Cataplexy Detection: Neurologists, You Are Not Alone!

(Discussion Paper)

Ilaria Bartolini[1], Andrea Di Luzio[1]

[1]Department of Computer Science and Engineering (DISI), Alma Mater Studiorum, University of Bologna, Italy

Abstract

Narcolepsy with cataplexy is a severe lifelong disorder characterized, among the others, by the sudden loss of bilateral face muscle tone triggered by emotions (cataplexy). In this extended abstract, we present two methodologies for the automatic analysis of patients' videos able to assist neurologists in diagnosing the disease and/or detecting attacks. Indeed, recent findings demonstrated that the detection of abnormal motor behaviors in video recordings of patients undergoing emotional stimulation is effective in characterizing the disease symptoms. Such motor behaviors (ptosis, mouth opening, head drop) are however to be discovered by neurologists through manual inspection of patients' videos. Automatic content-based video analysis is clearly of immediate help here. Experimental results conducted on real data support the effectiveness of the presented automated techniques.

Keywords

Video-based classification of cataplexy, automatic video content analysis, motor behavior patterns, data analysis for health

1. Introduction

Narcolepsy with cataplexy is a rare disorder mainly arising in young adults/children characterized by daytime sleepiness, sudden loss of muscle tone while awake triggered by emotional stimuli (cataplexy), hallucinations, sleep paralysis, and disturbed nocturnal sleep [13]. A recent approach for the detection of the disease is based on an analysis of video recordings of patients undergoing emotional stimulation made *on-site* by medical specialists [16]. According to this methodology, cataplexy is present if any of three abnormal motor behaviors is detected in the patient video: *ptosis* (a drooping or falling of the upper eyelid), *head drop*, and *smile/mouth opening* [13]. Such patterns are, however, still to be manually detected by neurologists through visual inspection of videos. This is due to the complete absence of automatic technological solutions able to properly support neurologists in such a delicate task.

It is evident that a tool able to detect the "correct" facial expression changes (i.e., the disease symptoms) from video recordings of patients would be able to automatically identify the presence of the disease. This could be extremely helpful, not only to support neurologists in diagnosing the disease, but also in monitoring everyday activities in a non-invasive way to

SEBD 2021: The 29th Italian Symposium on Advanced Database Systems, September 5-9, 2021, Pizzo Calabro (VV), Italy

ilaria.bartolini@unibo.it (I. Bartolini); andrea.diluzio2@unibo.it (A. Di Luzio)

provide early warnings in the event of the insurgence of a crisis. Indeed, it is well known that the synergistic use of Machine Learning (ML) techniques can help in alleviating the burden of the medical specialist in analyzing patient data, thus improving diagnostic consistency and accuracy [10].

In [2], we introduced the CAT-CAD (Computer Aided Diagnosis for CATaplexy) tool, which exploits ML techniques for the automatic analysis of video recordings made on patients undergoing emotional stimulation through the vision of funny movies designed to evoke the laughter. By means of a user friendly GUI, CAT-CAD effectively supports neurologists with (1) the automatic detection of disease symptoms, and thus in the disease recognition/monitoring, and (2) advanced functionalities for video playback and browsing/retrieval. CAT-CAD is the first tool to allow the automatic recognition of cataplexy symptoms based on the analysis of patients' video recordings.

In this extended abstract, we report details on the video analyzer for the automatic detection of cataplexy symptoms. This component of the CAT-CAD system is built on top of SHIATSU, a general and extensible framework for video retrieval which is based on the (semi-)automatic hierarchical semantic annotation of videos exploiting the analysis of their visual content [3], and exploits features managed through the Windsurf software library [4].

After reviewing some background information, we detail the methodologies used to automatically analyze videos: a pattern-based technique able to recognize facial patterns and a novel approach based on convolutional neural networks (Section 2). Finally, we provide results obtained from an extensive experimental evaluation to compare the performance of the two video analysis approaches, using a benchmark containing recordings from real patients (Section 3) and conclude (Section 4).

1.1. Background and Related Work

Narcolepsy with cataplexy usually arises in adolescence or young adulthood, but the diagnosis is typically established after a long period with a mean delay (across Europe) from symptom onset to diagnosis of 14 years [12]. The diagnosis delay is due not only to the failure to recognize the symptoms of the disease, but also to the misinterpretation of cataplexy phenomena as the expression of other disorders, such as episodes of loss of consciousness of epileptic nature, force reductions due to neuromuscular disorder or behavioral disorders of childhood psychiatric or neuropsychiatric relevance.

