence of fMRI. Most work has been performed on the decoding of information from the visual cortex, with the most successful methods using fMRI to measure the activity. Pioneering research was done by Haxby et al. (2001), who investigated the ventral visual pathway. In the study, subjects were presented with pictures of faces, cats, man-made objects and control scrambled objects in an fMRI scanner. It was shown that each category caused a distinct response pattern in the visual cortex, and they were therefore able to predict which category the participants were seeing. Interestingly, the patterns were also predicted correctly when highly specialized areas like the fusiform face area (FFA) or the parahippocampal place area (PPA) were left out from the analysis. This indicated that representations of objects are distributed across the entire visual cortex and that they are unique even in earlier stages of the visual pathway. In 2005, Kamitani and Tong confirmed this finding by showing that they could decode the orientation of a stimulus out of eight possible options by just looking at the the fMRI activity in the primary visual cortex (V1) (Kamitani and Tong, 2005).

A big breakthrough in the field was made by Thirion et al. (2006), as they were able to make an actual reconstruction of the shape of the image that had been presented to their participants. Until that time, all studies about visual decoding had done image classification, where an image would be chosen from a predefined set of options. By showing their participants flickering checkerboard patterns, Thirion et al. (2006) made a retinotopic mapping. A generative model was then created by mapping the retinotopic visual stimuli patterns to the fMRI data. By inverting this model using a Bayesian framework, unseen fMRI data could be used to reconstruct the corresponding pattern on a retinotopic map. They were also able to reproduce their findings on mental imagery, albeit not as well as in the visual task, supporting the hypothesis that imagery activates the early visual cortex (Chen et al., 1998; Kosslyn et al., 1995). Miyawaki et al. (2008) build upon this research by extending to a multi-voxel, multi-scale approach in order to be able to reconstruct luminance in an image. They showed that using multiple voxels was beneficial for the decoding performance, suggesting that visual infor-

mation is not only represented by renotopic mapping, but also by the correlations between voxels. A multi-scale approach was chosen as conventional retinotopy only creates a location-to-location mapping, and it is thought that the visual cortex represents visual information at multiple scales. In contrast to Thirion et al. (2006), they did not invert the encoding model to receive a decoding model, but instead directly computed it. For each patch in the image a decoder was trained that predicted the luminance based on a weighted sum of the fMRI signals. With this approach they were able to reconstruct geometric shapes and letters quite clearly.

Naselaris et al. (2009) took it one step further by trying to reconstruct natural images. They used a Bayesian framework that used two different encoding models to integrate both low-level and high-level information: the structural model and semantic model. Furthermore a natural image prior was used to create preexisting knowledge about the structure of natural images. The structural model defined a mapping between the fMRI data of the most predictive voxels in the visual cortex and natural images filtered by two-dimensional Gabor filters. The semantic encoding model defined a mapping between the fMRI data of the most predictive voxels and the semantic category of natural images as labeled by human observers. The reconstructions are created by choosing an image from the prior that has the highest posterior probability, which was proportional to the multiplication of the likelihoods of the two encoding models. The reconstructions were able to accurately depict the spatial structure and semantic category of the stimuli.

One of the most recent work on reconstructions was done by Schoenmakers et al. (2013, 2015). The ideas from Thirion et al. (2006) and Naselaris et al. (2009) were combined in Schoenmakers et al. (2013) to create an approach that could reconstruct handwritten letters. An encoding model based on image features together with a suitable image prior was used to explicitly invert the encoding model to create the decoding model. The encoding model was constructed as a regularized linear Gaussian model and the image prior used was a multivariate Gaussian. During encoding a mapping was learned between the pixel values of the input image and the fMRI data. High quality reconstructions were made, even of unseen letters, making this a very universal method. Schoenmakers et al. (2015) expanded on this research by adding higher-level semantic information to the model. This was done by adding information from higher-order brain areas and using a Gaussian mixture model as a prior. This lead to even better reconstructions then their previous work.

The most important work for the current study was the work by Nishimoto et al. (2011). They extended their research in Naselaris et al. (2009) from reconstructing natural images to the reconstruction of natural movies. Decoding the brain over time is a difficult problem as using fMRI to measure BOLD responses is relatively slow in comparison with the rate that vision is processed in the brain. However, fMRI is currently the best tool for noninvasive measurement of brain activity. Nishimoto et al. (2011) solved this problem by designing an encoding model that includes a set of hemodynamic response filters spanning various temporal delays to fit each voxel to their distinctive hemodynamic delay. The encoding model also makes use of motion-energy (Adelson and Bergen, 1985; Watson and Ahumada,

1985) as their feature model, which uses three-dimensional Gabor filters at its basis, to make a mapping between the movies and the fMRI data (for a more elaborate explanation of motion-energy see section 2.2.1). Reconstructions were constructed by using a Bayesian approach to combine the encoding model with a natural movie prior that consisted of approximately 18 million seconds of clips sampled from the internet. The top 100 videos with the highest posterior probability were averaged to create the final reconstruction. Their results were the first successful reconstructions of natural movies created from human brain activity.

2 Materials and Methods

2.1 Dataset

This research will use the data collected by Seeliger et al. (2019). An overview of the details described in that study will be presented here. Seeliger et al. (2019) collected 24 hours of fMRI data from a single subject (male, age 27) who watched 31 episodes of the television series *Doctor Who* (BBC). The data was collected especially to have a sufficient amount of free parameters in order to train machine learning models for decoding and encoding analysis. Only a single subject was used to be able to analyze a brain at voxel resolution without having to compensate for the phenotypic diversity between brains by smoothing and normalizing across them. In total, 121.360 volumes of training data and 1.178 volumes of test data were recorded.

In contrast with Nishimoto et al. (2011) who only collected slices covering the posterior occipital cortex and two hours of data, the data by Seeliger et al. (2019) covers the entire brain and was collected over a ten times bigger time frame. This means we have a lot more data to incorporate into our models and see whether performance of both the motion-energy and the deep neural network encoding models can be increased using more areas than the brain than just the posterior occipital cortex.

2.1.1 Experimental procedure

The volumes were recorded over a six month period. In each session one episode of *Doctor Who* was shown. All episodes were split into four clips, where the first three clips were all twelve minutes long and the last clip was the remainder of the episode with a variable length. Two clips were shown per fMRI recording to support the attention of the participant. Additionally seven short clips (ranging from one to four minutes with a total length of approximately 12 minutes) were concatenated and presented each session in an extra fMRI recording to be included in the test set. The test set clips featured mini-episodes from different seasons than the episodes from the train set, featuring different actors but very similar stories and surroundings. The test clips were presented in total 22 times over all sessions and the final test volumes were averaged over all repetitions. The training set was presented just once. All fMRI recordings ended with a black screen lasting 16 seconds to account for the hemodynamic delay. The videos were presented on a mirror in the MRI machine reflecting the projection of a video projector on the outside. The videos were shown at 20 horizontal and vertical degrees of the visual field. To make sure the participant was able to see as much as possible of the original video, the videos were first resized to 696×1264 and then cropped to 696×732 to

fill the screen within the 20 degrees. Black margins were added to the cropped video to fill the screen to the projector resolution of 768×1024. Audio was presented using earphones and the dynamic range of the audio was compressed in order to have the speech not overshadowed by the scanner noise, but have a tolerable volume for the louder sounds. The subject was instructed to fixate on the fixation cross placed in the center of the video. A custom-made foam head cast, a chin rest and constant distances within the set-up were used to ensure stable positions across the sessions. It was always ensured that the participant was comfortable during the sessions.

2.1.2 fMRI parameters

The measurements were taken in a Siemens 3T MAGNETOM Prisma with a 32-channel head coil. The functional scans had a TR of 700 ms, a TE of 39 ms and a flip angle of 75 degrees. Volumes were recorded with a voxel size of 2.4 mm^3 using 64 transversal slices. The videos were measured with a multiband acceleration factor of 8. Next to functional scans, structural scans were carried out in order to localize areas related to the visual and auditory systems. Specifically, localizers were collected for V1, V2, V3, MT, LOC, FFA, OFA, AC and M1. The first three were mapped separately for the dorsal and ventral stream, and all areas were provided for the different hemispheres. These structural scans had a TR of 2300 ms, a TE of 3.03 ms and a flip angle of 8 degrees. These volumes were recorded with a voxel size of 1 mm^3 using 192 saggital slices.

2.1.3 Data preprocessing

After the data was recorded, preprocessing was applied. The preprocessing only consisted of realignment and standardization. No slice time correction was necessary due to the fast multiband protocol. The series of volumes of both the train and test set were realigned per twelve minutes to their middle volume and then all volumes were realigned to the middle volume from the first twelve minutes of the first episode to ensure alignment across all clips. Each voxel was standardized with zero mean and unit variance.

2.2 Feature models

Two feature models were used in this research to train the encoding models and learn the mapping to the BOLD responses of the train set. The original research by Nishimoto et al. (2011) used motion-energy as their feature model and we expanded on their research by including convolutional neural networks as a feature model. Both models will be explained below. Code for the feature models and the rest of this research can be found on `https://github.com/LeoniekevandenBulk/Thesis`.

2.2.1 Motion-energy

The concept of motion-energy was originally proposed as a model of the human perception of motion (Adelson and Bergen, 1985; Watson and Ahumada, 1985). Motion-energy is designed to be selective to different spatiotemporal frequencies and is considered to be similar to the processing that occurs in the early visual pathway. The output of a motion-energy model gives information about the direction of motion at a given location in a given moment in time. Nishimoto et al. (2011) based their computations for the motion-energy model on the work by Adelson and Bergen (1985). The motion-energy features are created in several steps. The first step is to resize the input

videos to 96×96 and discard the color information by transforming the RGB values of the input to CIELAB color space. Color information is removed as it does not improve predictions, making it computationally better to remove it (Nishimoto et al., 2011). The videos are then passed through three-dimensional spatiotemporal Gabor filters. Gabor filters are filters that are sensitive to different positions, orientations and frequencies (e.g. a vertical bar moving to the right). The filters are made by multiplying a three-dimensional sinusoid with a three-dimensional Gaussian envelope. Two of the three dimensions represent space, the last dimensions represent time. The research by Nishimoto et al. (2011) used 6555 separate Gabor filters, containing filters with a range of eight directions (0, 45, 90, 135, 180, 225, 270 and 315 degrees), six spatial frequencies (0, 2, 4, 8, 16 and 32 cycles per frame) and three temporal frequencies (0, 2 and 4 Hz). Filters at each spatial frequency are positioned such that they are separated by 3.5 standard deviations. Each of the 6555 filters is created at two quadratic phases (0 and 90 degrees), causing one to have cosine phase and the other to have sine phase. The output of the two phases is squared and summed, taking advantage of the fact that $sin^2(\theta) + cos^2(\theta) = 1$. This is done to ensure that the resulting filter is contrast invariant and phase insensitive and is thus not influenced by the alignment of the movement to the receptive field of the filter. The summed filters are compressed using a logarithm and downsampled to the sampling rate of the measured BOLD signals by averaging over the TR. Each downsampled signal is normalized to have a mean of zero and a standard deviation of one by normalizing across time using a Z-score transformation. Outliers of more than three times the standard deviation are truncated to improve stability. Note that the motion-energy features of the test set and movie prior made use of the mean and standard deviation of the motion-energy features of the train set to be normalized. Nishimoto et al. (2011) made their implementation of their framework available online at `https://github.com/gallantlab/motion_energy_matlab`, which we used to obtain the motion-energy features for our dataset.

2.2.2 Convolutional neural network

We propose to use the features created by convolutional neural networks next to the motion-energy features to include more higher-order information. Zeiler and Fergus (2014) have shown that the layers of a neural network respond to increasingly complex features. The first couple of layers resemble Gabor-like filters, but in the deeper layers the features start to resemble (parts of) object shapes. We therefore believe the neural network features to be a better fit for the higher-order areas of the visual cortex than the motion-energy features.

