Forecasting fMRI Images From Video Sequences: Linear Model Analysis

A Preprint

Dorin Daniil
dorin.dd@phystech.edu

Kiselev Nikita
kiselev.ns@phystech.edu

Grabovoy Andrey
grabovoy.av@phystech.edu

February 7, 2024

Abstract

The problem of reconstructing the dependence between fMRI sensor readings and human perception of the external world is investigated. The dependence between the sequence of fMRI images and the video sequence viewed by a person is analyzed. Based on the dependence study, a method for approximating fMRI readings from the viewed video sequence is proposed. The method is constructed under the assumption of the presence of a time invariant hemodynamic response time dependence of blood oxygen level. A linear model is independently constructed for each voxel of the fMRI image. The assumption of markovality of the fMRI image sequence is used. To analyze the proposed method, a computational experiment is performed on a sample obtained during tomographic examination of a large number of subjects. The dependence of the method performance quality on the hemodynamic response time is analyzed on the experimental data. Hypotheses about invariance of model weights with respect to a person and correctness of the constructed method are tested.

Keywords neuroimaging · fMRI · video sequences · correlation analysis · hypothesis testing · linear model · forecasting

1 Introduction

A set of techniques that visualize the structure and function of the human brain, is called *neurovisualization*. Neuroimaging [1] techniques such as ECG, CT, MRI, and fMRI, are used to study the brain and to detect disease and mental disorders.

Functional magnetic resonance imaging, or *fMRI*. is a type of MRI based on changes in blood flow, caused by neural activity in the brain [2]. These changes don't occur instantaneously, but with a delay, which is 4–8 seconds [3]. It's because it takes the vascular system a long time to respond to the brain's need for glucose [4, 5, 6].

In fMRI imaging, sequences of echoplanar imaging (EPI) [7, 8, 9]. Processing of areas with varying signal intensity depending on the method of activation, the type of artifact. depending on the method of activation, the type of artifacts and duration is carried out using special methods and programs [3, 10, 11]. The processed results are formalized as activation maps, which are combined with the localization of anatomical structures of the cerebral cortex.

The fMRI method plays a major role in neuroimaging, but has some important limitations. The works of [12, 13] discuss the temporal and spatial resolutions of fMRI. Temporal resolution is a significant disadvantage of this method. Another disadvantage of fMRI — the inevitable noise, associated with movement of the object in the scanner, human heartbeat and respiration, thermal fluctuations of the device itself, etc. In the work [14] proposed methods of graph-based methods for suppressing the above-mentioned noises and demonstrates their effectiveness in the task of epilepsy and depression detection.

In fMRI, the subject is given a variety of test tasks and external stimuli that induce activation of specific external stimuli are applied, causing activation of certain localized areas of the brain that are responsible for performing of the brain responsible for its functions. Various test tasks are applied: movements of fingers and limbs [15, 16], image finding and examination of a chessboard [17, 18], listening to non-specific noises, single words or coherent text [19, 20]. Changes in human brain activity during fMRI examinations could also be caused by viewing the video of [21], which is the subject of this paper.

The most well-known video processing methods are based on 3D convolutions [22]. The difference between 3D and 2D convolutions is the simultaneous work with the spatial and temporal part of the information. of information. A significant disadvantage of these methods is the strong increase in the number of model parameters and the large computational costs. One of the most modern and improving architectures of neural networks for image processing is the ResNet [23] residual neural network. It allows training deep neural networks (up to 152 layers) with high accuracy, overcoming the problem of gradient fading that arises when training deep networks.

The real work is devoted to recovering the dependence between fMRI images and video sequences. The assumption that such a dependence exists is used. In addition, it is assumed that there is a constant time delay between the image and the video sequence [6]. The dependence of the fMRI image on one image and the previous image is tested. The delay time acts as a hyperparameter of the model. Based on the dependence analysis, a method is proposed to approximate the fMRI readings from the the viewed video sequence.

According to a study [24], when patients undergo fMRI, viewing a video sequence activates a specific cortical network. This network includes including the occipital lobe and frontal regions. This network is located predominantly in the right hemisphere. In this paper, we're considering an approach that uses these parts of the brain to analyze latency time.

The data on which the dependence hypothesis is tested and the demonstration of the work of the constructed method are presented in the paper [25]. This data set was obtained from examination of a group of 63 subjects. Thirty of them underwent fMRI examination. They were asked to perform the same task — viewing a short audiovisual movie. For it, annotations were generated in the paper under review, containing, among other things, information about the time of appearance and disappearance of of individual words, objects and characters. The methods of audio and video annotation are described in detail in the [26] and [27].

2 Problem statement

Frame rate $\nu \in \mathbb{R}$ and duration $t \in \mathbb{R}$ of the video sequence are set. The video sequence is set as

$$\mathbf{P} = [\mathbf{p}_1, \ldots, \mathbf{p}_{\nu t}], \quad \mathbf{p}_\ell \in \mathbb{R}^{W \times H \times C}, \tag{1}$$

with width, height and number of image channels W, H and C, respectively.

Let us denote the frequency of fMRI images by $\mu \in \mathbb{R}$. We set the sequence of images

$$\mathbf{S} = [\mathbf{s}_1, \ldots, \mathbf{s}_{\mu t}], \quad \mathbf{s}_\ell \in \mathbb{R}^{X \times Y \times Z}, \tag{2}$$

where X, Y, and Z — the dimensions of the voxel image.

The problem is to construct a mapping that accounts for the Δt delay between the fMRI image and the video sequence, as well as previous tomographic readings. fMRI image and the video sequence, as well as previous tomographic readings. Formally, it is necessary to find such a mapping $\mathbf{g}$ that

$$\mathbf{g}(\mathbf{p}_1, \ldots, \mathbf{p}_{k_\ell - \nu\Delta t}; \mathbf{s}_1, \ldots, \mathbf{s}_{\ell-1}) = \mathbf{s}_\ell, \ \ell = 1, \ldots, \mu t, \tag{3}$$

where for the ℓ-th fMRI image the number of the corresponding image k_ℓ is determined by the formula

$$k_\ell = \frac{\ell \cdot \nu}{\mu}. \tag{4}$$

3 Proposed forecasting method

The scheme of the proposed fMRI image reconstruction method is shown in Figure 1.

