Convolutional Neural Networks for Real-Time Epileptic Seizure Detection

Felix Achilles[1,2], Federico Tombari[1,3], Vasileios Belagiannis[1,4],
Anna Mira Loesch[2], Soheyl Noachtar[2], Nassir Navab[1,5]

[1] *Technische Universität München, Computer Aided Medical Procedures*;
[2] *Ludwig-Maximilians-University of Munich, Department of Neurology*;
[3] *University of Bologna, DISI*;
[4] *University of Oxford, Department of Engineering Science*;
[5] *Johns-Hopkins-University, Computer Aided Medical Procedures*;

pre-print version (author's copy)
published online: July 13th, 2016

Abstract

Epileptic seizures constitute a serious neurological condition for patients and, if untreated, considerably decrease their quality of life. Early and correct diagnosis by semiological seizure analysis provides the main approach to treat and improve the patients' condition. To obtain reliable and quantifiable information, medical professionals perform seizure detection and subsequent analysis using expensive video-EEG systems in specialized epilepsy monitoring units. However, the detection of seizures, especially under difficult circumstances such as occlusion by the blanket or in the absence of predictive EEG patterns, is highly subjective and should therefore be supported by automated systems. In this work, we conjecture that features learned via a convolutional neural network provide the ability to distinctively detect seizures from video, and even allow our system to generalize to different seizure types. By comparing our method to the state of the art we show the superior performance of learned features for epileptic seizure detection.

1 Introduction

Epileptic seizures are characterized by stereotyped motion patterns. Individual patients show different variations of these motion patterns due to the specific neuroanatomical configurations and localization of the epileptogenic zone (see Noachtar and Peters, 2009). It is therefore complex to derive a general model describing which motions indicate a seizure and which represent normal, non-seizure related behavior. However, reliable detection of epileptic seizures is

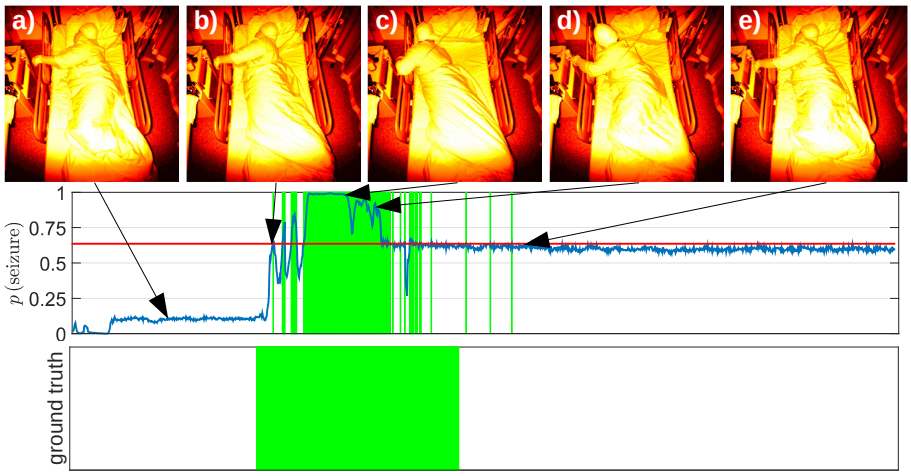


Figure 1: Example of a seizure that is detected with our proposed method. The first row shows frames from the IR stream, colored in red. In frame a), the patient is still asleep. Seizure starts in frame b), then develops into a strong version and uncontrolled head and leg motion in c) and d); end of the seizure in e). The second row shows the raw detector output probability for the *seizure* event (blue line). Frames whose output probability lies above a defined threshold (red line) are counted as *seizure* (highlighted in green). Here, the detection threshold was chosen for maximum sensitivity, while sacrificing specificity.

essential for the process of quantification and the basis for further semiological analysis and diagnosis. This analysis needs to be performed by expert clinicians in epilepsy monitoring units (EMUs), as many seizures cannot be recalled by the patients and go unnoticed by their families or friends. (Blum, Eskola, et al., 1996). A descriptive modeling of seizure motions would hence be a valuable asset for performing automatic classification into ictal phases (during seizures) and interictal phases (in between seizures). Such information can be used by clinicians in order to count seizures by detection of specific patterns. Automated seizure detection can furthermore reduce the time required to review video recordings by highlighting critical events in the sequence. The gold standard for continuous patient monitoring in EMUs are video-EEG systems, which require trained personnel for configuration, maintenance and reviewing of the recordings. Seizure analysis is performed visually and is prone to considerable interobserver variability. (Bleasel, Kotagal, et al., 1997) Furthermore, the staff needs to perform attachment of EEG electrodes to the patient's scalp, which constitutes a complicated and time-intensive procedure. Conditions exist in which no EEG signal can be used to support the detection, e.g. when the seizure is non-epileptic (psychogenic), when the epileptogenic zone is too distant from the recording electrodes or when movement artifacts obscure the EEG. To overcome