We used a convolutional neural network designed for the classification of videos. These so called action recognition networks are state-of-the-art when it comes to identifying and detecting activities in video. One of the best models of the past year is the R(2+1)D network (Tran et al., 2018), which is a variation on the successful ResNet architecture (He et al., 2016). The network uses spatiotemporal convolutions that are in-between 2D and 3D, called (2+1)D convolutions. It splits a full 3D convolution in two successive operations: a spatial 2D convolution followed by a temporal 1D convolution. This decomposition doubles the

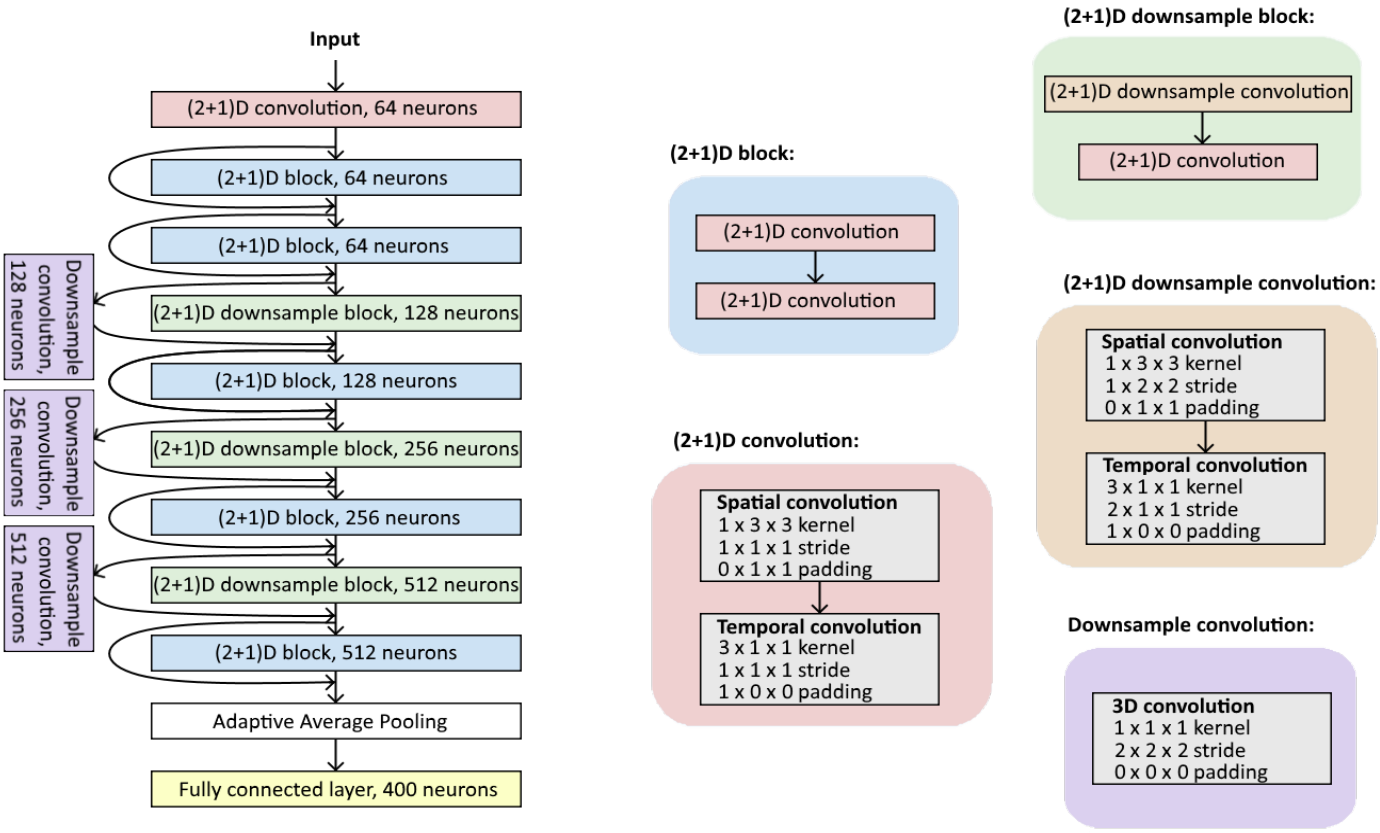

Figure 1: *The architecture of the R(2+1)D network. The different elements of the architecture are color-coded, their specifics can be found in their corresponding colored boxes on the right. The curved arrows denote the skip-connections that carry over the input of a block to be summed with the output of that block. Note that in the architecture only the number of neurons in the temporal layers are displayed, the number of neurons in the spatial layers can be computed via equation 1.*

amount of nonlinearities in comparison with a 3D convolution for the same number of parameters, making the network capable of learning more complex representations. The decomposition also makes the network easier to optimize, yielding a better performance.

The network used for the current study is the 18-layer R(2+1)D network from Tran et al. (2018) and is depicted in Figure 1. We will give an overview of the architecture here, the specific details, however, can be found in the original paper. The network is equal to an 18-layer 3D ResNet where the 3D convolutions have been replaced with the decomposed convolutions as described above. The network starts out with a (2+1)D convolution, which is followed by eight (2+1)D blocks, an average pooling layer and ends with a fully convolutional layer which maps to the final labels by means of a softmax function. A (2+1)D block is defined as two sequential (2+1)D convolutions that get summed with the initial input of that block via a skip connection. Tran et al. (2018) designed the (2+1)D convolutions such that they have the same amount of parameters as a full 3D convolution, which results in that the temporal 1D convolutions have the same number of neurons as there respective 3D counterparts, but the number of neurons in the spatial 2D convolution are determined

via the following formula:

$$N_s = \frac{k_s^2 k_t N_{t-1} N_t}{k_s^2 N_{t-1} + k_t N_t} \quad (1)$$

where k_s and k_t are the kernel sizes for the spatial 2D and temporal 1D convolutions respectively, and N_{t-1} and N_t are the the number of neurons in the previous and current temporal convolution respectively.

The number of neurons for the temporal convolutions is 64 for the first separate (2+1)D convolution and 64, 64, 128, 128, 256, 256, 512, 512 for the (2+1)D blocks. The spatial kernel sizes are set to 1×3×3 and the temporal kernel sizes are 3×1×1. For each convolution, the input is padded with 1 in the dimensions of the convolution. The stride is set to 1×1×1 to keep the dimensions of the input the same, except in the first spatial and temporal convolution of blocks 3, 5 and 7. There the stride is set to 1×2×2 and 2×1×1 respectively to half the dimension of the input. In these blocks the input that gets summed via the skip-connections is also downsampled with the use of a normal 3D convolution with a kernel size of 1×1×1, a stride of 2×2×2 and a padding of 0×0×0 to avoid summing two volumes with different dimensions. Batch normalization and a ReLu activation function are always applied after each convolution.

We used the pretrained network from Tran et al. (2018), which can be downloaded from `https://github.com/facebookresearch/VMZ/blob/master/tutorials/models.md` and is implemented in *PyTorch* (Paszke et al., 2017). The network was trained on the Kinetics dataset (Kay et al., 2017), which consists of videos in which human actions are performed (e.g. playing certain sports, shaking hands or laughing). The dataset contains 240,000 ten second long training videos at 15 FPS divided over 400 classes. The network expects an input size of 112 pixels × 112 pixels × 16 frames. Training videos were scaled to a size of 128 × 171 pixels and input was created by randomly cropping windows of size 112 × 112 pixels from 16 random consecutive frames and mirrored horizontally with a 0.5 probability. By doing this spatial and temporal jittering, the training size was artificially enlarged meaning that training could continue for longer before hitting a plateau in the loss. Training was done over 45 epochs with an epoch size of 1,000,000 clips. During training, an initial learning rate of 0.01, a momentum of 0.9, a weight decay of 0.0004 and a mini-batch size of 32 were used. Every 10 epochs the learning rate was divided by 10.

Although the Kinetics dataset uses a frame rate of 15 FPS, we should use an FPS that fits the speed of the BOLD response recordings with a TR of 0.7 seconds. In order to create 16 frames over 0.7 seconds, we thus needed an FPS of 22.86 (note that this FPS was also used for the motion-energy features to be consistent across features). We decided to finetune the network to train with the new FPS and also to train the network to be sensitive to the full input instead of cropped windows, as this is more useful when trying to reconstruct entire scenes. Training parameters were kept mostly the same, only the videos were directly resized to 112 × 112 pixels, omitting the spatial jittering. Furthermore, the finetuned network was trained for 30 epochs with an epoch size of 240,000 clips and a mini-batch size of 10. Both training and finetuning was done with stochastic gradient descent using cross-entropy loss.

The final features from our R(2+1)D model

are created by taking the output from the average pooling after the last (2+1)D convolution, which is a feature vector of size 1×512. This layer was taken as the most informative higher order information, as fully connected layers are proven not to be very useful in the encoding of BOLD responses from movies (Güçlü and van Gerven, 2017).

2.3 Encoding model

Encoding models were trained for both feature models between the features created with the training data and their associated BOLD responses. The data was fit per voxel. Let $\mathbf{x^t}$ be a 0.7 second video from the training data consisting of pixels at timepoint t and $\mathbf{y_i^{t+d}}$ the BOLD response of the ith voxel at timepoint t plus a hemodynamic delay of d. Furthermore, let $\mathbf{f}(\mathbf{x^t})$ be the feature representation of the video input at timepoint t. Ridge regression, which is a form of regularized linear regression, was used to fit the individual voxels, giving us the following equation:

$$y_i^{t+d} = \beta_j^\top f(x^t) + \epsilon_i \tag{2}$$

where β_j represents a regression coefficient and ϵ_i is residual noise equal to $\mathcal{N}(0, \sigma_i^2)$. The regression coefficients can be determined by estimating:

$$\hat{\beta}_i = (F^\top F + \lambda I)^{-1} F^\top Y \tag{3}$$

where $F = (f(x^1), f(x^2), ..., f(x^t))$, $Y = (y_i^1, y_i^2, ..., y_i^t)$, I is the identity matrix and λ is a regularization parameter. Leave-one-out cross-validation was used to optimize the regularization parameters. The code that was used for the regression can be found at `https://github.com/alexhuth/ridge`.

To account for the hemodynamic delay, five encoding models were trained for each feature model. Each of the encoding models corresponded to a different delay, covering the feature responses from 2.8 - 5.6 seconds before the BOLD response, with a 0.7 second interval. This in contrast to the original work by Nishimoto et al. (2011), which used a fixed delay of 4 seconds for their reconstructions. For each voxel, it was determined on a separate validation set which delay lead to the most accurate predictions of the BOLD responses. Accuracy was based on Pearson's correlation coefficient between the observed and predicted BOLD responses. From the set of voxels and their optimal delays, a selection was made to only include voxels with a high prediction accuracy as adding voxels with low predictive power leads to a decrease in performance (Kay et al., 2008). Similar to Nishimoto et al. (2011), we collected the top 2000 voxels for both the motion-energy and the R(2+1)D pooling layer encoding model from a total of 117,010 collected voxels. A combined voxel selection was also created in order to be able to capture both the voxels sensitive to low-level features and the voxels sensitive to high-level features. This was done by extracting the voxels (and their respective weights) that were unique across both voxel selections (i.e. can only be found in one of the voxel selections) and adding these to the combined selection. For all voxels that were shared between the two voxel selections, the voxel and weights from the feature model that has the best predictive power is added to the combined selection as well.

2.4 Decoding model

A Bayesian approach was used to reconstruct the test set from the corresponding BOLD

responses. Although it is possible to train a separate decoding model that maps back from the BOLD response to the respective feature, common practice is to derive the decoding model from the learned encoding model (Naselaris et al., 2011). This can be achieved through Bayes' rule, which states that the decoding model is proportional to the encoding model multiplied with a prior. The prior that was used consisted of ∼5 million natural video clips sampled from the internet via two different datasets: the Youtube 8M dataset (Abu-El-Haija et al., 2016) and the Moments in Time dataset (Monfort et al., 2019). The Youtube 8M dataset is a big video dataset created to train machine learning models on, consisting of 6.1 million videos divided over 3862 classes extracted from Youtube based on an automatic quality assessment. The Moments in Time dataset was created as an action recognition set consisting of a million videos labeled with 339 different classes. We randomly sampled 4453362 and 765000 0.7 second videos from the videos in those datasets respectively. All prior clips were assigned a uniform probability. Note that the original work by Nishimoto et al. (2011) used a prior set of ∼18 million clips, however, because of time constraints, we had to use a more limited set.

Each of the prior videos was transformed by both feature models into a set of features. Using the trained encoding model weights from the voxel selections, the BOLD responses for the top 2000 selected voxels were predicted for both feature sets of the prior videos. These were compared to the observed BOLD responses for the test set by taking the Pearson's correlation coefficient between the two. Just like in Nishimoto et al. (2011), we took the top 100 prior videos that had the highest likelihood per TR, based on the correlation coefficient. Note that originally instead of the correlation coefficient, the likelihood estimation was based on a multivariate Gaussian. However, since our training data was only collected once, it was not possible to estimate the covariance matrix, so Pearson's correlation coefficient was taken instead.

The top 100 best prior videos per TR were averaged to create the final reconstruction. Each clip was normalized to have unit standard deviation before averaging to make sure each clip had an equal contribution to the average. After averaging, the mean and standard deviation of the reconstruction were normalized to be equal to the average mean and standard deviation of the top 100 videos.