We denote the fMRI image as $\mathbf{s}_\ell = [v^\ell_{ijk}] \in \mathbb{R}^{X \times Y \times Z}$, where $v^\ell_{ijk} \in \mathbb{R}_+$ — value of the corresponding voxel. In order to reduce the running time of the method, we propose to use compression of fMRI images by dimensionality reduction. Compression by a factor of 2 is represented in the form of a mapping

$$\boldsymbol{\chi} : \mathbb{R}^{X \times Y \times Z} \to \mathbb{R}^{X/2 \times Y/2 \times Z/2}.$$

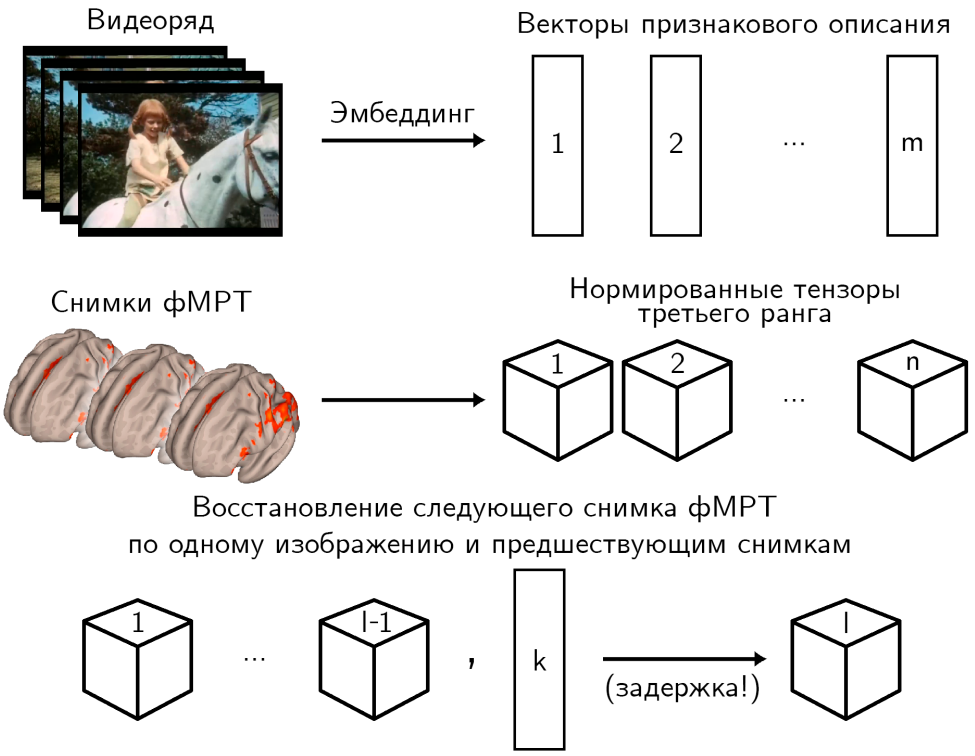


Figure 1: Method scheme

A compression of 2^k times is obtained by applying $\boldsymbol{\chi}$ successively k times. In the following, for simplicity, we keep the notation of image dimensions $X \times Y \times Z$.

Suppose that the Markov property is satisfied for the sequence of snapshots, i.e., each snapshot depends only on one image and the previous snapshot. Then the corresponding mapping is written in the form

$$\mathbf{g}(\mathbf{p}_{k_\ell - \nu\Delta t}) = \mathbf{s}_\ell - \mathbf{s}_{\ell-1} = \boldsymbol{\delta}_\ell, \ \ell = 2, \ldots, \mu t. \tag{5}$$

where $\boldsymbol{\delta}_\ell = [v^\ell_{ijk} - v^{\ell-1}_{ijk}] = [\delta^\ell_{ijk}] \in \mathbb{R}^{X\times Y\times Z}$ — the difference between two consecutive snapshots.

Mapping $\mathbf{g} : \mathbf{P} \to \mathbf{S}$ is represented as a composite of the other two:

$$\mathbf{g} = \boldsymbol{\varphi} \circ \boldsymbol{\psi},$$

$$\boldsymbol{\psi} : \mathbf{P} \to \mathbb{R}^d \text{ — image vectorization,}$$
$$\boldsymbol{\varphi} : \mathbb{R}^d \to \mathbf{S} \text{ — restorable mapping.}$$

For each image from the video sequence, we have an embedding vector of dimension d:

$$\mathbf{x}_\ell = [x^\ell_1, \ldots, x^\ell_d]^\mathsf{T} \in \mathbb{R}^d, \ \ell = 1, \ldots, \nu t.$$

The ResNet152 neural network architecture without the last linear layer is used.

Given (4), the total number of pairs (image, snapshot) is $N = \mu(t - \Delta t)$. Thus, for each voxel a sample is given

$$\mathfrak{D}_{ijk} = \{(\mathbf{x}_\ell, \delta^\ell_{ijk}) \mid \ell = 2, \ldots, N\}.$$

The regression task is set

$$y_{ijk} : \mathbb{R}^d \to \mathbb{R}. \tag{6}$$

A linear model is used with a vector of parameters

$$\mathbf{w}_{ijk} = [w^{ijk}_1, \ldots, w^{ijk}_d]^\mathsf{T} \in \mathbb{R}^d :$$

$$f_{ijk}(\mathbf{x}, \mathbf{w}_{ijk}) = \langle \mathbf{x}, \mathbf{w}_{ijk} \rangle. \tag{7}$$

For the model f_{ijk} with its corresponding parameter vector $\mathbf{w}_{ijk} \in \mathbb{R}^d$ define a quadratic loss function with L_2 regularization:

$$\mathcal{L}_{ijk}(\mathbf{w}_{ijk}) = \sum_{\ell=2}^{N} \left(f_{ijk}(\mathbf{x}_\ell, \mathbf{w}_{ijk}) - \delta_{ijk}^\ell \right)^2 + \alpha \|\mathbf{w}_{ijk}\|_2^2, \tag{8}$$

where $\alpha \in \mathbb{R}$ — regularization coefficient.