these problems and assist the detection procedure, a robust method needs to be developed that is non-invasive and easy to maintain. Such method could possibly even allow an extended automated home monitoring, where EEG systems cannot be easily used. Advances in processing hardware, feature extraction and machine learning methods have made real-time image analysis attractive, even pushing it to near-human and superhuman performance in specific tasks (see Krizhevsky, Sutskever and Hinton, 2012; Szegedy, Liu, et al., 2015). In recent years, it was subsequently attempted to provide methods purely relying on visual information, with the goal of keeping all of the benefits but none of the disadvantages of expensive video-EEG systems.

In this work, we present a novel seizure detection method based on convolutional neural networks, introduced by LeCun, Bottou, et al. (1998). The main intuition behind our approach is that such a network can learn discriminative features from video frames which distinguish normal patient poses and appearances from those characteristic of a seizure. This is conversely to the state of the art, which relies on hand-designed features. Furthermore, existing methods are designed to only detect specific seizure types, while our approach is more general and can be used to detect various types of epileptic seizures from video. We train our CNN in a supervised fashion, by supplying frames obtained from a combined depth and infrared (IR) sensor (see Fig. 1). During testing, the input is processed in real-time. We wish to point out that the employed data modalities have an important practical advantage, since illumination with IR eliminates the need for room lighting during the patient's sleep phase as required by Lu, Pan, et al. (2013). Our method can generally be used in different types of monitoring units, as it is not dependent on room lighting or a special room setting as required for neonate recordings. We will review related work and state of the art in the next section, followed by a detailed explanation of our method, experimental results and concluding remarks.

2 Related Work

Quantification of epileptic seizures through video analysis was first shown by Li, Da Silva and Cunha (2002), who evaluated patient motions during seizures by marker-based tracking of limb movements in 2D video recordings. An extension of this work by O'Dwyer, Cunha, et al. (2007) revealed that after successful seizure detection, quantitative analysis of versive head movements allows for correct lateralization of the epileptogenic zone, thus providing a significant clinical value for subsequent epilepsy surgery.

These advances have spawned several approaches to a full automation of the detection procedure through video analysis. Karayiannis, Xiong, et al. (2006) evaluate motion-strength and motion-trajectory features based on optical flow analysis, which are classified with a single-layer neural network. Key differences with respect to our approach are that the authors provide hand-designed instead of learned features to their classifier as well as the need for computing optical flow. Pisani, Spagnoli, et al. (2014) perform a frequency analysis of the average

luminance in order to detect clonic seizures which are characterized by rhythmical twitching motions. Such an approach however depends on the presence of specific motion frequencies. Conversely, our approach is general and can even detect seizure-related static and slow patient motions arising from e.g. tonic seizures. Another limitation of (Karayiannis, Xiong, et al., 2006; Pisani, Spagnoli, et al., 2014) is their sole evaluation on neonate patient recordings, which are not representative for a general EMU monitoring scenario. Indeed, detecting seizures within recordings of adult or pediatric patients is more complex, as they include various motions that arise from activities unrelated to seizures such as leaving the bed, interacting with the staff or using laptops, books or phones.

Cuppens, Chen, et al. (2012) use the Spatio-Temporal Interest Point detector on recordings of pediatric patients in order to find relevant keypoints inside a spatio-temporal window. At these locations, Histogram-of-Flow features are computed and subsequently classified via SVM. The authors report the dependency on a sufficiently large amount of detected keypoints, a limitation that does not apply to our method as it densely extracts features from every input pixel. Finally, Kalitzin, Petkov, et al. (2012) detect clonic seizures of adult patients. They derive robust motion frequency features from optical flow and compute the relative spectral energy inside a fixed interval of 2 Hz-6 Hz. Their algorithm has same limitation as the one by Pisani et al. as it again requires the presence of specific motion frequencies during the seizure.