2.5 Analyses

In order to evaluate the quality of the reconstructions, a method is necessary that quantifies the similarity between the test set videos and their reconstructions. Research surrounding video similarity is mostly focused on content similarity such that videos with the same context can be easily grouped together. However, since we are investigating whether a video can be reconstructed pixel-by-pixel, we chose to resort to a frame-by-frame comparison using an image similarity measure. Recently, the most successful methods for image similarity have been using features from convolutional neural network to compare images (Wang et al., 2014; Jing et al., 2015). For our comparisons, we will use AlexNet (Krizhevsky et al., 2012). AlexNet is a eight-layer convolutional neural network consisting of five convolutional layers (with 64, 192, 384, 252 and

252 neurons respectively) and three fully connected layers (with 4096, 4096 and 1000 neurons). AlexNet was one of the first successful artificial neural networks used for image classification and the ImageNet challenge and has been shown to have a high accuracy and display clear hierarchical object features across the layers (Zeiler and Fergus, 2014). The network maps to a 1000 object categories, containg both animate and inanimate objects. We used the pretrained network from the *PyTorch* platform (Paszke et al., 2017), which can be found on `https://pytorch.org/docs/stable/torchvision/models.html`.

The test video and all three reconstructions were fed frame-by-frame through AlexNet. For each of the layers of the network the features from each of the reconstructions were compared with the features from the test video by using Pearson's correlation coefficient. In order to have a baseline for the correlation coefficients, we also created a random reconstruction by selecting random top 100 videos for each TR. To test the significance of the reconstructions, we compared the correlation coefficients of each of the reconstructions with the random reconstruction by means of a Wilcoxon rank-sum test for each of the layers in the network. To test whether there was any significant difference between the reconstructions themselves, a Wilcoxon signed-rank test was used between the correlation coefficents for each of the layers of the network.

2.6 Imagery

We can take the decoding of the visual cortex one step further by trying to decode mental imagery from the BOLD responses, as previously successfully tried by Thirion et al. (2006). Research has shown that mental imagery engages the visual cortex just like visual information, albeit in a weaker manner (Chen et al., 1998; Kosslyn et al., 1995). We will apply our learned encoding models to the BOLD responses of a mental imagery task, to see if this can be reconstructed as well. The data used was collected by Seeliger et al. (2019), just as the rest of the data used for this research. The same participant that collected the earlier described train and test data was used for the mental imagery task and it was collected with the same protocol and preprocessing. The participant was asked to imagine the two different intro sequences of the *Doctor Who* series appearing in the test set for 10 seconds, in which a blue telephone box flies either through a blue or orange colored space. Both these conditions were repeated 24 times. The final BOLD responses were created by taking the average over these repetitions. As it is unclear what the hemodynamic delay is between the onset of imagining and the BOLD response, data was only collected during the 10 seconds of imagery and the signals were not shifted forward in time. To make the imagery reconstructions, the same encoding and decoding model as for the test set reconstructions were applied. This resulted in a top 100 of prior videos for each of the 14 TRs, which were averaged to create the final reconstruction.

3 Results

3.1 Voxel selection

Five encoding models were trained for each feature model, where each of the encoding models represented a delay in a range of 2.8-5.6 seconds between the stimuli onset and the BOLD response using 0.7 second intervals. Across all delays, the 2000 most predic-

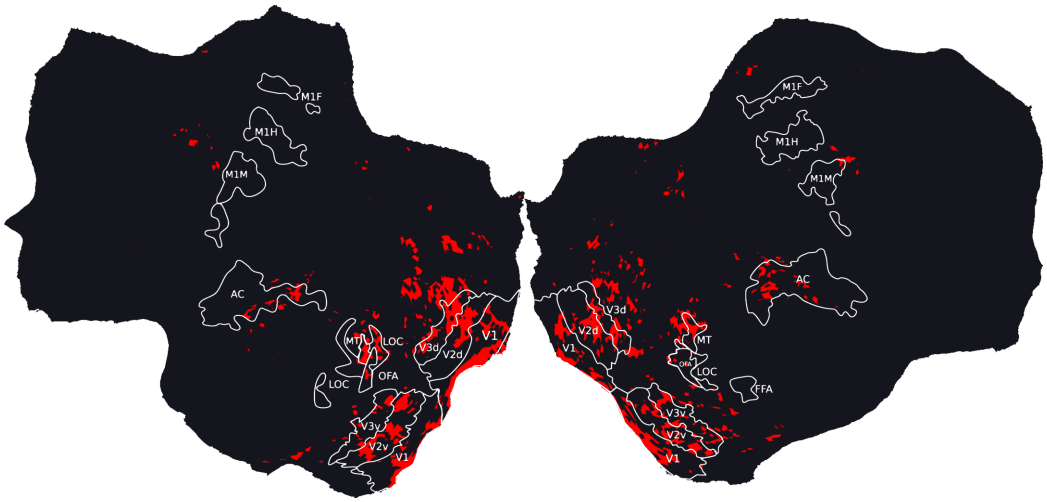

Figure 2: *Top 2000 voxel selection from the motion-energy encoding model presented on a flat-map of the brain.*

tive voxels were chosen for the motion-energy encoding models and the R(2+1)D pooling layer encoding models. Figures 2 and 4 display the voxel selections for motion-energy and the R(2+1)D pooling layer respectively across flat-maps of the brain. All flat-maps presented in this paper were created using *Pycortex* software (Gao et al., 2015). Figures 3 and 5 display Pearson's correlation coefficients for those selected top 2000 voxels on the validation set that was used for the selection procedure. These figures show that there is a distinct difference between the locations of the voxels selected through the two feature models. The motion-energy features are mainly localized in brain areas V1, V2 and V3, while the pooling layer features from the R(2+1)D network are more localized in the higher-order layers of the visual areas and can even be found strongly in the auditory cortex. However, both voxel selections occur in both the lower- and higher

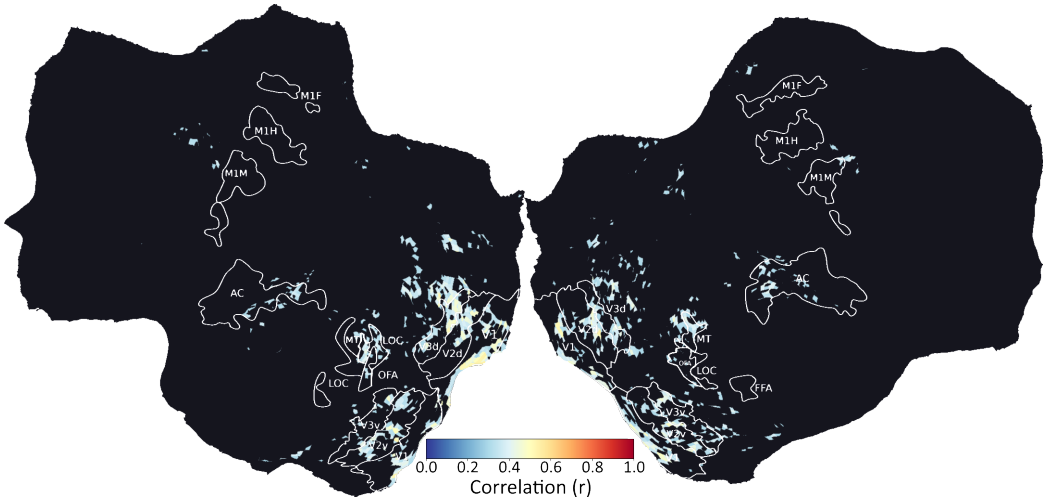

Figure 3: *The validation correlation coefficients for the voxel selection of the motion-energy encoding model presented on a flat-map of the brain.*

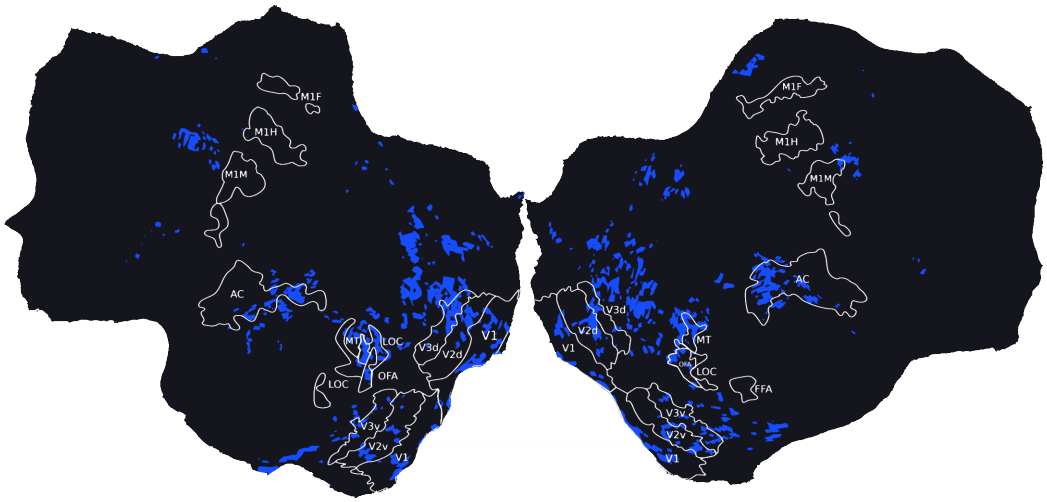

Figure 4: *Top 2000 voxel selection from the R(2+1)D pooling layer encoding model presented on a flat-map of the brain.*

order visual areas and are not exclusively found in one region. The correlation coefficients do show that the motion-energy voxel selection distinctly has the best validation performance in the early visual cortex, while the pooling layer voxel selection seems to have a preference for the higher-order visual areas, but only slightly. Overall the correlations are significantly higher for the motion-energy voxel selection with an average of 0.36 and a standard deviation of 0.052 in contrast to a average of 0.28 and a standard deviation of 0.039 for the R(2+1)D pooling layer voxel selection.

These two voxel selections were merged to create the combined voxel selection. There are 1300 voxels that were shared across the motion-energy voxel selection and the R(2+1)D pooling layer voxel selection, of which 1171 voxels had the best predictive power in the motion-energy voxel selection

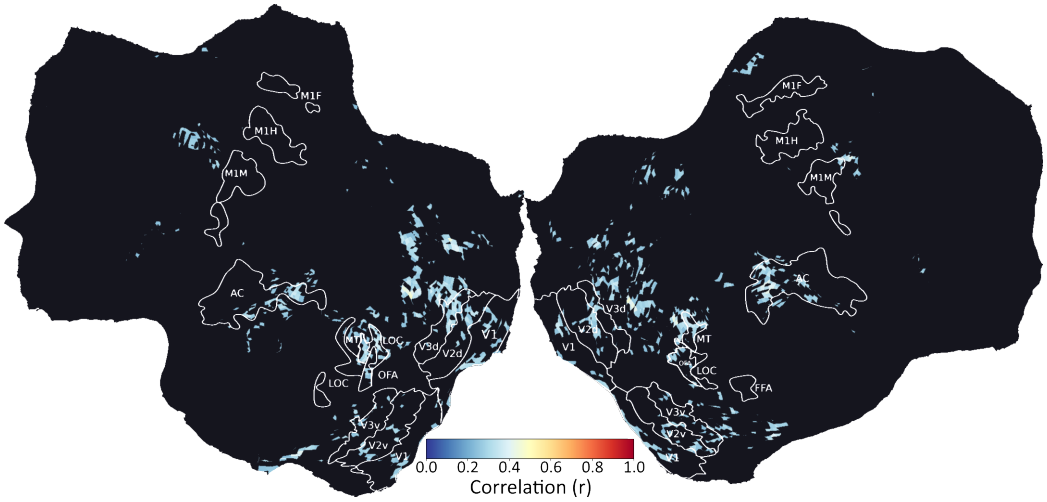

Figure 5: *The validation correlation coefficients for the voxel selection of the R(2+1)D pooling layer encoding model presented on a flat-map of the brain.*

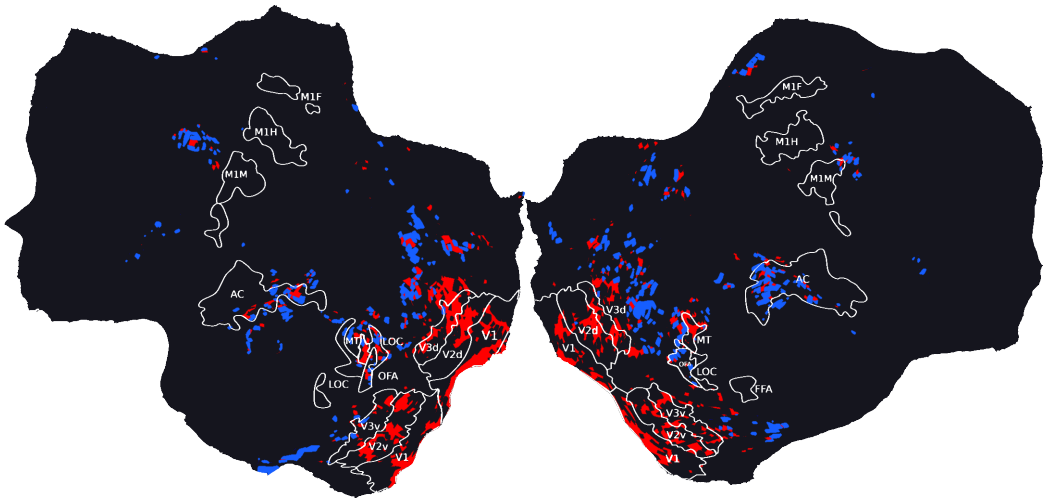

Figure 6: *The resulting voxel selection by combining the voxel selections from the motion-energy and the R(2+1)D pooling layer presented on a flat-map of the brain. It includes 2700 voxels, the blue colored ones were taken from the pooling layer selection and the red colored ones ones were taken from the motion-energy selection.*

and 129 voxel were picked from the R(2+1)D pooling layer selection. The remaining voxels consist of 700 voxels uniquely occurring in the motion-energy selection and 700 voxels uniquely occurring in the R(2+1)D pooling layer selection, giving a total of 2700 voxels for the combined voxel selection. The final voxel selection for the combined model can be found in Figure 6 and its respective Pearson's correlation coefficients on the validation set can be found in Figure 7. Figure 6 shows that the voxels in visual areas V1, V2 and V3 almost entirely come from the motion-energy selection, while the higher-order visual areas are predominantly voxels from the R(2+1)D pooling layer selection. Area MT, however, also seems to be heavily influenced by the motion-energy selection.