It is required to find the parameters that give a minimum to the loss functional $\mathcal{L}_{ijk}(\mathbf{w}_{ijk})$ for given hyperparameters Δt and α:

$$\hat{\mathbf{w}}_{ijk} = \arg\min_{\mathbf{w}_{ijk}} \mathcal{L}_{ijk}(\mathbf{w}_{ijk}). \tag{9}$$

The minimum of the loss function is found by the least squares method. Let's define the matrix of objects-features

$$\mathbf{X} = [\mathbf{x}_2, \ldots, \mathbf{x}_N]^\mathsf{T} = [x_j^i] \in \mathbb{R}^{(N-1)\times d} \tag{10}$$

and a vector whose components are the differences of values of the same voxel in different images,

$$\boldsymbol{\Delta}_{ijk} = [\delta_{ijk}^2, \ldots, \delta_{ijk}^N]^\mathsf{T} \in \mathbb{R}^{N-1}. \tag{11}$$

The solution is written in the form

$$\hat{\mathbf{w}}_{ijk} = (\mathbf{X}^\mathsf{T}\mathbf{X} + \alpha \mathbf{I})^{-1} \mathbf{X}^\mathsf{T} \boldsymbol{\Delta}_{ijk}. \tag{12}$$

Let us obtain a formula for reconstructed fMRI images. Let's introduce a matrix of weights

$$\hat{\mathbf{W}} = [\hat{\mathbf{w}}_1, \ldots, \hat{\mathbf{w}}_{XYZ}]^\mathsf{T} = [\hat{w}_j^i] \in \mathbb{R}^{XYZ \times d}. \tag{13}$$

Let's introduce for tensors $\mathbf{s}_\ell, \boldsymbol{\delta}_\ell \in \mathbb{R}^{X \times Y \times Z}$ vectors

$$\mathbf{s}_\ell^R = [v_1^\ell, \ldots, v_{XYZ}^\ell]^\mathsf{T}, \; \boldsymbol{\delta}_\ell^R = [\delta_1^\ell, \ldots, \delta_{XYZ}^\ell]^\mathsf{T} \in \mathbb{R}^{XYZ}.$$

Then the vector of the forecasted image is found by the formula

$$\hat{\mathbf{s}}_\ell^R = \mathbf{s}_{\ell-1}^R + \hat{\boldsymbol{\delta}}_\ell^R = \mathbf{s}_{\ell-1}^R + \hat{\mathbf{W}}\mathbf{x}_\ell. \tag{14}$$

4 Numerical experiment

To analyze the performance of the proposed method and test the hypotheses a computational experiment was carried out.

The sample presented in [25] was used as data. The dataset contains the results of examination of 63 subjects. For thirty of them fMRI readings are known. There are 16 males and 14 females, ranging in age from 7 to 47 years. The mean age of the subjects — 22 years.

Characteristics of the sample: duration of examination, frame rates of fMRI video sequences and images, and their dimensions are summarized in Table 1.

Table 1: Dataset Description

Name	Notation	Value
Duration of examination	t	390 s
Video frame rate	ν	25 Hz
fMRI frame rate	μ	1.64 Hz
Video dimensions	W, H, C	640, 480, 3
fMRI dimensions	X, Y, Z	40, 64, 64

The sample was divided into training and test samples in the ratio of 70% and 30%, respectively. The quality criterion for fMRI image reconstruction is MSE — the sum of squares of deviations between the true and reconstructed images, averaged over all voxels of each image. from the test sample.

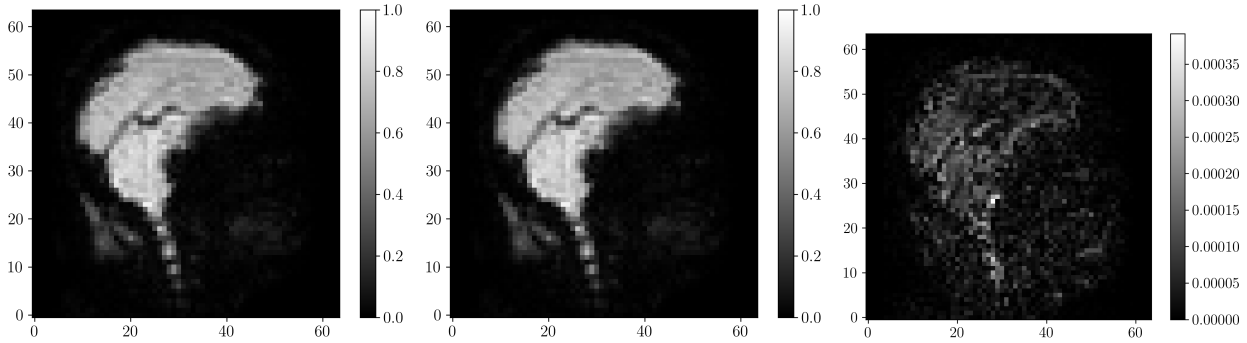
To reduce the running time of the algorithm, the fMRI image is precompressed using MaxPool3D layer. Compression ratios of 1, 2, 4 and 8 are considered. The voxel values are normalized to $[0; 1]$ by the MinMaxScale procedure.

Table 2 summarizes the specifications of the computer on which the computational experiment was on which the computational experiment was performed.