Few scientific studies have considered the video-polygraphic features of cataplexy in adult age and only recently the motor phenotype of childhood cataplexy has been described exclusively using video recordings of the attacks evoked by watching funny cartoons [13]. These studies showed that in the context of the physiological response to the laughter there are the distinctive elements of cataplexy, called *motor behaviors patterns*, particularly evident at the level of the *facial expression changes*. In particular, the three most recurrent motor phenomena (often displayed by patients affected by the disease) are ptosis, head drop, and smile/mouth opening [13].

To the best of our knowledge, CAT-CAD is the first study about the automatic recognition of cataplexy by exploiting patient video recordings. However, automatic detection of facial motor phenomena similar to the ones used to diagnose cataplexy is commonly used in other

contexts. For example, the detection of eyelid closure, head pose, or mouth opening is useful for the automatic recognition of fatigue/drowsiness in vehicle drivers [11, 9, 6]. The "verbatim" use of such techniques in the context of cataplexy diagnosis is however inappropriate, since the peculiar motor patterns are somewhat different, even if they can be detected using similar facial features.

2. The CAT-CAD Video Analyzer

The core of the CAT-CAD tool is the real-time analysis of patients' videos to detect the presence of disease symptoms (i.e., ptosis, head drop, and smile/mouth opening). Two different approaches were developed to perform video analysis, with the idea of comparing their relative performance and possibly combining them to achieve the best possible result in recognizing the different motor phenomena:

- The *Pattern-Based* approach (Section 2.1) is built on the automatic detection of facial features in video frames.
- The *Deep Learning* approach (Section 2.2) exploits three convolutional neural networks, each trained to detect a specific motor phenomenon.

2.1. Video Analyzer: Pattern-Based Approach

The first video analyzer to be implemented in CAT-CAD for the detection of cataplexy motor phenomena exploits facial landmarks, as detected by OpenFace [1]. The first step of the pattern characterization process consists in detecting and extracting patients' facial landmarks of interest from each video frame; this is necessary because it is safe to assume that different patients have different facial features. Using OpenFace on each single video frame we are thus able to extract, for each video, a time series of multi-dimensional feature vectors, each vector characterizing the facial landmarks extracted from a single frame.

2.1.1. Ptosis

Ptosis is a drooping or falling of the upper eyelid. This, however, should not be mistaken as a (regular) eye blink. For this, ptosis is detected whenever eyes are closed for a period of time longer than a typical blink. For each frame, a 12-dimensional feature vector is extracted, containing the (x, y) coordinates of six landmarks characterizing the shape of the eye. The *Eye Aspect Ratio* (EAR) can then be defined as the ratio of the eye *height* to the eye *width* (averaged for left and right eye). EAR is partially invariant to head pose and fully invariant to uniform image scaling and in-place face rotation. The semantics of EAR are as follows: when an eye is closing, EAR approaches zero, whereas when the eye is completely open, EAR attains its maximum value (which varies from person to person). Therefore, we define the presence or absence of ptosis by measuring the length of the time series corresponding to a "long enough" sequence of frames with closed eyes (EAR lower than a threshold).

2.1.2. Head Drop

For head drop, a 8-D feature vector is extracted from each frame, including the (x, y) coordinates of the two landmarks characterizing the external corner of each eye and the one that is immediately below the tip of the nose and the rotation of the head around X and Z axes. The Center of Gravity (CoG) of the three landmarks is then used to measure rotation around the Y axis. Head drop is then detected if rotation around one of the three axes exceeds a threshold.

2.1.3. Smile/Mouth Opening

For the third motor phenomenon, a 8-D feature vector is extracted for each frame, containing (x, y) coordinates of four landmarks characterizing the shape of the mouth. The *Mouth Aspect Ratio* (MAR) can then be defined as the ratio of the mouth *width* to the mouth *height*. Like EAR, MAR is partially invariant to head pose and fully invariant to uniform image scaling and in-place face rotation. When the mouth is closed, MAR attains its maximum value (which varies from person to person), while if the mouth is completely open, MAR reaches its lowest value; intermediate values characterize various types of smile. We thus consider the cataplectic mouth opening as present if the current MAR is lower than a threshold, indicating that the patient is smiling widely or opening her mouth.

2.2. Video Analyzer: Deep Learning Approach

The alternative video analysis tool is based on convolutional neural networks (CNNs). The CNN architecture used in this work is based on the DeXpression Network [7], which achieves excellent performance in expression recognition, and has been implemented using TensorFlow (https://www.tensorflow.org/).