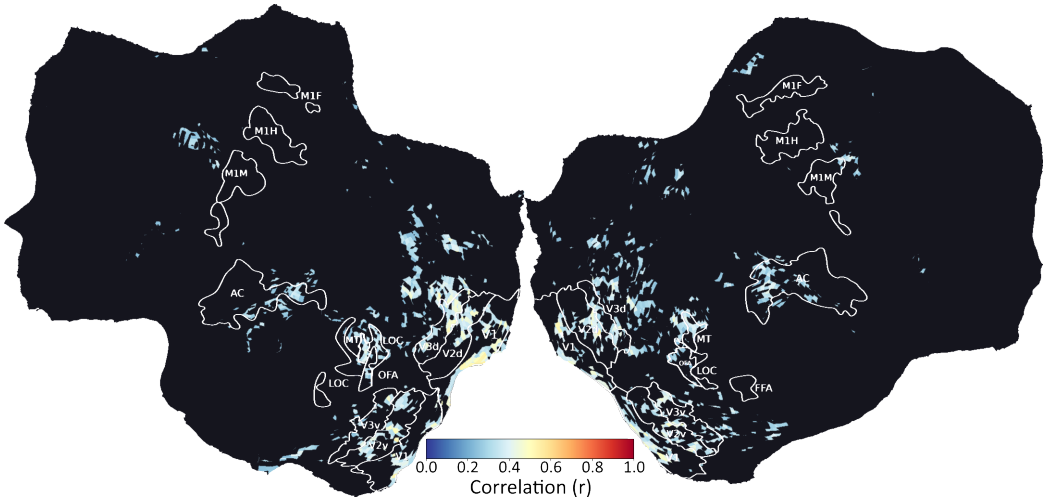

Figure 7: *The validation correlation coefficients for the combined voxel selection presented on a flat-map of the brain.*

Delay	Motion-energy	R(2+1)D Pooling	Combined
2.8 seconds	212/2000 voxels	80/2000 voxels	247/2700 voxels
3.5 seconds	454/2000 voxels	331/2000 voxels	560/2700 voxels
4.2 seconds	608/2000 voxels	675/2000 voxels	842/2700 voxels
4.9 seconds	246/2000 voxels	462/2000 voxels	445/2700 voxels
5.6 seconds	480/2000 voxels	452/2000 voxels	606/2700 voxels

Table 1: *The distribution over the selected delays for the three different voxel selections.*

The correlation coefficients of the combined selection have an average of 0.33 and a standard deviation of 0.062.

Table 1 shows the distribution over the selected delays across the three voxel selections. A delay of 4.2 seconds was selected most across all three selections with percentages of 30.4%, 33.8% and 31.2% for the motion-energy selection, the R(2+1)D pooling layer selection and the combined selection. A delay of 2.8 seconds is chosen least, with percentages of 10.6%, 4.0% and 9.1% respectively. The average delay for the three models comes down to 4.3 seconds for the motion-energy selection, 4.5 seconds for the (2+1)D pooling layer selection and 4.4 seconds for the combined selection. Figures 8, 9 and 10 show the distribution over the delays on an individual voxel level for the motion-energy, R(2+1)D pooling layer and the combined voxel selections respectively. In Figure 8, the motion-energy selection shows a systematic change in the delay selection from the lower-order visual areas with smaller delays to the higher-order visual areas with bigger delays. In contrast, the R(2+1)D pooling layer selection in Figure 9 does not show such a change and displays mainly bigger delays in every area of the visual pathway. Interestingly however, there can be found smaller delays in the

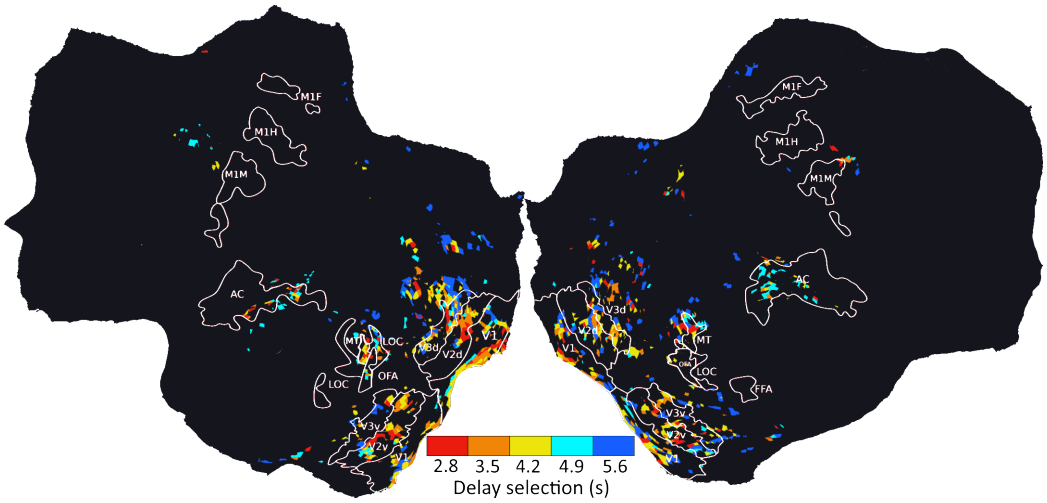

Figure 8: *The distribution over delays for the voxel selection of the motion-energy encoding model presented on a flat-map of the brain.*

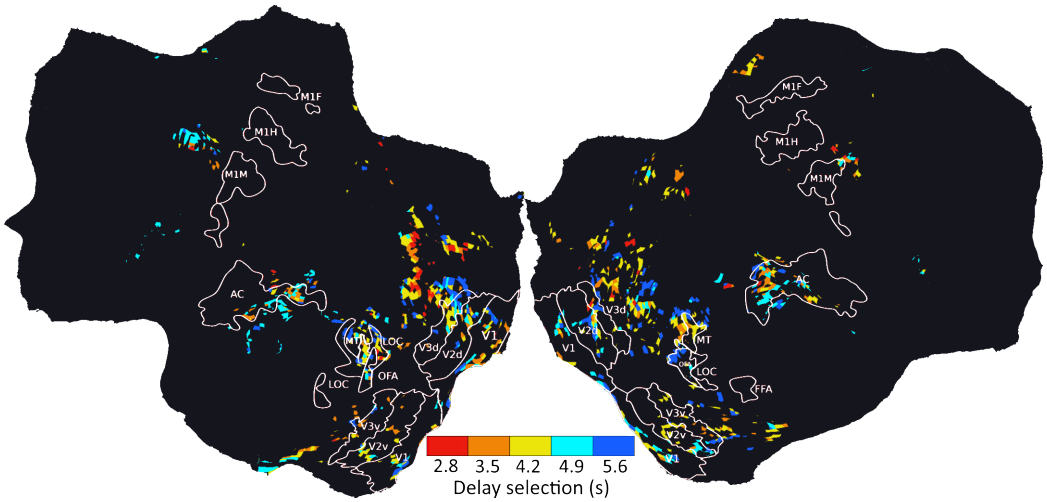

Figure 9: *The distribution over delays for the voxel selection of the R(2+1)D pooling layer encoding model presented on a flat-map of the brain.*

elongated area to the top-right of the LOC in the right hemisphere, which seems to be the intraparietal sulcus (IPS), although we do not have a localizer that identifies it. The combined voxel selection in Figure 10 shows the change in delay selection from the lower-order visual areas to the higher-order visual areas quite well again, with the exception of the IPS region, as the lower-order areas are mainly composed of voxels from the motion-energy selection.

3.2 Reconstructions

The three voxel selections described above were used to create the reconstructions of the test set by predicting the respective BOLD responses for all prior videos and selecting the top 100 videos per TR. The top 100 videos were averaged to create the final reconstructions. Figures 11 - 16 shows examples of the created reconstructions. The figures show different scenes from the test set, spanning four TR's, together with the recon-

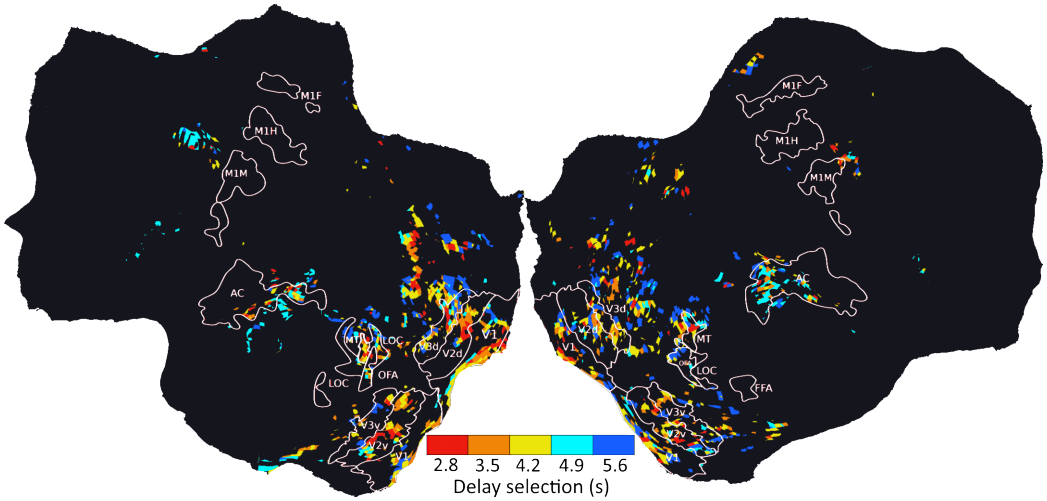

Figure 10: *The distribution over delays for the combined voxel selection presented on a flat-map of the brain.*

structions and the first five videos from the top 100 that formed these reconstruction. The scenes are the same across all models to make comparison easier. The full reconstruction videos can be viewed on `https://github.com/LeoniekevandenBulk/Thesis`.

Figures 11, 13 and 15 show reconstructions of three scenes from the test set that were successfully reconstructed for the motion-energy, R(2+1)D pooling layer and combined model respectively. We define this as successful reconstructions as we can recognize the scenes in the reconstructions. The first scene is depicted in the first big column of these figures and shows the intro sequence from the episodes used in the test set, the second column shows the second scene depicting one of the main characters talking and the last column shows a scene where two characters are sitting across a table from each other. Scene one gets reconstructed quite well by the motion-energy and combined model, we can clearly see centralized white text against a darker background. In the R(2+1)D pooling layer model, the text is a lot less present, but becomes more of a centralized spotlight. However, it seems to focus quite a lot on the color blue with its prior videos, which could be linking to the background color. Scene two gets reconstructed as a human form by all three models, albeit less sharply by the R(2+1)D pooling layer model. Interesting to see is that although the character in the scene is not upright, all its reconstructions are. Furthermore, even though a different character is shown in the fourth frame, the reconstructions do not seem to change a lot from the previous reconstruction frames. The reconstructions for the third scene are quite clearly two human shapes next to each other for all three models, but it is again hard to pick up the scene transition to a different perspective from the reconstructions.