3 Method

We use a CNN to model the relation between epilepsy patient recordings as input and the probability of a seizure event as output. Conversely to previous methods, we rely on a single-frame approach. Patients show unnatural postures during clonic, tonic or general convulsive seizures, which can be detected by our method even if no motion is present. In Sec. 4 we show that the detection accuracy improves upon state-of-the-art by a large margin, without leveraging temporal consistency. To create our model, we preprocess the input data (Sec. 3.1), define the network architecture (Sec. 3.2) and finally train the CNN in a supervised fashion (Sec. 3.3).

3.1 Preprocessing

Convolution is a linear operation, which results in the output being inherently sensitive to peak input values. In other terms, feature extraction by convolution will always favor bright image regions. We give an equal *a priori* weight to both input modalities by normalizing their domains. The depth sensor provides a video stream I_{D} whose pixel values represent distances between 500 mm and 4500 mm at 1 bit/mm resolution. In the IR stream I_{IR} however, values are not limited and extend to the full 16-bit range [0-65525]. In order to decrease the high dynamic range, I_{IR} is preprocessed with a natural logarithm, resulting in

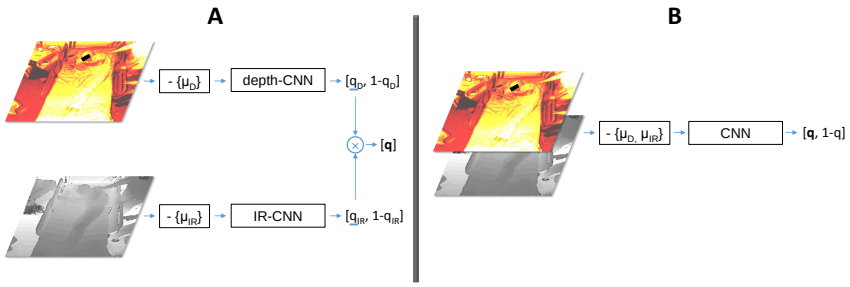


Figure 2: Combining depth and IR input. (A): Depth and IR are processed with individual CNNs only trained on the respective modality. Both CNNs follow the same architecture, detailed in Fig. 3. Final seizure probability q is obtained by multiplying q_{D} from the depth-CNN and q_{IR} from the IR-CNN. (B): The CNN is trained on a 2-channel input built by combining a depth-frame and an IR-frame. Network architecture is the same as in type A but processes both modalities in its first layer of filters.

$\hat{I}_{\mathrm{IR}} = \log(1 + I_{\mathrm{IR}})$. Spurious noise in the depth stream I_{D} (e.g. at reflective surfaces or occluding edges) is removed by applying a $[3 \times 3]$ median filter, obtaining $\hat{I}_{\mathrm{D}}$. Finally, the intensity values of $\hat{I}_{\mathrm{D}}$ and $\hat{I}_{\mathrm{IR}}$ are normalized to the range [0-255].

3.2 Network Architecture

The network architecture that is best suited to the task at hand is depicted in Fig. 3. Input to the network are IR and depth frames $\{\hat{I}_{\mathrm{IR}}, \hat{I}_{\mathrm{D}}\}$. Depending on the combination scheme that is used, they are either processed individually (type A) or stacked to build up one single frame with two channels (type B), see Fig. 2. The input frames originally have a resolution of $512\,\mathrm{px} \times 424\,\mathrm{px}$. For training and testing, they are center-cropped to $424\,\mathrm{px} \times 424\,\mathrm{px}$ and downscaled to $100\,\mathrm{px} \times 100\,\mathrm{px}$. We compute an average frame for IR and depth respectively by summing over all available samples in the training set and dividing by the number of samples.

After subtracting the average frame $\{\mu_{\mathrm{IR}}^{\mathrm{train}}, \mu_{\mathrm{D}}^{\mathrm{train}}\}$ for each modality, the input is transformed into a feature map through the first computational block consisting of 1) a convolutional layer with stride 1, 2) a rectified linear unit (ReLU), 3) a max pooling layer with stride 12 and 4) a local response normalization layer as the one used by Krizhevsky, Sutskever and Hinton (2012). In the second computational block, two subsequent fully connected layers with ReLU activation units reduce the size of the feature map so to build up a two-element vector. This vector is normalized into $[0, 1]$ by a softmax operation, see equation (1). Finally, the CNN output is $[q,\ 1-q]$, where q represents the probability for the seizure event to be *true* and reciprocally $1-q$ describes the