Our CNN architecture consists of three different types of blocks:

1. an *Input Block*, which performs image pre-processing,
2. a *Feature Extraction Block*, inspired by the architectural principles introduced by GoogleNet [15], which is repeated four times, and
3. an *Output Block*, which is used to produce the result class from the features extracted by previous layers.

We have trained three different networks, one for each motor phenomenon to be recognized. The three networks share the same architecture, but the learned weights are clearly different, due to the use of different training classes. It is clear that, for this approach, each frame is analyzed per se, and no considerations on frame sequences, like duration of eyelids closing or of head drop, can be extrapolated, contrary to the pattern-based approach. The three neural networks have been trained for 8 epochs each, for a total time of about 12 hours.

For each video frame, the face of the patient is first detected by means of OpenFace and then cropped. The resulting image is converted to grayscale and downsized to produce images of 320×320 pixels. Cropping of the images was necessary in order to provide the CNN with only face details (thus avoiding that the surrounding environment would distract the learning).

3. Experimental Evaluation

The benchmark used for our experimental evaluation consists of a population of patients of the Outpatient Clinic for Narcolepsy of the University of Bologna, who were assessed for the presence of cataplexy by way of a neurophysiological and biological diagnosis [16].

The first (experimental) group of patients includes 14 subjects displaying symptoms of the disease. Training of video analyzers has been performed using an inter-patient separation scheme, where patients have been randomly assigned to non-overlapping training and test sets, by respecting sex and age distribution. In particular, 11 patients have been included in the training set (thus, their entire videos have been used to train each analyzer), while the remaining 3 patients have been exploited to test the accuracy of the tool.

The second group includes 44 different subjects that show no sign of the disease. Among those, 14 patients have been selected as a control group so as to follow the same sex and age distribution of the experimental group.

For the deep-learning approach, data augmentation was performed by adding, to each training set frame, seven additional images by performing: (i) 3 *rotations* with a random angle between -45° and $+45^\circ$, (ii) 3 *translations* with a random shift between -50 and 50 pixels, and (iii) 1 *horizontal flipping*. The final training sets consists of 191140 labeled images for ptosis, 61216 labeled images for head-drop, and 108196 labeled images for mouth opening.

3.1. Performance Measures

To objectively evaluate the performance of our analyzers, each frame can be labeled according to a confusion matrix as correctly and incorrectly recognized for each of the two classes available (in our case, motor phenomenon actually present or absent). The four possible outcomes are tp (true positive, a frame where the symptom is correctly detected as present), fn (false negative, symptom wrongly not detected), tn (true negative, symptom correctly not detected) and fp (false positive, symptom wrongly detected as present). From the confusion matrix, the performance measures used in our experiments are defined as follows.

Recall/Sensitivity (R) is defined as the fraction of the frames showing signs of the disease (positives) that are correctly identified: $R = \frac{tp}{tp+fn}$. R is therefore used to measure the accuracy of a technique in recognizing the presence of the disease.

Specificity (S) is the fraction of frames not showing the disease (negatives) that are correctly classified: $S = \frac{tn}{tn+fp}$. S thus expresses the ability of a technique to avoid false alarms (which can lead to expensive/invasive exams).

Precision (P) is another popular metric, besides R and S, which are the fundamental prevalence-independent statistics. P is defined as the fraction of correct positively classified frames and assesses the predictive power of the classifier: $P = \frac{tp}{tp+fp}$.

Accuracy (A) measures the fraction of correct decisions, to assess the overall effectiveness of the algorithm: $A = \frac{tp+tn}{tp+fp+fn+tn}$.

Balanced Score (F_1) is a commonly used measure, combining P and R in a single metric computed as their harmonic mean: $F_1 = 2\frac{1}{\frac{1}{P}+\frac{1}{R}} = \frac{2 \cdot tp}{2 \cdot tp + fp + fn}$.

For the pattern-based approach, thresholds used for the detection of ptosis, head drop, and mouth opening were chosen as the ones providing the best classifying performance on the test set [8]. To this end, a Receiver Operating Characteristics (ROC) graph is used for each threshold, and the threshold value maximizing the harmonic mean of R and S measures is chosen as the optimal one: this represents a metric for imbalanced classification, seeking an equilibrium between the two measures [14].