Figures 12, 14, 16 show unsuccessful reconstructions from four different scenes from the test set for the motion-energy, R(2+1)D pooling layer and combined model respectively. The reconstructions are deemed unsuccessful as they do not resemble the scenes sufficiently. Note that the majority of the reconstructed scenes from the test set was unsuccessful. Each of these scenes was selected to show the limitations of the current reconstruction method. Column one depicts a scene with one of the main characters in the foreground and multiple soldiers in the background. The top five prior videos in all models show several videos with groups of people, with some even showing a person more clearly in the foreground. However, because the locations of these people do not align with each other, the reconstructions become very noisy and not recognizable as people. Scene two depicts a humanoid alien who gets reconstructed as an elongated lighter blob. Even though the character looks quite similar to a human, the prior videos display almost no human faces. This in contrast to the second column of the successful reconstructions which shows quite a similar perspective and shows almost solely human faces in the prior videos. The prior videos instead seem to focus on the elongated shape of the alien. The third column shows the three main characters inside a space ship, which contains a lot of futuristic details. The reconstructions are again mostly noise. The prior videos from the models seem to focus on quite different videos. The motion-energy model seems to depict mostly water or grassy terrains, the R(2+1)D pooling layer model seems to focus more on big groups of people and the combined model shows mostly big machin-

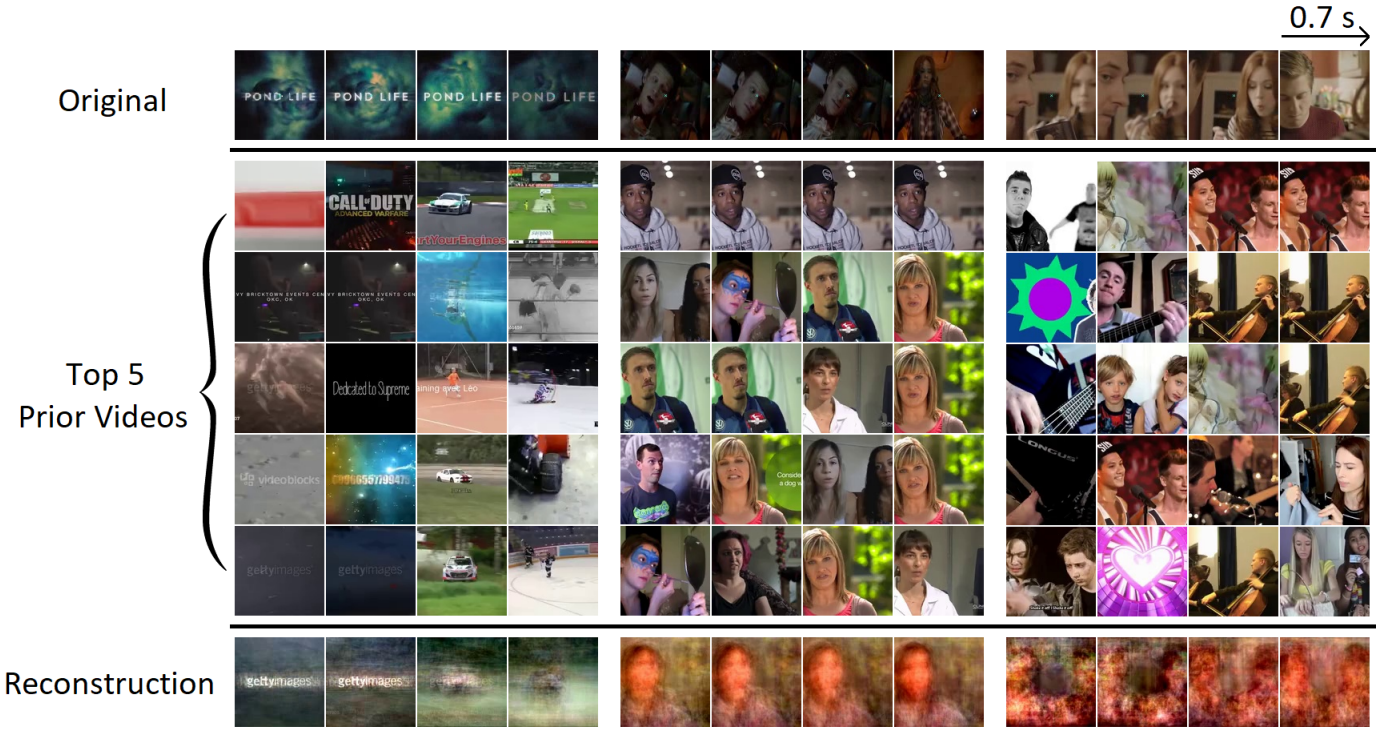

Figure 11: *Examples of successful reconstructions of three different scenes by the motion-energy model. The second through sixth rows display the first five videos from the top 100 prior videos that create the reconstruction. Each frame within a scene is separated by 0.7 seconds.*

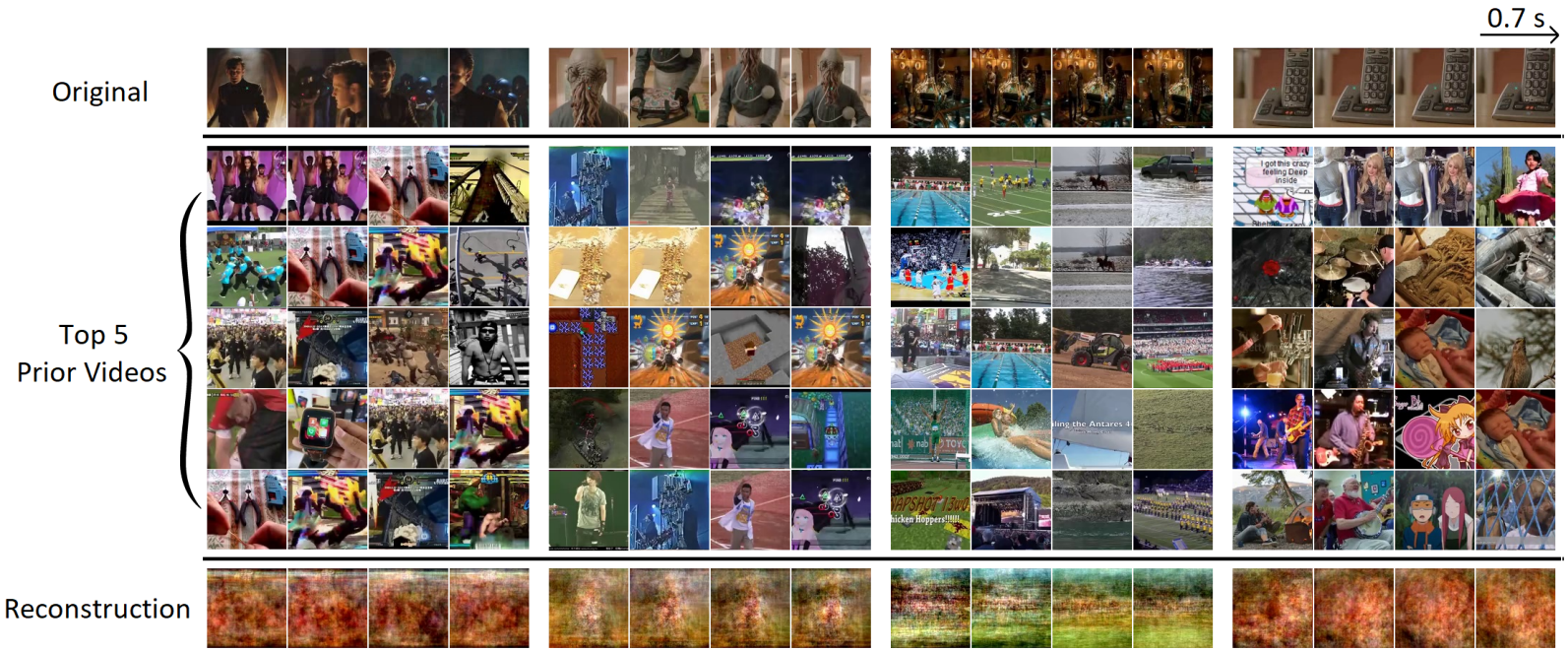

Figure 12: *Examples of unsuccessful reconstructions of four different scenes by the motion-energy model. The second through sixth rows display the first five videos from the top 100 prior videos that create the reconstruction. Each frame within a scene is separated by 0.7 seconds.*

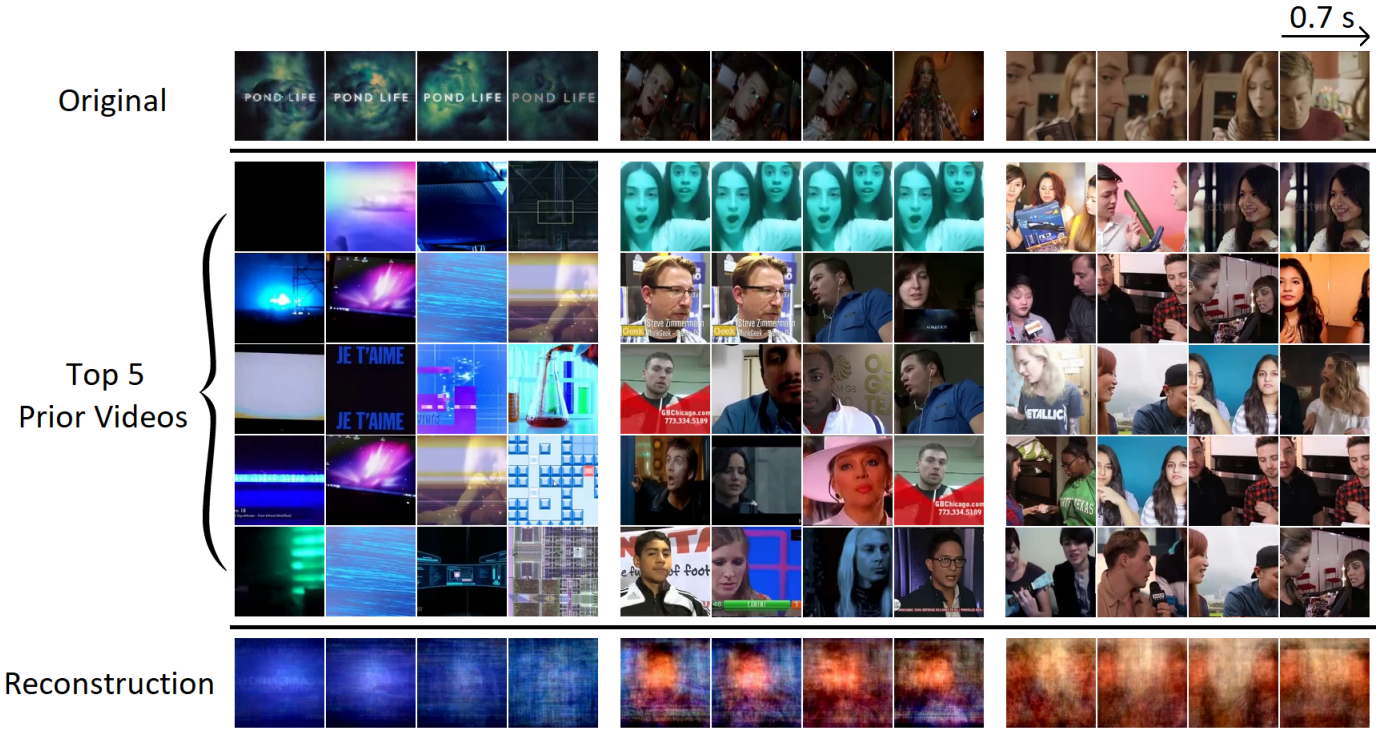

Figure 13: *Examples of successful reconstructions of three different scenes by the R(2+1)D pooling layer model. The second through sixth rows display the first five videos from the top 100 prior videos that create the reconstruction. Each frame within a scene is separated by 0.7 seconds.*