Table 2: PC Specification

Element	Description
CPU	Intel Core i7-7700 3.6 GHz
GPU	NVIDIA GeForce GTX 1060 3 GB
RAM	16 GB 2400 MHz
Hard Drive	M.2 SSD
OS	Windows 10

Method performance. Figure 2 shows slices of the true and reconstructed images from the test sample. Figure 2.(c) shows the difference between them. To demonstrate the performance of the algorithm, the 7th subject was selected, $\Delta t = 5$s, compression factor 1, regularization factor $\alpha = 1000$. The 20th slice along the first coordinate of the 37th image in the sequence was considered. Since the voxel values are normalized to the segment $[0; 1]$, an error of the order of 10^{-3} indicates a fairly accurate prediction.



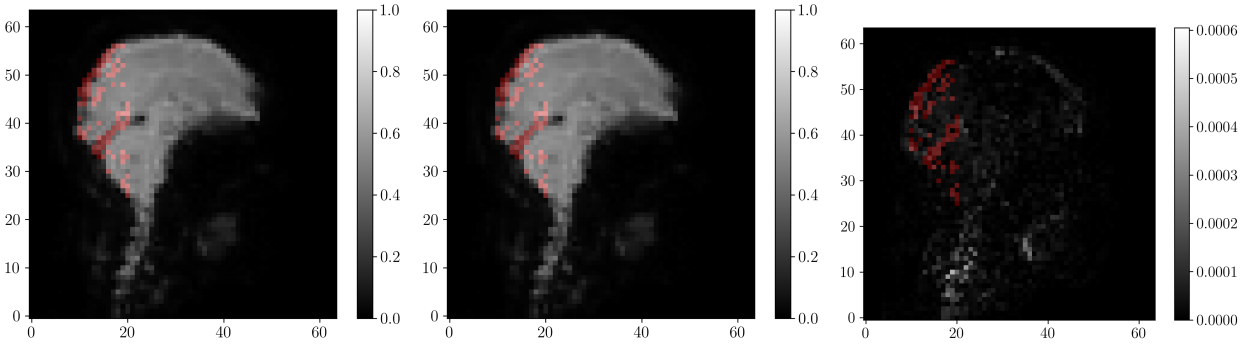
(a) Test **(b)** Predicted **(c)** Difference

Figure 2: Slices of fMRI images from the test sample

Delay time analysis. The dependence of recovery quality on delay time was investigated. The 47th subject and 4x compression were chosen for the example. The left graph in Figure 4 shows the dependence of the MSE metric on the delay time Δt. The study confirms that the most active part of the brain is the most active part of the brain in this examination — the occipital lobe. The other parts contribute noise to the considered dependence. In the present work, the above-mentioned region is localized, as shown in Figure 3. To localize the region, the lower third and the right two thirds of the volumetric of the tomographic image. The area highlighted in red is the area that contains the 3% of the most variable voxels in the occipital lobe. For this purpose, all voxels of the localized area were ordered by descending order of the total absolute change in values. Then 3% of voxels with the largest changes were selected. The MSE metric was recalculated exactly on this part of the image. The corresponding graph is shown on the right side of Figure 4. There is a more distinct minimum at $\Delta t \approx 5$ seconds.

Optimal regularization parameter. The dependence of MSE on the regularization parameter α was analyzed. Compression factors 1, 2, 4, and 8 were considered. The corresponding graphs are shown in Figure 5. Averaging over the subjects was performed to construct the graph. The limits of standard deviation are marked. The graphs show that the optimal value of the coefficient $\alpha \approx 1000$. The curve shape is preserved regardless of the compression factor of fMRI images.

Effect of image compression ratio on method runtime. We compare the training time of the model when using different compression coefficients of fMRI images. Coefficients 1, 2, 4 and 8 are considered.



(a) True **(b)** Predicted **(c)** Difference

Figure 3: Localization of the most active zone

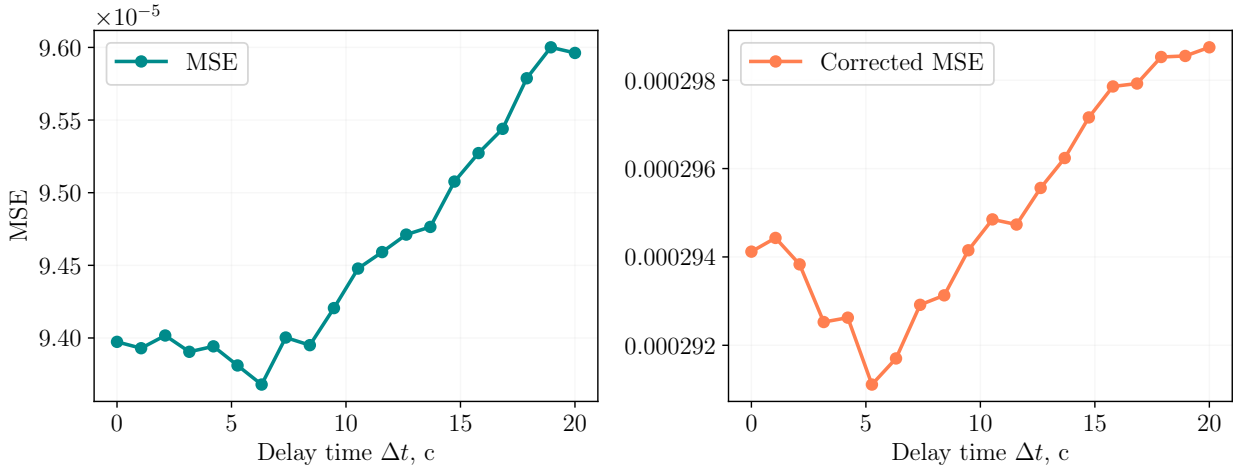


Figure 4: Dependence of MSE on delay time

For each value of the compression ratio, the average value of the model training time for all subjects is calculated. value of the model training time. The standard deviation is calculated. The experimental results are summarized in Table 3. The running time of the method is significantly reduced when using pre-compression of fMRI images. The experiment with the selection of the optimal regularization coefficient confirms that the compression of the images does not change the dependences.