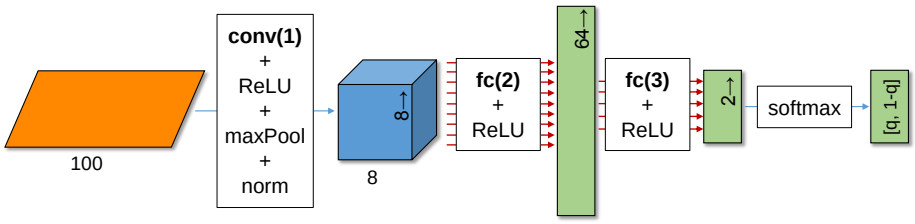


Figure 3: Network architecture. Layer conv(1) extracts local features through convolution with eight $[5 \times 5]$ filter kernels. Max pooling with large $[12 \times 12]$ receptive fields downscales the input to a $8\,\text{px} \times 8\,\text{px}$ feature map. Fully connected layers fc(2-3) condense the feature map into a binary output, yielding the seizure probability q.

probability for the event to be *false*.

3.3 Training

In order to learn the connection weights and biases for the convolutional and the fully connected layers, we append a *logarithmical softmax loss* layer to the network output:

$$L(x, y) = -\frac{1}{n} \sum_i \log \left(\frac{e^{x_{i,y(i)}}}{\sum_c e^{x_{i,c}}} \right) \tag{1}$$

where x is the last feature map of the network, y contains the ground truth class for each output variable and n is the number of output variables. Through the loss layer, decisions for the wrong binary class are penalized and the derivative with respect to the output x is calculated. This derivative is back-propagated through the network, using stochastic gradient descent (SGD) for loss minimization. For all iterations, a batch-size of 250 samples and a learning-rate of $10^{-3.25}$ are chosen. The rest of the parameters is set according to Krizhevsky, Sutskever and Hinton (2012). In particular, the momentum is set to 0.9 and weight decay for L_2 regularisation is $5 \cdot 10^{-4}$. During training, we apply a dropout of 0.5 before fc(2) and fc(3) to reduce overfitting.

We augment the training data in order to increase the generalization properties of our model. More specifically, we horizontally flip half of the images in each batch, and randomly shuffle the training set before each SGD epoch. Using real patient recordings for training, we cannot guarantee a 50/50 ratio between positive and negative training samples: indeed, the problem at hand is intrinsically unbalanced, since negative frame samples outnumber the positive ones on typical recordings. In order to achieve a balanced training set, one could select the class with fewer available samples and select an equal amount of samples from the over-represented class. This however results in discarding a lot of valuable samples, hence distorting the input distribution which may lead to a

loss of generality. In contrast, we determine the class $c_- \in \mathbf{C} = \{c_{\text{true}}, c_{\text{false}}\}$ which has fewer available training samples and randomly pick the same amount of samples from class $c_+ = \mathbf{C} \setminus c_-$ before each epoch. After one epoch is trained using SDG, a different set of samples is randomly picked from c_+ and combined with the samples from c_-. This way, the network is trained along gradients from positive and negative samples at an equal rate while at the same time using all of the available training data.

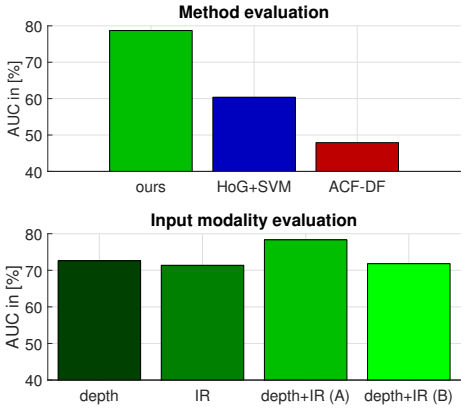
4 Experiments

4.1 Data Acquisition

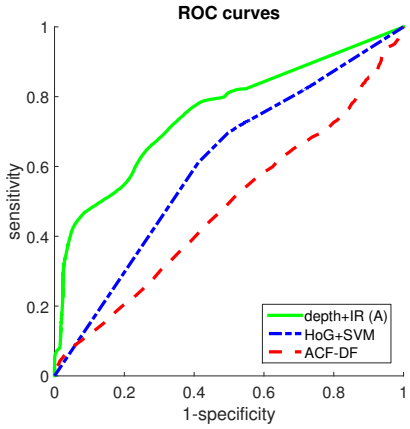
The data used for our experiments was acquired from adult patients admitted to an EMU for advanced epilepsy diagnostics. During the stay they were monitored via video-EEG and an additional Kinect v2 consumer depth camera that was installed at the foot-end of the patient bed, similar to Cunha, Paula, et al. (2012). From the video-EEG recordings, medical professionals determined beginning and end of each seizure. Sequences were selected in which the patient presented at least one clinical sign, regardless if it was subtle, non-rhythmic, occluded by the blanket or by clinical staff. In total, we acquired 52 sequences recorded from 10 patients, including clonic, tonic and automotor seizures. Each sequence on average lasts 1:46 min at a 15 FPS frame rate. The entire database consists of $82,666$ frames.