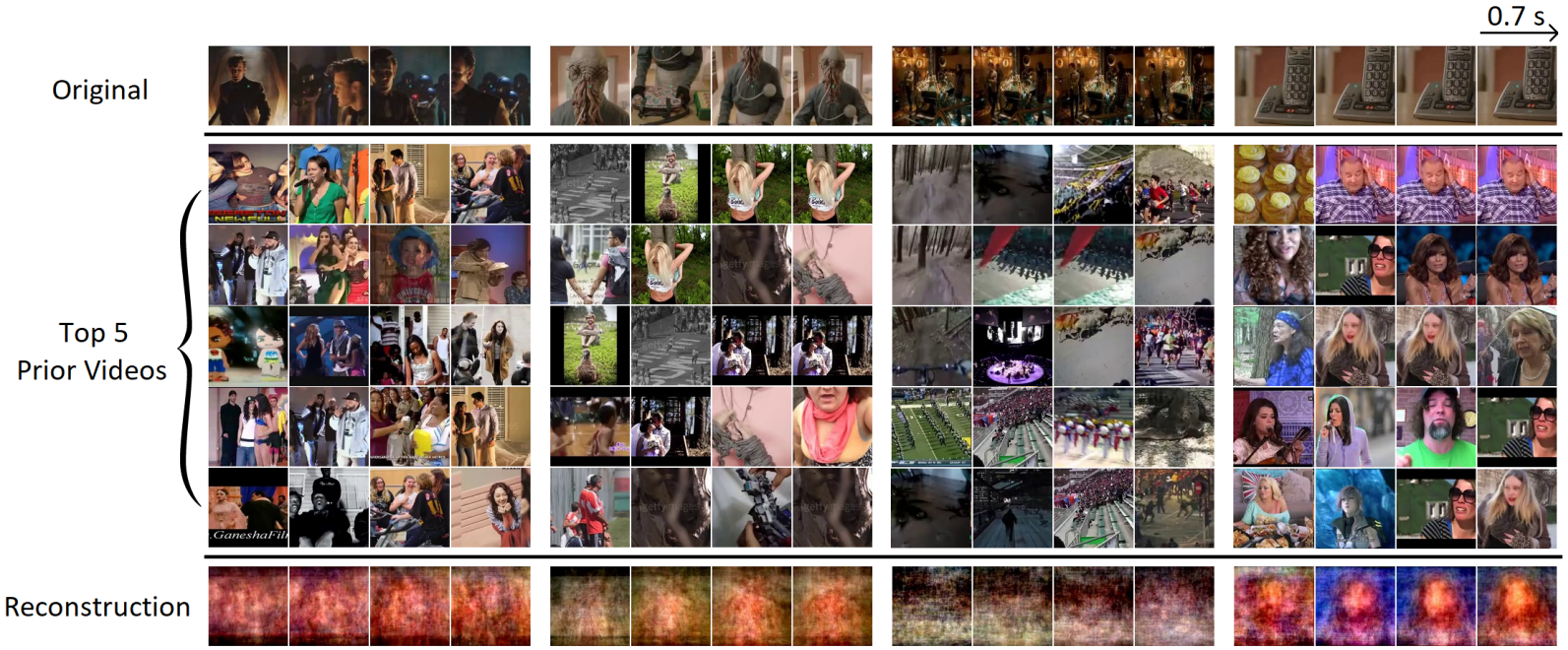

Figure 14: *Examples of unsuccessful reconstructions of four different scenes by the R(2+1)D pooling layer model. The second through sixth rows display the first five videos from the top 100 prior videos that create the reconstruction. Each frame within a scene is separated by 0.7 seconds.*

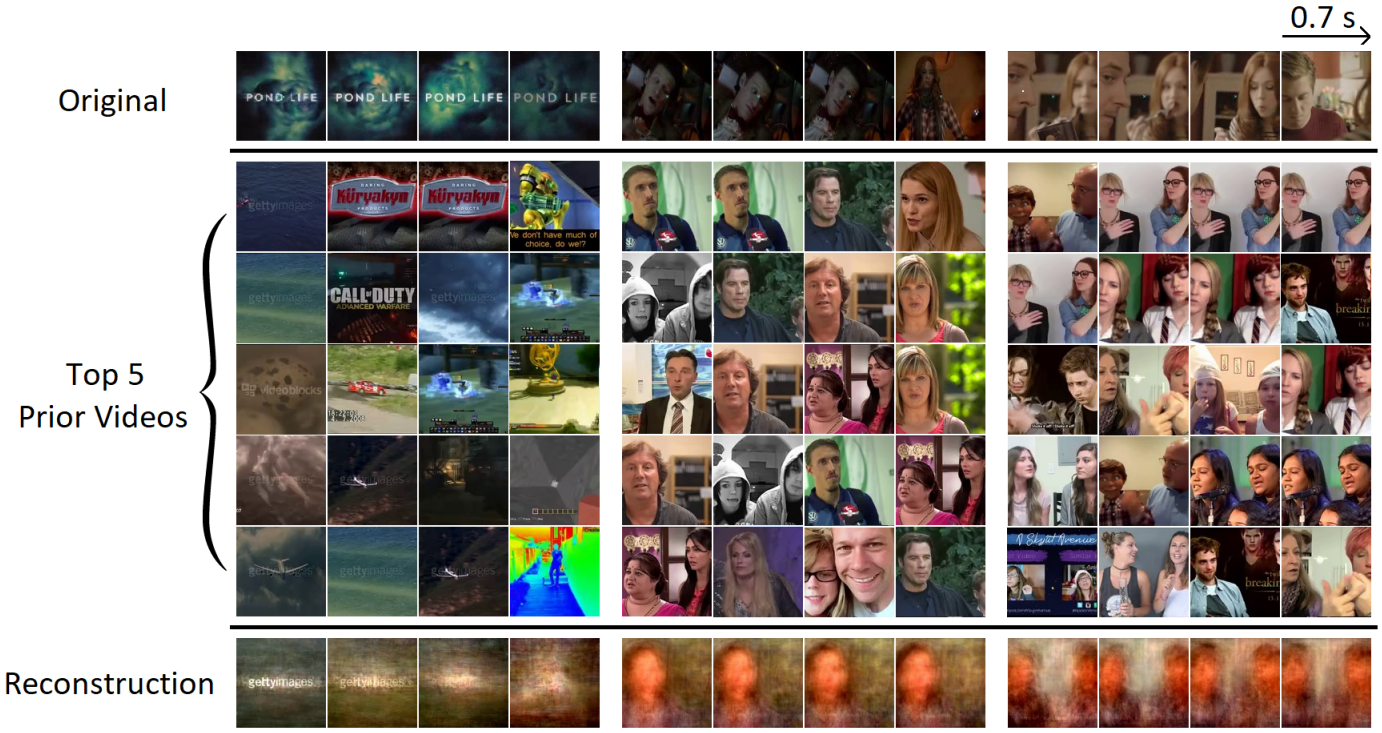

Figure 15: *Examples of successful reconstructions of three different scenes by the combined model. The second through sixth rows display the first five videos from the top 100 prior videos that create the reconstruction. Each frame within a scene is separated by 0.7 seconds.*

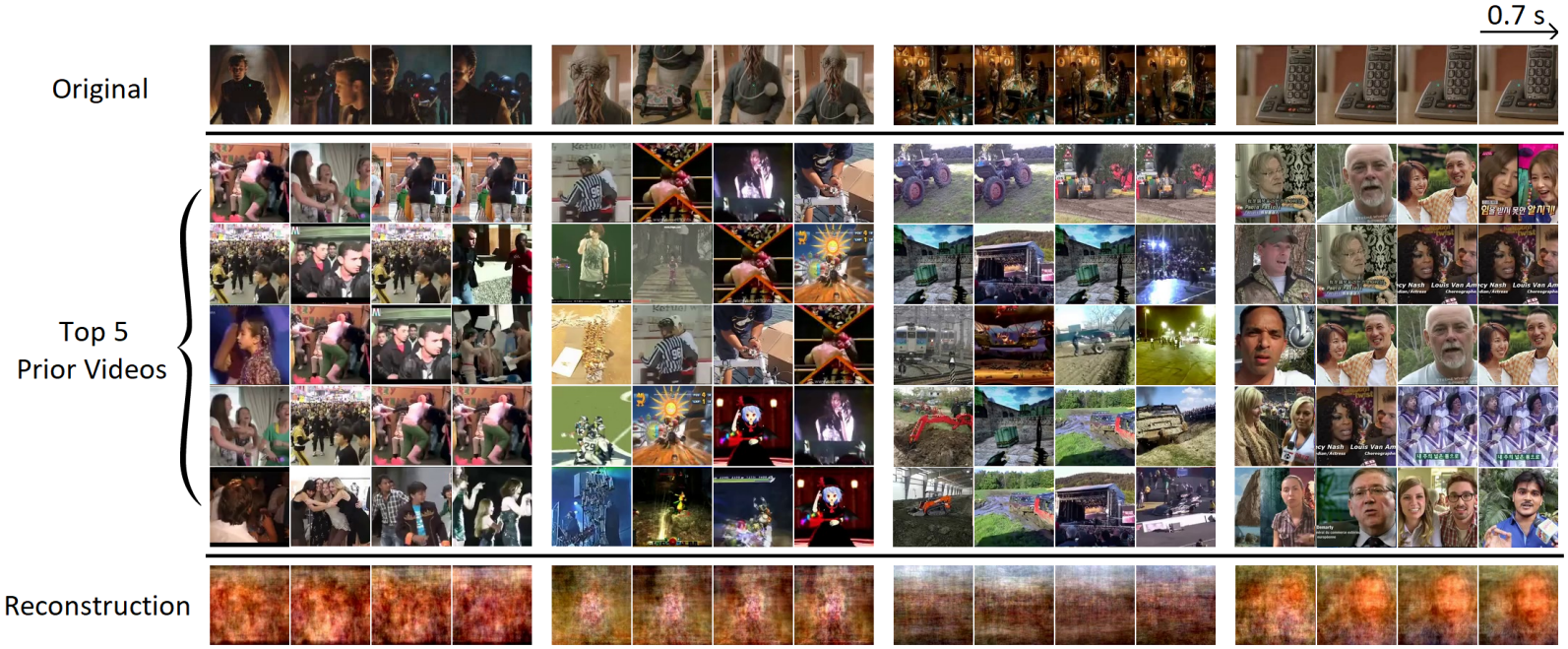

Figure 16: *Examples of unsuccessful reconstructions of four different scenes by the combined model. The second through sixth rows display the first five videos from the top 100 prior videos that create the reconstruction. Each frame within a scene is separated by 0.7 seconds.*

ery. The last column visually shows just a telephone, but the audio was that of someone leaving a voice-mail. Interestingly, although the motion-energy reconstruction is mainly just noise, the reconstructions of the R(2+1)D pooling layer and combined model depict a face, reconstructing a person behind the voice instead of the telephone itself.

When comparing the selected prior videos across the three models, it becomes clear that they focus on different elements in the original test videos. The motion-energy model seems to focus more on individual visual features in the prior videos, while the R(2+1)D pooling layer model focuses more on the entire scenes. This can for example be seen in the first scene of the successful reconstructions, where the motion-energy model separately tries to represent either the text or the background in the prior videos, while the R(2+1)D pooling layer model tries to represent those two visual features more at the same time. Another example of this can be seen in the second scene of the unsuccessful reconstructions, where the motion-energy model uses different prior videos to incorporate the alien face, its body, the round button and the background. In contrast, the R(2+1)D model seems to focus on the shape of the entire scene with its prior videos. The combined model seems to strike a balance between these two approaches such that the prior videos displaying important visual features still keep the overall shape in the scene into account.

To analyze the reconstructions, they were fed to the AlexNet neural network (Krizhevsky et al., 2012). In each of the eight layers of the network, the features of the reconstruc-