Table 3: Dependence of model training time on compression ratio

Compression coefficient	Mean time, s	Std, s
1	36.3	6.1
2	6.7	0.5
4	1.6	0.1
8	1.4	0.3

Analyzing the distribution of model weights. A graph of the distribution of the values of the components of the model weight vector was plotted. To construct it, we averaged over all voxels for the 4th subject. The result is shown in Figure 6. The model weights do not lie in the neighborhood of any particular value, that is, their distribution is not degenerate. This result is quite consistent with reality, because a human being, while viewing pays attention to certain parts of the frame, such as characters or other details. other details.

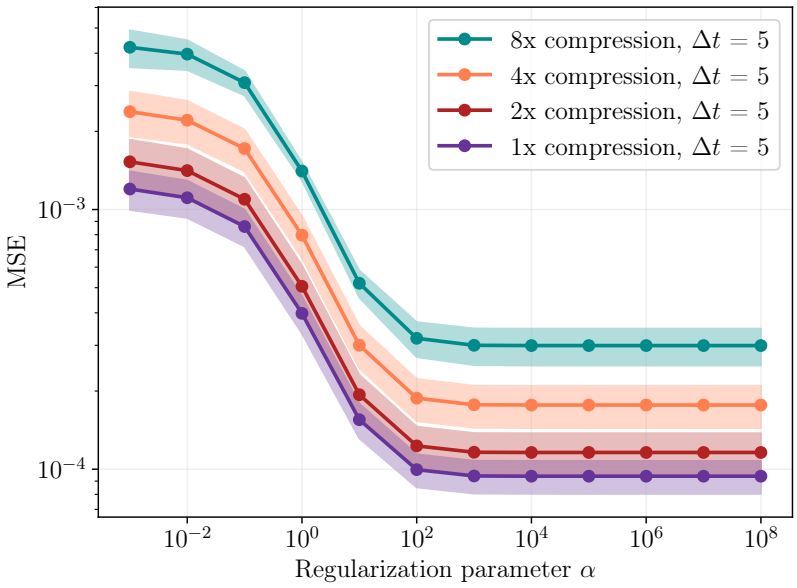


Figure 5: Dependence of MSE metric on regularization parameter α on images from the test sample

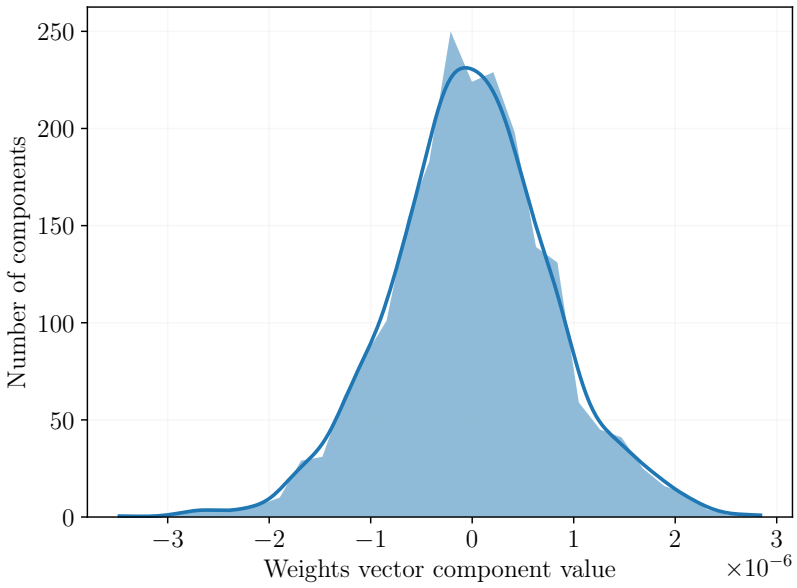


Figure 6: Weights vector component distribution

Hypothesis of invariance of model weights with respect to humans. The hypothesis of invariance of the model weights with respect to the person was tested: Using one subject's weight matrix to reconstruct another subject's fMRI images. The MSE metric on the test sample was used. The results are presented in Table 4. The 4th and 7th subjects were considered. The weight matrix of the 4th was used to reconstruct the of the 7th subject's images. The MSE values are almost the same.

Table 4: Testing the hypothesis of invariance of model weights with respect to humans

Weights matrix	True	Mixed	Difference
MSE	$9.7494 \cdot 10^{-5}$	$9.7498 \cdot 10^{-5}$	$3.96 \cdot 10^{-9}$

A similar experiment was conducted for each pair of subjects. The obtained results are presented in Figure 7, which was obtained as follows. Some subject (corresponding to a row of the matrix) is considered, MSE — «true» is calculate.d for him. Next, another subject is considered (corresponding to a column of the matrix), its matrix of weights is taken, and a prediction is made for the first subject, then the MSE — «subtracted» is calculated. The difference between the resulting MSE as a percentage of the «true» is entered into the matrix. A positive value means that the «mixed» MSE is greater than the «true». A negative — that the «substituted» is smaller. That is, there is a MAPE on the heatmap. The ideal model should result in only positive deviation values, however, as can be seen in Figure 7, there are negative values in the matrix. Nevertheless, they are rather small, namely, they correspond to deviations of the order of 1%. This is explained by the fact that the model is quite simple, and therefore has a high generalizing ability. However, this does not prevent us from concluding that the data do not contradict the hypothesis about the invariance of the model weights with respect to humans.