4.2 Evaluation

We build up a training and test split of our labeled data by assigning half of the patients as training and the other half as testing set. This split is used for cross-validation, so that we can evaluate the average performance in detecting seizures from unseen patients (*cross-subject* performance). The method of Pisani, Spagnoli, et al. (2014) was implemented selecting a threshold of 20 to binarize temporal difference frames. Summing over the binarized frame i yields $\bar{L}[i]$, which the authors name the *average motion signal*. Rhythmical motion is detected in $\bar{L}[i]$ by combining its auto-correlation function and its difference function, hence we refer to the method as ACF-DF. In order to apply ACF-DF on depth and IR data, we independently extract $\bar{L}[i]$ and its correlation signal on both modalities, apply the decision threshold and combine both decisions with a logical *OR*, which yielded the best results. As baseline comparison to a learning method that relies on hand-crafted features, we extract HoG features with 2×2 cells, 8×8 block size and 9 gradient bins from each frame. Feature vectors of corresponding depth and IR frames are concatenated and the resulting 8,712-element descriptor is classified via linear SVM. Hyperparameter optimization of the SVM is performed as cross-validated grid search, yielding an optimal C parameter of $10^{0.2}$. During testing, SVM classification scores are transformed into posterior probabilities using a sigmoid function.



(a) Upper chart: Comparison, in terms of AUC, of our proposed method with HoG+SVM and ACF-DF by Pisani, Spagnoli, et al. (2014). Lower chart: Evaluation of our method using depth only, IR only and depth+IR with network types A and B.



(b) ROC curve of the proposed method depth+IR (A) (green), compared with the curves obtained by HoG+SVM (blue) and ACF-DF (Pisani, Spagnoli, et al., 2014) (red).

Figure 4: Evaluation of our method in terms of *cross-subject* performance. Comparison of AUC values in (a), ROC curves in (b).

For each experiment, we have evaluated the performance of our seizure detector by sweeping the decision threshold over the range of $[0, 1]$ and computing, at each value, the True Positive Rate and False Positive Rate, so to build up the Receiver Operating Characteristic (ROC) curve. As a figure of merit, we compute the Area Under the Curve (AUC).

The upper chart of Fig. 4(a) reports the AUC, based on using combined depth+IR input, for *our* best method (CNN with network type A) versus *HoG+SVM* and *ACF-DF*. The best CNN architecture depth+IR (A) achieves an AUC of 78.33%, yielding an absolute improvement of 17.94% over HoG+SVM and 30.43% over ACF-DF. Interestingly, we can here witness how the proposed CNN-learned features outperform handcrafted generic features such as HoG. At the same time, results show that the ACF-DF method hardly generalizes to different types of seizures. More specifically, it was designed to detect clonic seizures, on the subset of which it achieves an AUC of 60.07%, similar to HoG+SVM. To complement previous results, we show in Fig. 4(b) the average ROC curves of the three methods. Additionally, in the lower chart of Fig. 4(a), we compare four variants of our CNN, using as input either *depth only*, *IR only* or both *depth+IR* with late (type A) or early (type B) modality combination. While in the early fusion setup only one network is required to generate seizure probability q, the late fusion uses one network for each modality and generates the final probability by multiplying the respective network outputs.

The late combination approach *depth+IR (A)* yields an improvement of

5.72% over *depth only*, an improvement of 6.96% over *IR only* and an improvement of 6.53% over the early combination approach *depth+IR (B)*. Best results are obtained through a straightforward fusion of the infrared-CNN and the depth-CNN responses by multiplication of their respective output. In order to evaluate if this way of fusing the modalities discards too much fine-grained detail about the scene, we trained an SVM classifier on the concatenated 64-element vectors taken from *fc(2)* of both networks. This way, an average AUC of 74.34% was achieved, not improving on the 78.33% that were reached by multiplication of CNN outputs.