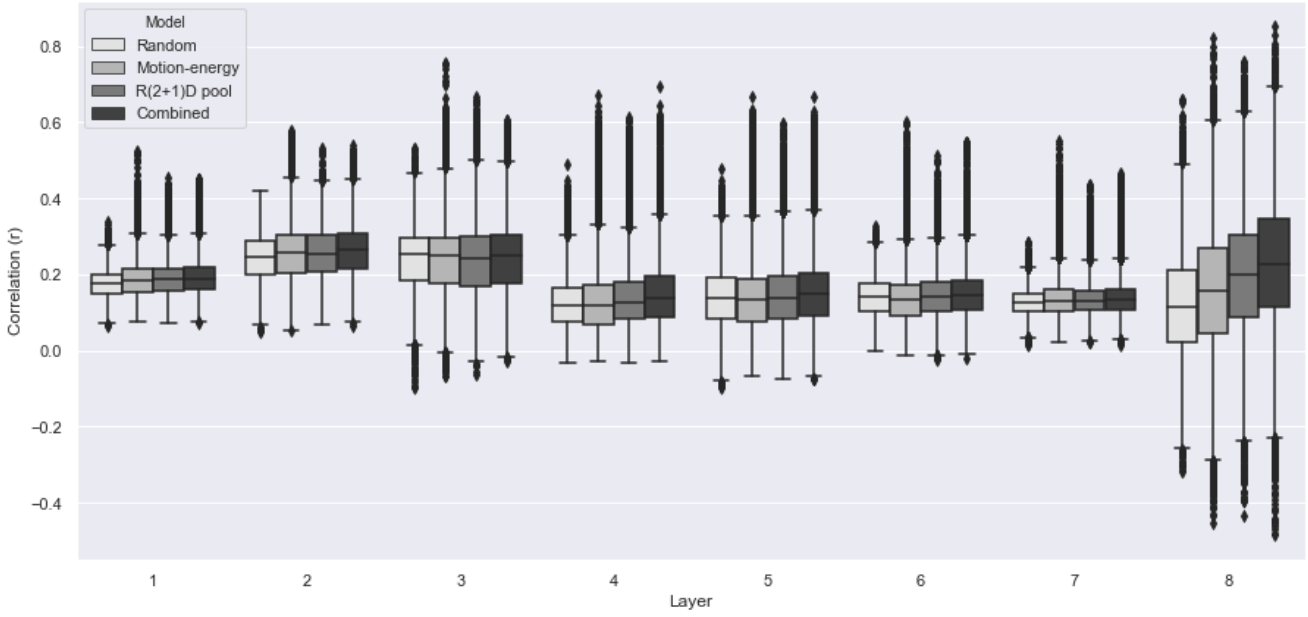

Figure 17: *Results of the correlation analyses based on the AlexNet network features for all eight layers between the original test video and a random reconstruction, the motion-energy model reconstruction, the R(2+1)D pooling layer model reconstruction and the combined model reconstruction.*

tions were compared to the features of the test video in the same layer using Pearson's correlation coefficients. To create a baseline, a random reconstruction was also created and compared to the test video. Figure 17 shows the results of this analysis in a boxplot. Only in layers 1, 2, 7 and 8 are all models significantly better than chance, in layer 4 the R(2+1)D pooling layer and the combined model are significantly better than the random model and in layers 5 and 6 it is just the combined model that is significantly better than chance ($p < 0.001$, Wilcoxon rank-sum test). In every layer except layer 3, it holds that the combined model is significantly better than both the motion-energy and R(2+1)D pooling model. In layer 3 there is no significant difference between any of the models. In layers 1, 4, 5, 6 and 8 it holds that the R(2+1)D pooling model is significantly better than the motion-energy model, in layer 2 the motion-energy model is significantly better than the R(2+1)D pooling model and in layer 7 there is no significant difference between the motion-energy model and the R(2+1)D pooling model ($p < 0.001$, Wilcoxon signed-rank test). The highest average score in the correlation coefficients is achieved with the combined model in layer 2 with a score of 0.26, but the average score with the biggest difference with the other models within a layer is achieved by the combined model in layer 8 with a difference of 0.11 with the random model, 0.07 with the motion-energy model and 0.03 with the R(2+1)D pooling model. Note that the standard deviation is much larger in layer 8 than in the other layers, this is probably caused by the fact that the purpose of this layer is to classify the images to one of the object categories. If the reconstruction is successful, it is probably recognizable enough to be classified in the right category. On the other hand, if it is unsuccessful and thus very noisy, it could be classified as anything and create a feature vector that is not correlated with the original frame at all.

3.3 Imagery

The voxel selections were also used to create reconstructions of imagery data that was collected. Two different intro sequences of the *Doctor Who* series appearing in the test set were imagined for ten seconds by the participant, in which a blue telephone box flies either through a blue or yellow colored space. For each TR, a top 100 of prior videos was collected in the same manner as for the other reconstructions. The intro sequences plus the reconstructions for each of the models can be found in Figure 18 and 19. Note that because it is unclear what the hemodynamic delay is between the process of imagining and the resulting BOLD response, the fMRI data was not shifted forward in time. It is thus very well possible that the intro sequence and the reconstruction do not align in time. Overall, the imagery reconstructions are very noisy. For the first imagery task, the motion-energy and combined model have a bright spotlight in the middle of the reconstructions from the fourth through the eighth frame, which could point to the participant focusing on the telephone box in the middle of his imagined scene. The R(2+1)D pooling layer model mainly seems to display noise. In the second imagery task, the bright spotlight for the motion-energy and combined model is visible as well, only the fifth frame clearly shows text, possibly representing the text that was indeed visible in the intro sequence. Interestingly, the R(2+1)D pooling layer model clearly shows a human shape in the first three frames. It is unclear to why this is. Possibly was the participant

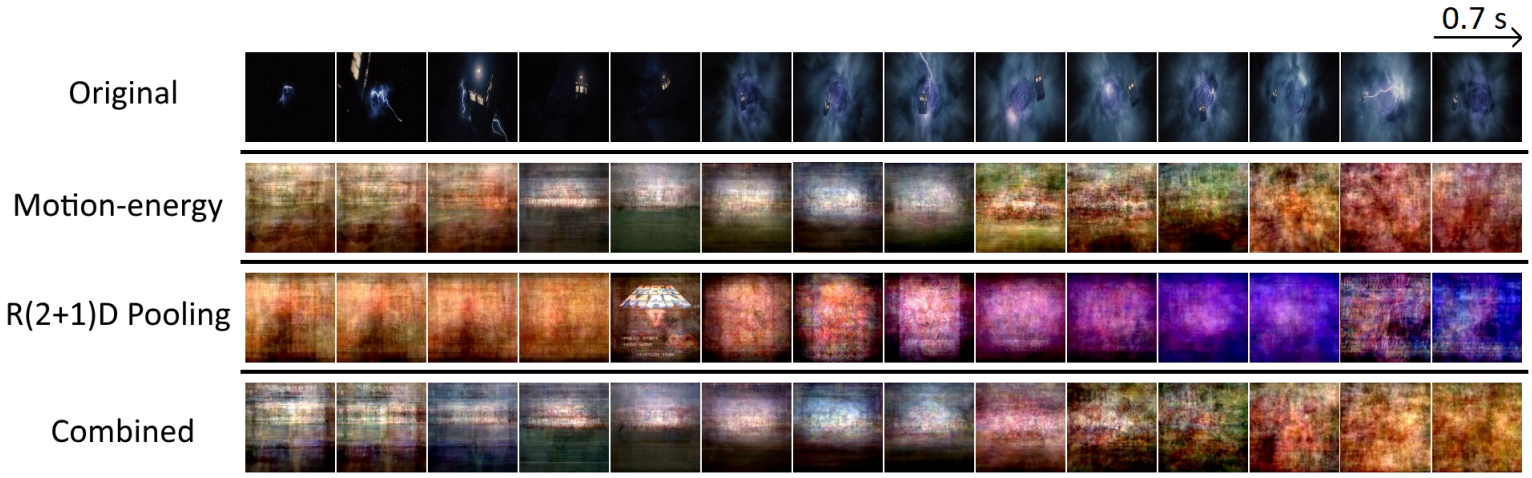

Figure 18: *The reconstructions of the first imagery task in which a blue telephone box flies through blue space by the motion-energy, the R(2+1)D pooling layer and the combined model. Each frame shown is separated by 0.7 seconds.*

thinking about someone right before the imagery tasks, or maybe the imagining of the words "Doctor Who" elicited the participant to think of the actual character that portrays Doctor Who. Nevertheless, the rest of the frames mainly display noise again. A clear difference between the two imagery tasks was the color of the background. However, since the color of the reconstructions appears to be mostly the same across the two imagery tasks, color does not seem to be a feature that was reconstructed correctly.

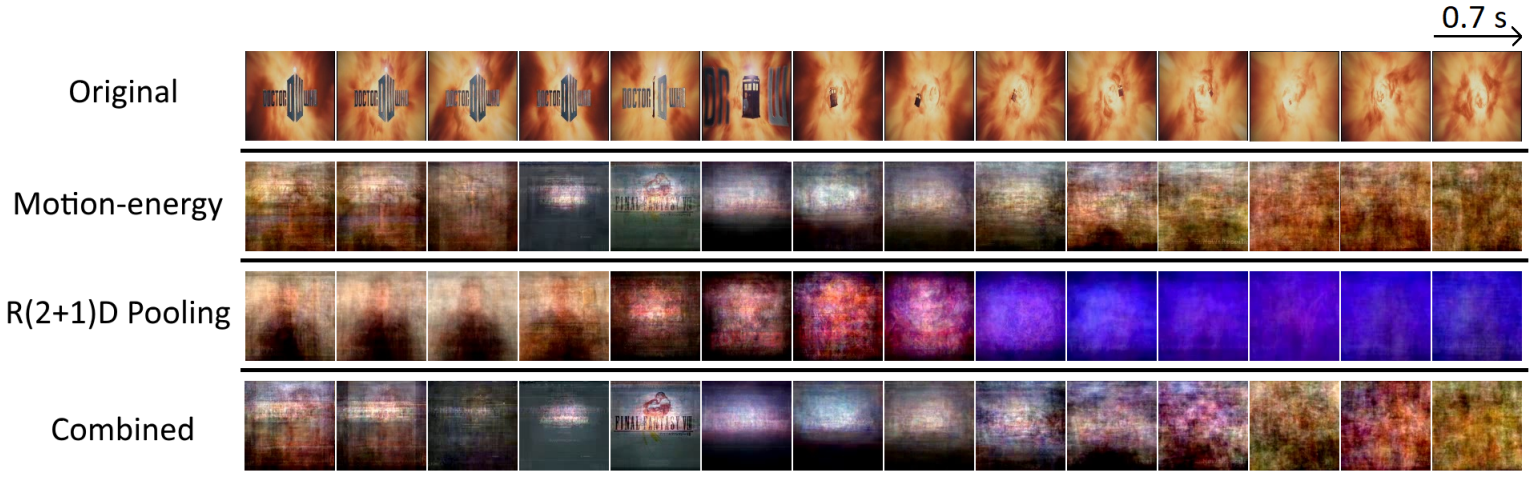

Figure 19: *The reconstructions of the second imagery task in which a blue telephone box flies through orange space by the motion-energy, the R(2+1)D pooling layer and the combined model. Each frame shown is separated by 0.7 seconds.*

4 Discussion

The present study elaborated on the work by Nishimoto et al. (2011) who successfully reconstructed naturalistic movies from cortical activity in the visual cortex by training an encoding model between motion-energy features and BOLD responses and combining this with a naturalistic movie prior in a Bayesian decoder. Our addition to their work was three-fold. In contrast to the two hours of fMRI data that was originally collected of the visual cortex from three participants, we used a dataset of 24 hours of fMRI data covering the entire brain from a single subject. A bigger dataset ensures that there are enough free parameters for the encoding model to be trained more adequately. Furthermore, a convolutional neural network was introduced as feature model to be able to capture the higher-order information in the visual cortex next to the motion-energy features that capture just lower-order information. Lastly, for each voxel in the encoding models it was decided for which hemodynamic delay its prediction performance was best, ensuring a more optimal mapping between features and BOLD responses.

By selecting the top 2000 most predictive voxels per encoding model, we were able to differentiate between the visual areas more sensitive to lower-order information and the areas more sensitive to higher-order information. In line with previous research, we found that lower-order information is mainly localized and best predicted in visual areas V1, V2 and V3, while the higher-order information is mainly localized and best predicted in the higher-order visual areas (Huth et al., 2012; Güçlü and van Gerven, 2015). However, a minority of voxels in the motion-energy voxel selection still also appeared in the higher-order visual areas in the combined model and a minority of voxels in the R(2+1)D pooling layer voxel selection appeared in the lower-order areas of the combined model. This can be explained by findings that there are higher-order voxels that are responsive to more basic features, and lower-order voxels that are responsive to more complex features (Hegdé and Van Essen, 2006). Interestingly, higher-order area MT is mostly encoded by the motion-energy voxels in our combined model, while research would suggest that it should be well predicted by higher layers in a convolutional neural network (Güçlü and van Gerven, 2017). Nonetheless, motion-energy was developed to be a model of human motion perception and previous work has shown it to be a good model of MT neurons (Nishimoto and Gallant, 2011). The fact that motion-energy revolves around motion specifically probably outweighs the higher-level features from the R(2+1)D model.

Through the selection of the hemodynamic delay per voxel, we have shown that there is a trend in increasing delays along the visual pathway, which is in line with the findings that the motion-energy model has a lower average in delays than the R(2+1)D pooling layer model. This result makes sense as we know that visual signals start in the early visual cortex and travel along the visual pathway to higher-order areas. The exception to this rule was found in an area thought to be the IPS, which could be explained by the fact that the IPS is involved in visual-spatial attention (Corbetta et al., 1995; Corbetta and Shulman, 2002), which is a relatively fast process. As the participant was fixating on the centre of the video, good visual-spatial attention was necessary to follow the events in the episodes, possibly

explaining why the voxels in the IPS were among the top voxels to be predicted.