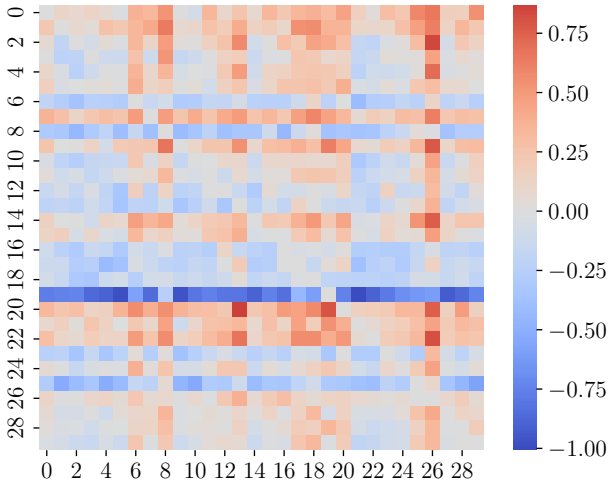


Figure 7: MAPE of MSE changing when predicting on the mixed weight matrix

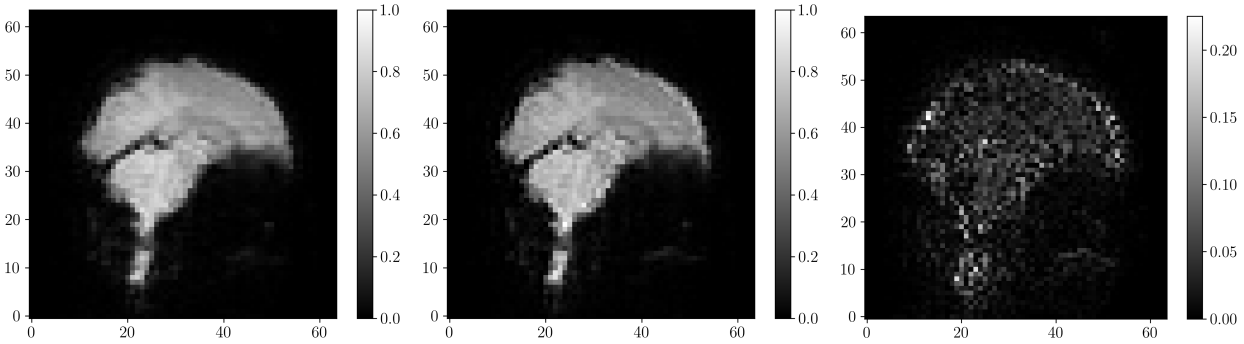
Method corectness. The quality of the method performance on uninformative data is considered. A matrix consisting entirely of units was taken as a matrix of objects-signs $\mathbf{X}$. Comparison with the results on the present feature-description matrix was made. To the first snapshot of the 35th subject, all the reconstructed changes in voxel values. As a result, we have the last snapshot of the sequence. Figure 8 are slices of the last true and recovered snapshots from the test sample. Figure 8.(c) shows the difference between them. The results on the uninformative ones are demonstrated in Figure 9. The difference between the true and recovered images when working with uninformative data is much higher, which confirms that there is a correlation between the sensor readings and the images from the video sequence. The numerical results are summarized in Table 5.

Table 5: Quality of method performance on uninformative data

Data	True	Uninformative	Difference
MSE	$4.87 \cdot 10^{-4}$	$1.76 \cdot 10^{-3}$	$1.27 \cdot 10^{-3}$

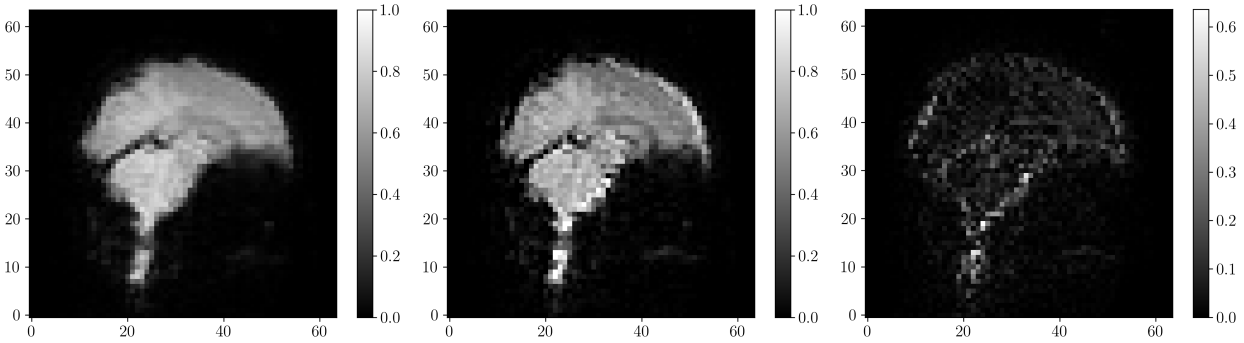
Correlation data analysis. Let us carry out an additional study of the data presented in the sample [25]. As mentioned earlier, the dataset also contains information about the time of appearance and disappearance of individual objects in the frame. We use this information to compose a feature-based description of the images of the video sequence of lower dimensionality. In total, there is information about 135 different objects. Let's encode each image with a vector of 0 and 1 of dimension 135, where 0 corresponds to the absence of an object in the frame and 1 to the absence of an object in the frame. object in the frame, and 1 — its presence.

The object that most often appears and disappears from the frame — Pippi, the main character in the movie shown to the subjects. The number of different fragments where she is in the frame is 26. We investigate the cross-correlation of the time series corresponding to this object with the with the time series of fMRI images.