4.3 Performance

Finally, we report that the algorithm runs at 10 ms per frame (i.e., 100 frames per second) on a desktop PC equipped with a GeForce GTX 660 graphics card using the MatConvNet library (Vedaldi and Lenc, 2014).

5 Conclusion

We have presented a novel seizure detection algorithm that builds on convolutional neural networks and that, compared to the state of the art, is able to achieve superior results on a competitive dataset including different types of epileptic seizures. It was shown that training individual CNNs for each input modality results in higher accuracy than an early combination of modalities. This could be attributed to the very different statistics of depth and IR data and hints at the conclusion that combining geometry and texture information in a single CNN is not a trivial task. We note that the presented network is shallow in comparison to popular image classification networks, but still outperforms state-of-the-art methods for epilepsy detection from video yielding an AUC of 78.33%. At the same time the network is fast at test time, such that we achieve real-time performance. Furthermore, the lightweight design allows for the use of standard processing hardware or even mobile devices. Our CNN for epilepsy detection can greatly benefit patients and neurologists, as it facilitates the reviewing process of large video databases and is able to give real-time feedback on the patient status without the use of invasive monitoring equipment.

Acknowledgments

The authors would like to thank Professor João Paulo Cunha as well as Christian Vollmar for fruitful discussions and continuous support.

Funding

This work has been funded by the German Research Foundation (DFG) through grants NA 620/23-1 and NO 419/2-1.

References

Bleasel A, Kotagal P, Kankirawatana P, Rybicki L. 1997. Lateralizing value and semiology of ictal limb posturing and version in temporal lobe and extratemporal epilepsy. Epilepsia. 38(2):168–174.

Blum D, Eskola J, Bortz J, Fisher R. 1996. Patient awareness of seizures. Neurology. 47(1):260–264.

Cunha J, Paula LM, Bento VF, Bilgin C, Dias E, Noachtar S. 2012. Movement quantification in epileptic seizures: a feasibility study for a new 3D approach. Medical engineering & physics. 34(7):938–45.

Cuppens K, Chen CW, Wong KY, Van de Vel A, Lagae L, Ceulemans B, Tuytelaars T, Van Huffel S, Vanrumste B, Aghajan H. 2012. Using spatio-temporal interest points (stip) for myoclonic jerk detection in nocturnal video. In: Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE. IEEE; p. 4454–4457.

Kalitzin S, Petkov G, Velis D, Vledder B, Lopes da Silva F. 2012. Automatic segmentation of episodes containing epileptic clonic seizures in video sequences. Biomedical Engineering, IEEE Transactions on. 59(12):3379–3385.

Karayiannis NB, Xiong Y, Tao G, Frost JD, Wise MS, Hrachovy RA, Mizrahi EM. 2006. Automated detection of videotaped neonatal seizures of epileptic origin. Epilepsia. 47(6):966–980.

Krizhevsky A, Sutskever I, Hinton GE. 2012. Imagenet classification with deep convolutional neural networks. In: NIPS.

LeCun Y, Bottou L, Bengio Y, Haffner P. 1998. Gradient-based learning applied to document recognition. Proceedings of the IEEE. 86(11):2278–2324.

Li Z, Da Silva AM, Cunha JPS. 2002. Movement quantification in epileptic seizures: a new approach to video-eeg analysis. Biomedical Engineering, IEEE Transactions on. 49(6):565–573.

Lu H, Pan Y, Mandal B, Eng HL, Guan C, Chan DW. 2013. Quantifying limb movements in epileptic seizures through color-based video analysis. Biomedical Engineering, IEEE Transactions on. 60(2):461–469.

Noachtar S, Peters AS. 2009. Semiology of epileptic seizures: a critical review. Epilepsy and Behavior. 15:2–9.

O'Dwyer R, Cunha JP, Vollmar C, Mauerer C, Feddersen B, Burgess RC, Ebner A, Noachtar S. 2007. Lateralizing significance of quantitative analysis of head movements before secondary generalization of seizures of patients with temporal lobe epilepsy. Epilepsia. 48(3):524–530.

Pisani F, Spagnoli C, Pavlidis E, Facini C, Ntonfo GMK, Ferrari G, Raheli R. 2014. Real-time automated detection of clonic seizures in newborns. Clinical Neurophysiology. 125(8):1533–1540.

Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A. 2015. Going deeper with convolutions. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR.

Vedaldi A, Lenc K. 2014. Matconvnet – convolutional neural networks for matlab. CoRR. abs/1412.4564.