The goal of this research was to create reconstructions of naturalistic movies from fMRI brain responses. Although shown to be possible by Nishimoto et al. (2011) and again in the present study, the quality of the reconstructions appears to be very sensitive to the particular naturalistic movies and the viewing conditions of the fMRI participant. Ultimately, the task seems to be too complex for the current method. Most naturalistic movies contain too many details that can't be reconstructed well by taking an average over prior videos, current successful reconstructions mainly seem to be single objects against a relatively simple background. As soon as there appear multiple objects in the same scene or objects that are not represented well by the prior videos, the quality of the reconstructions drops drastically. The same holds for the imagery task, it was probably a too complex scene to reconstruct well with this technique. Even though there were some characteristic features that should have been relatively easy to reconstruct like the text in the second task, because there was a lot more to imagine around it, it might have influenced the focus on specifically the text. When looking at the reconstructions across the three models, it becomes clear from the selected prior videos that each model focuses on different elements. The motion-energy model focuses on smaller shapes, while the R(2+1)D pooling layer model appears to be more influenced by the entire picture. The combined model seems to combine these two approaches. Overall, combining lower-order and higher-order information does appear to have an advantage in comparison to just using either one of these types of information. The AlexNet analyses showed that it was significantly better than the other models in all layers but one, pointing to the fact that there is merit to the approach of combining lower- and higher-order information.

4.1 Future work

There are some limitations to the current research when compared to the work by Nishimoto et al. (2011). Our prior set of 5 million videos is quite small compared to the prior set of Nishimoto et al., that was more than three times our size with 18 million. Logically, the more videos present in the prior set, the more variety in objects and events are available to make better reconstructions with. However, we still stand by our finding that that having many objects or uncommon objects in a scene will always be hard to reconstruct with the current method, even with a very large prior set. Secondly, our likelihood calculation was different. We did not use a multivariate Gaussian to express the likelihood between the observed and predicted BOLD response, but just took Pearson's correlation between the two. It is unclear whether this has had a big impact on the prior selection. Lastly, the fMRI data collected by Nishimoto et al. was based on videos with 15 Hz, or 15 frames per second (FPS). The standard FPS of video, however, is normally 24 Hz. This results in videos of 15 Hz feeling like they have been slowed down as each frame is stretched out of a longer period of time. The dataset that was used for this research had a normal FPS to simulate natural viewing conditions as much as possible, but this likely resulted in more noisy BOLD signals relative to the work by Nishimoto et al. (2011). This would explain why our correlation coefficients in the early visual cortex on the validation set in the motion-energy model, next to the fact

that our validation set was not an average over trials and thus contains a higher signal-to-noise ration, are about 0.2 lower than the corresponding correlation coefficients on the test set of the original work.

Relatively low prediction accuracies are overall a common problem when training encoding models on fMRI data. This is caused by two important factors. First, the features used are not good enough in representing the voxel activity. For this research, it would be a good step to not just use the last convolutional layer from the R(2+1)D model, but determine for each voxel which of all the convolutional layers had the best predictive power as in Güçlü and van Gerven (2015, 2017). This way all levels of information (low-, mid- and high-level) would be represented as best as possible. Alternatively, we could adopt recurrent neural networks instead of the feed-forward networks that normal convolutional neural networks belong to. This way the top-down influence and long-term dependencies in the neural information flow can be modelled and therewith hopefully improve the encoding prediction accuracies. Recurrent convolutional neural networks have already been successfully used in EEG research (Bashivan et al., 2015).

Another way of improving the prediction accuracies is to develop more complex encoding models. Currently, most encoding models are trained on a single voxel level. However, research has shown that the firing patterns between voxels have an important role in the decoding of stimuli from brain data since visual features elicit responses across multiple voxels (Chen et al., 2006; Engel et al., 1997). Thus, a encoding model that trains on a multi-voxel level might prove beneficial for the prediction accuracy as it is able to pick up more in-depth cortical signals. Such an approach was already taken by Miyawaki et al. (2008), who used a multi-voxel decoding model to reconstruct geometric shapes from the visual cortex and showed that it outperformed the decoding model based on single voxels.

As the quality of the reconstructions is one of the bigger concerns after analyzing our results, there is quite some future work possible in making accurate, detailed video reconstructions from brain data. A promising method from the last few years that could be adopted for creating such reconstructions are generative adversarial networks (GANs) (Goodfellow et al., 2014). GANs became popular because of their powerful ability to create original images that look very realistic. Recent research has already had success in the reconstruction of natural images from fMRI data (Güçlütürk et al., 2017; Seeliger et al., 2018; Shen et al., 2019), so it seems like the next logical step to test this method for the reconstruction of naturalistic video.

4.2 Conclusion

We have shown that combining higher-level and lower-level information is significantly better than using either of these levels of information in the reconstruction of naturalistic videos created from fMRI data. However, it must also be concluded that the current method of reconstruction is not sufficient enough to create scenes that have the complexity that regular naturalistic videos display and that there is still a lot of performance to be gained in the mapping between voxel activity and feature representations. Future work is necessary to create methods for more detailed reconstructions and better feature and encoding models.

References

Abu-El-Haija, S., Kothari, N., Lee, J., Natsev, P., Toderici, G., Varadarajan, B. and Vijayanarasimhan, S. (2016), 'Youtube-8m: A large-scale video classification benchmark', *arXiv preprint arXiv:1609.08675* .

Adelson, E. H. and Bergen, J. R. (1985), 'Spatiotemporal energy models for the perception of motion', *Journal of the Optical Society of America* **2**(2), 284–299.

Bashivan, P., Rish, I., Yeasin, M. and Codella, N. (2015), 'Learning representations from EEG with deep recurrent-convolutional neural networks', *arXiv preprint arXiv:1511.06448* .

Chen, W., Kato, T., Zhu, X.-H., Ogawa, S., Tank, D. W. and Ugurbil, K. (1998), 'Human primary visual cortex and lateral geniculate nucleus activation during visual imagery', *NeuroReport* **9**(16), 3669–3674.

Chen, Y., Geisler, W. S. and Seidemann, E. (2006), 'Optimal decoding of correlated neural population responses in the primate visual cortex', *Nature Neuroscience* **9**(11), 1412.

Corbetta, M. and Shulman, G. L. (2002), 'Control of goal-directed and stimulus-driven attention in the brain', *Nature reviews neuroscience* **3**(3), 201.

Corbetta, M., Shulman, G. L., Miezin, F. M. and Petersen, S. E. (1995), 'Superior parietal cortex activation during spatial attention shifts and visual feature conjunction', *Science* **270**(5237), 802–805.

Engel, S. A., Glover, G. H. and Wandell, B. A. (1997), 'Retinotopic organization in human visual cortex and the spatial precision of functional MRI', *Cerebral Cortex* **7**(2), 181–192.

Fukushima, K. (1980), 'Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position', *Biological Cybernetics* **36**(4), 193–202.

Gao, J. S., Huth, A. G., Lescroart, M. D. and Gallant, J. L. (2015), 'Pycortex: an interactive surface visualizer for fMRI', *Frontiers in Neuroinformatics* **9**, 23.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y. (2014), Generative adversarial nets, *in* 'Advances in Neural Information Processing Systems 27', pp. 2672–2680.

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

Güçlütürk, Y., Güçlü, U., Seeliger, K., Bosch, S., van Lier, R. and van Gerven, M. A. (2017), Reconstructing perceived faces from brain activations with deep adversarial neural decoding, *in* 'Advances in Neural Information Processing Systems 30', pp. 4246–4257.

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L. and

Pietrini, P. (2001), 'Distributed and overlapping representations of faces and objects in ventral temporal cortex', *Science* **293**(5539), 2425–2430.

He, K., Zhang, X., Ren, S. and Sun, J. (2016), Deep residual learning for image recognition, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition', pp. 770–778.

Hegdé, J. and Van Essen, D. C. (2006), 'A comparative study of shape representation in macaque visual areas v2 and v4', *Cerebral Cortex* **17**(5), 1100–1116.

Huth, A. G., Nishimoto, S., Vu, A. T. and Gallant, J. L. (2012), 'A continuous semantic space describes the representation of thousands of object and action categories across the human brain', *Neuron* **76**(6), 1210–1224.

Jing, Y., Liu, D., Kislyuk, D., Zhai, A., Xu, J., Donahue, J. and Tavel, S. (2015), Visual search at Pinterest, *in* 'Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining', ACM, pp. 1889–1898.

Kamitani, Y. and Tong, F. (2005), 'Decoding the visual and subjective contents of the human brain', *Nature Neuroscience* **8**(5), 679.

Kay, K. N., Naselaris, T., Prenger, R. J. and Gallant, J. L. (2008), 'Identifying natural images from human brain activity', *Nature* **452**(7185), 352.

Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P. et al. (2017), 'The kinetics human action video dataset', *arXiv preprint arXiv:1705.06950* .

Kosslyn, S. M., Thompson, W. L., Klm, I. J. and Alpert, N. M. (1995), 'Topographical representations of mental images in primary visual cortex', *Nature* **378**(6556), 496.

Krizhevsky, A., Sutskever, I. and Hinton, G. E. (2012), Imagenet classification with deep convolutional neural networks, *in* 'Advances in Neural Information Processing Systems 25', pp. 1097–1105.

McCulloch, W. S. and Pitts, W. (1943), 'A logical calculus of the ideas immanent in nervous activity', *The Bulletin of Mathematical Biophysics* **5**(4), 115–133.

Miyawaki, Y., Uchida, H., Yamashita, O., Sato, M., Morito, Y., Tanabe, H. C., Sadato, N. and Kamitani, Y. (2008), 'Visual image reconstruction from human brain activity using a combination of multiscale local image decoders', *Neuron* **60**(5), 915–929.

Monfort, M., Andonian, A., Zhou, B., Ramakrishnan, K., Bargal, S. A., Yan, Y., Brown, L., Fan, Q., Gutfreund, D., Vondrick, C. et al. (2019), 'Moments in time dataset: one million videos for event understanding', *IEEE Transactions on Pattern Analysis and Machine Intelligence* .

Naselaris, T., Kay, K. N., Nishimoto, S. and Gallant, J. L. (2011), 'Encoding and decoding in fMRI', *NeuroImage* **56**(2), 400–410.

Naselaris, T., Prenger, R. J., Kay, K. N., Oliver, M. and Gallant, J. L. (2009), 'Bayesian reconstruction of natural images from human brain activity', *Neuron* **63**(6), 902–915.

Nishimoto, S. and Gallant, J. L. (2011), 'A three-dimensional spatiotemporal receptive field model explains responses of area MT neurons to naturalistic movies', *Journal of Neuroscience* **31**(41), 14551–14564.

Nishimoto, S., Vu, A. T., Naselaris, T., Benjamini, Y., Yu, B. and Gallant, J. L. (2011), 'Reconstructing visual experiences from brain activity evoked by natural movies', *Current Biology* **21**(19), 1641–1646.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L. and Lerer, A. (2017), Automatic differentiation in PyTorch, *in* '31st Conference on Neural Information Processing Systems'.

Schoenmakers, S., Barth, M., Heskes, T. and van Gerven, M. A. J. (2013), 'Linear reconstruction of perceived images from human brain activity', *NeuroImage* **83**, 951–961.

Schoenmakers, S., Güçlü, U., Van Gerven, M. A. J. and Heskes, T. (2015), 'Gaussian mixture models and semantic gating improve reconstructions from human brain activity', *Frontiers in Computational Neuroscience* **8**, 173.

Seeliger, K., Güçlü, U., Ambrogioni, L., Güçlütürk, Y. and van Gerven, M. A. (2018), 'Generative adversarial networks for reconstructing natural images from brain activity', *NeuroImage* **181**, 775–785.

Seeliger, K., Sommers, R. P., Güçlü, U., Bosch, S. E. and van Gerven, M. A. J. (2019), A large single-subject fMRI dataset for probing brain responses to naturalistic stimuli in space and time. Submitted. Dataset published on *Donders Data Repository*. Persistent identifier: `11633/di.dcc.DSC_2018.00082_134`.

Shen, G., Horikawa, T., Majima, K. and Kamitani, Y. (2019), 'Deep image reconstruction from human brain activity', *PLOS Computational Biology* **15**(1), 1–23.

Thirion, B., Duchesnay, E., Hubbard, E., Dubois, J., Poline, J.-B., Lebihan, D. and Dehaene, S. (2006), 'Inverse retinotopy: inferring the visual content of images from brain activation patterns', *NeuroImage* **33**(4), 1104–1116.

Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y. and Paluri, M. (2018), A closer look at spatiotemporal convolutions for action recognition, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition', pp. 6450–6459.

Wang, J., Song, Y., Leung, T., Rosenberg, C., Wang, J., Philbin, J., Chen, B. and Wu, Y. (2014), Learning fine-grained image similarity with deep ranking, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition', pp. 1386–1393.

Watson, A. B. and Ahumada, A. J. (1985), 'Model of human visual-motion sensing', *JOSA A* **2**(2), 322–342.

Zeiler, M. D. and Fergus, R. (2014), Visualizing and understanding convolutional networks, *in* 'European conference on computer vision', Springer, pp. 818–833.