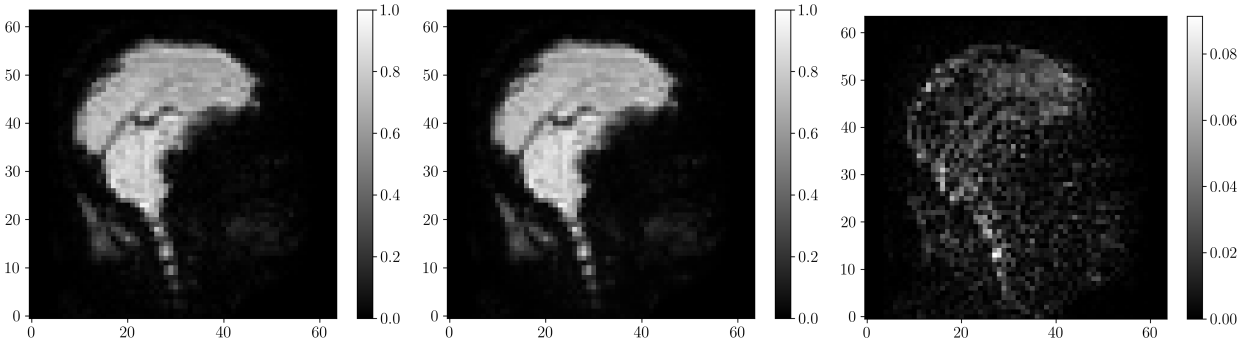
(a) True **(b)** Predicted **(c)** Difference

Figure 8: Slices of fMRI images from the test sample



(a) True **(b)** Predicted **(c)** Difference

Figure 9: Slices of fMRI images from the test sample (uninformative data)



(a) True **(b)** By cross-correlation **(c)** Difference

Figure 10: Slices of fMRI images from the test sample (compared to cross-correlation)

Figure 10.(a) shows a real fMRI image from the test sample. The 7th subject was selected, and the 20th slice on the first coordinate of the 37th image in the sequence. Figure 10.(b) presents the snapshot that was obtained from the cross-correlation function values. Namely, for each voxel, a time series corresponding to its change over time was obtained. Next, its cross-correlation function with the selected object was computed. Then, an offset was selected that corresponds to a delay time on the order of 5 seconds. The obtained number for each voxel was normalized to the segment $[0; 1]$.

As can be seen from Figure 10.(c), the images are almost identical. The MSE value is $7.8 \cdot 10^{-5}$. This further confirms the fact that the human brain reacts to the appearance and disappearance of specific objects in the frame.

5 Conclusion

In this paper we consider the task of restoring the dependence between the readings fMRI sensors and human perception of the external world. The method of approximation of fMRI images sequence by video sequence is proposed. The method takes into account the hemodynamic response time — the delay time between the images from the video sequence and the fMRI images. A linear model is independently constructed for each voxel of the fMRI image. Each linear model is constructed under the assumption that the sequence of fMRI images is Markovian. In the experiments, the optimal value of the delay time is selected for each subject. The optimal value is found from analyzing the plot of MSE versus delay time. The regularization coefficient is selected. The effect of fMRI image compression ratio on model training time is investigated. It is assumed that the occipital lobe of the brain is responsible for information from visual organs. MSE correction is performed based on localization of this region and selection of the most changing voxels. With this construction, the graph has a characteristic minimum corresponding to the optimal value of the delay time. The obtained value of the delay time is consistent with neurobiological information. The experimental MSE values are small, indicating that there is a correlation between the data. The variation of images in the video sequence is taken into account since the distribution of model weights is not degenerate. The hypothesis of invariance of model weights with respect to humans is tested. The correctness of the method is confirmed by experiments with random data.

References

[1] Ju V Puras and EV Grigorieva. The neurovisualization methods in diagnostics of head injury. part 1. computer tomography and magnetic resonance imaging. *Russian journal of neurosurgery*, (2):7–16, 2014.

[2] Gary H. Glover. Overview of functional magnetic resonance imaging. *Neurosurgery Clinics of North America*, 22(2):133–139, April 2011. doi:10.1016/j.nec.2010.11.001. URL https://doi.org/10.1016/j.nec.2010.11.001.

[3] Peter A. Bandettini, Eric C. Wong, R. Scott Hinks, Ronald S. Tikofsky, and James S. Hyde. Time course epi of human brain function during task activation. *Magnetic Resonance in Medicine*, 25(2):390–397, June 1992. ISSN 1522-2594. doi:10.1002/mrm.1910250220. URL http://dx.doi.org/10.1002/mrm.1910250220.

[4] S Ogawa, T M Lee, A R Kay, and D W Tank. Brain magnetic resonance imaging with contrast dependent on blood oxygenation. *Proceedings of the National Academy of Sciences*, 87(24):9868–9872, December 1990. ISSN 1091-6490. doi:10.1073/pnas.87.24.9868. URL http://dx.doi.org/10.1073/pnas.87.24.9868.

[5] Denis Le Bihan and Avi Karni. Applications of magnetic resonance imaging to the study of human brain function. *Current Opinion in Neurobiology*, 5(2):231–237, 1995. ISSN 0959-4388. doi:https://doi.org/10.1016/0959-4388(95)80031-X. URL https://www.sciencedirect.com/science/article/pii/095943889580031X.

[6] Nikos K. Logothetis. The underpinnings of the BOLD functional magnetic resonance imaging signal. *The Journal of Neuroscience*, 23(10):3963–3971, May 2003. doi:10.1523/jneurosci.23-10-03963.2003. URL https://doi.org/10.1523/jneurosci.23-10-03963.2003.

[7] A Connelly, G D Jackson, R S Frackowiak, J W Belliveau, F Vargha-Khadem, and D G Gadian. Functional mapping of activated human primary cortex with a clinical mr imaging system. *Radiology*, 188(1):125–130, July 1993. ISSN 1527-1315. doi:10.1148/radiology.188.1.8511285. URL http://dx.doi.org/10.1148/radiology.188.1.8511285.

[8] K K Kwong, J W Belliveau, D A Chesler, I E Goldberg, R M Weisskoff, B P Poncelet, D N Kennedy, B E Hoppel, M S Cohen, and R Turner. Dynamic magnetic resonance imaging of human brain activity during primary sensory stimulation. *Proceedings of the National Academy of Sciences*, 89(12):5675–5679, June 1992. ISSN 1091-6490. doi:10.1073/pnas.89.12.5675. URL http://dx.doi.org/10.1073/pnas.89.12.5675.

[9] S Ogawa, D W Tank, R Menon, J M Ellermann, S G Kim, H Merkle, and K Ugurbil. Intrinsic signal changes accompanying sensory stimulation: functional brain mapping with magnetic resonance

imaging. *Proceedings of the National Academy of Sciences*, 89(13):5951–5955, July 1992. ISSN 1091-6490. doi:10.1073/pnas.89.13.5951. URL http://dx.doi.org/10.1073/pnas.89.13.5951.

[10] Klaus Baudendistel, Lothar R. Schad, Michael Friedlinger, Frederik Wenz, Johannes Schröder, and Walter J. Lorenz. Postprocessing of functional mri data of motor cortex stimulation measured with a standard 1.5 t imager. *Magnetic Resonance Imaging*, 13(5):701–707, 1995. ISSN 0730-725X. doi:https://doi.org/10.1016/0730-725X(95)00016-A. URL https://www.sciencedirect.com/science/article/pii/0730725X9500016A.

[11] Robert W. Cox. Afni: Software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and Biomedical Research*, 29(3):162–173, 1996. ISSN 0010-4809. doi:https://doi.org/10.1006/cbmr.1996.0014. URL https://www.sciencedirect.com/science/article/pii/S0010480996900142.

[12] Ravi S Menon and Seong-Gi Kim. Spatial and temporal limits in cognitive neuroimaging with fmri. *Trends in cognitive sciences*, 3(6):207–216, 1999.

[13] Nikos K Logothetis. What we can do and what we cannot do with fmri. *Nature*, 453(7197):869–878, 2008.

[14] Maxim Sharaev, Alexander Andreev, Alexey Artemov, Alexander Bernstein, Evgeny Burnaev, Ekaterina Kondratyeva, Svetlana Sushchinskaya, and Renat Akzhigitov. fmri: preprocessing, classification and pattern recognition, 2018.

[15] F.E. Roux, J.P. Ranjeva, K. Boulanouar, C. Manelfe, J. Sabatier, M. Tremoulet, and I. Berry. Motor Functional MRI for Presurgical Evaluation of Cerebral Tumors. *Stereotactic and Functional Neurosurgery*, 68(1-4):106–111, 07 1998. ISSN 1011-6125. doi:10.1159/000099910. URL https://doi.org/10.1159/000099910.

[16] K. Papke, T. Hellmann, B. Renger, C. Morgenroth, S. Knecht, G. Schuierer, and P. Reimer. Clinical applications of functional mri at 1.0 t: motor and language studies in healthy subjects and patients. *European Radiology*, 9(2):211–220, 1999. doi:10.1007/s003300050658. URL https://doi.org/10.1007/s003300050658.

[17] Stephen A. Engel, David E. Rumelhart, Brian A. Wandell, Adrian T. Lee, Gary H. Glover, Eduardo-Jose Chichilnisky, and Michael N. Shadlen. fmri of human visual cortex. *Nature*, 369(6481):525–525, 1994. doi:10.1038/369525a0. URL https://doi.org/10.1038/369525a0.

[18] Walter Schneider, B. J. Casey, and Douglas Noll. Functional mri mapping of stimulus rate effects across visual processing stages. *Human Brain Mapping*, 1(2):117–133, January 1994. ISSN 1097-0193. doi:10.1002/hbm.460010205. URL http://dx.doi.org/10.1002/hbm.460010205.

[19] J. R. Binder, S. M. Rao, T. A. Hammeke, F. Z. Yetkin, A. Jesmanowicz, P. A. Bandettini, E. C. Wong, L. D. Estkowski, M. D. Goldstein, V. M. Haughton, and J. S. Hyde. Functional magnetic resonance imaging of human auditory cortex. *Annals of Neurology*, 35(6):662–672, June 1994. ISSN 1531-8249. doi:10.1002/ana.410350606. URL http://dx.doi.org/10.1002/ana.410350606.

[20] S. Dymarkowski, S. Sunaert, S. Van Oostende, P. Van Hecke, G. Wilms, P. Demaerel, B. Nuttin, C. Plets, and G. Marchal. Functional mri of the brain: localisation of eloquent cortex in focal brain lesion therapy. *European Radiology*, 8(9):1573–1580, 1998. doi:10.1007/s003300050589. URL https://doi.org/10.1007/s003300050589.

[21] Jean Decety, Julie Grezes, Nicolas Costes, Daniela Perani, Marc Jeannerod, Emmanuel Procyk, Franco Grassi, and Ferruccio Fazio. Brain activity during observation of actions. influence of action content and subject's strategy. *Brain: a journal of neurology*, 120(10):1763–1777, 1997.

[22] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015.

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

[24] Daniel Anderson, Katherine Fite, Nicole Petrovich, and Joy Hirsch. Cortical activation while watching video montage: An fmri study. *Media Psychology - MEDIA PSYCHOL*, 8:7–24, 02 2006. doi:10.1207/S1532785XMEP0801_2.

[25] Julia Berezutskaya, Mariska J. Vansteensel, Erik J. Aarnoutse, Zachary V. Freudenburg, Giovanni Piantoni, Mariana P. Branco, and Nick F. Ramsey. Open multimodal iEEG-fMRI dataset from naturalistic

stimulation with a short audiovisual film. *Scientific Data*, 9(1), March 2022. doi:10.1038/s41597-022-01173-0. URL https://doi.org/10.1038/s41597-022-01173-0.

[26] Paul Boersma and David Weenink. Praat: doing phonetics by computer [computer program]. version 6.0. 37. *Retrieved February*, 3:2018, 2018.

[27] Julia Berezutskaya, Zachary V. Freudenburg, Luca Ambrogioni, Umut Güçlü, Marcel A. J. van Gerven, and Nick F. Ramsey. Cortical network responses map onto data-driven features that capture visual semantics of movie fragments. *Scientific Reports*, 10(1), July 2020. doi:10.1038/s41598-020-68853-y. URL https://doi.org/10.1038/s41598-020-68853-y.