3.2. Overall Performance

Tables 1 and 2 show the performance of the proposed classification techniques over cataplectic/non-cataplectic patients, respectively: figures values were obtained by averaging individual values weighted by the recordings' length. Tables report classification results for the detection of the three symptoms as well for the overall cataplexy which, we remind, is recognized as present if any of the three motor pattern is detected for a particular frame. To compare the performance of the two alternative video analyzers, results in tables show in boldface the best value obtained in any of the five considered performance measures (specificity only for the control group).

	pattern-based					deep learning				
motor phenomenon	R	S	P	A	F_1	R	S	P	A	F_1
ptosis	**0.72**	**0.84**	**0.82**	**0.78**	**0.77**	0.71	0.67	0.68	0.69	0.70
mouth opening	**0.78**	0.76	0.75	**0.76**	**0.77**	0.72	**0.81**	**0.79**	0.76	0.75
head drop	0.60	**0.94**	**0.89**	**0.77**	**0.72**	**0.67**	0.81	0.78	0.74	0.72
overall	**0.75**	**0.79**	**0.79**	**0.77**	**0.77**	0.70	0.74	0.73	0.72	0.71

Table 1
Performance of the proposed approaches for cataplectic patients.

	pattern-based	deep learning
motor phenomenon	S	S
ptosis	**0.99**	0.83
mouth opening	**0.98**	0.83
head drop	**0.99**	0.81
overall	**0.98**	0.66

Table 2
Specificity of the proposed approaches for non-cataplectic subjects.

Above results led us to draw the following considerations:

- The pattern-based approach lead to significantly superior results with respect to its deep learning counterpart. In particular, for cataplectic subjects the former attains the best performance in 85% of the metrics (17 out of $4 \times 5 = 20$ performance measures).

- When considering specific motor phenomena, the pattern-based approach consistently outperforms the deep learning approach in detecting ptosis, while the latter sports superior measures only for specificity and precision in detecting mouth opening and for recall in head drop detection.
- The superior specificity of the pattern-based technique is confirmed in non-cataplectic subjects, with an overall specificity at 98%.

A possible explanation for the inferior performance of the deep learning approach is the fact that such approach cannot discriminate between quick and long eye blinks/head drops, due to the fact that each frame is analyzed individually by the CNN. It is therefore likely that the higher number of false positives is because the CNN wrongly detects "regular" eye blinks or head movements as ptosis or head drop.

For the case of non-cataplectic subjects, it is interesting to note that the performance of the deep learning approach for the overall detection of cataplexy is sensibly worse than those attained for the single motor phenomena. For such patients, false positives for ptosis, head drop, and mouth opening are present in different frames. Indeed, due to the absence of positive cases, the set of false positive frames for the overall cataplexy coincides with the union of frames wrongly classified by any specific motor phenomenon detector.

Finally, we include a brief discussion about efficiency of the proposed techniques. On our experimental setup, which involved a commodity (low-end) machine, we were able to extract EAR, CoG and MAR descriptors in real-time for each video frame. Clearly, this is the more time consuming operation for the pattern-based approach, thus it is proven that the whole process of automatic detection can be performed on-line during a single emotional stimulated video recording session. On the other hand, our current implementation of the deep learning approach is only capable to obtain a throughput of 18.5 frame/s, thus being unable to attain real-time performance (recall that the frame rate of videos is 30 frame/s). The reason for this measure is the following: when analyzing a single frame, about 50% of the time is spent in detecting the position of the patient face, about 25% in cropping the image (retaining only the face), and 25% for the classification of the frame by the three neural networks. The bottleneck of the whole computation is clearly the face detection phase, which we implemented using the OpenFace library, instead of using other faster methods (such as the well-known Haar-Cascade filter). This choice was carried out starting from the consideration that quicker filters often fail to identify the face within the image, especially in videos with excessive head movement, which is the common case for cataplectic subjects.

4. Conclusions

In this extended abstract, we reported details on the video analyzer of CAT-CAD for the automatic detection of cataplexy symptoms (ptosis, head drop, and smile/mouth opening). Two different approaches are introduced for the detection of disease symptoms: the Pattern-Based approach is based on analysis of facial features, using the OpenFace framework, while the Deep Learning approach uses CNNs, as implemented by TensorFlow. An extensive comparative experimental evaluation conducted on a benchmark of real patients recordings demonstrated the accuracy of the proposed techniques. When comparing the effectiveness of the two video

analyzers we introduced to detect cataplexy symptoms, the pattern-based approach achieves superior performance. One of the possible explanations for the inferior detection accuracy of the deep learning approach is the fact that 2D CNNs are unable to properly take into account the temporal dimension that correlates subsequent frames in a video. The use of 3D CNNs could be an interesting way to pursue, and we plan to consider their inclusion in CAT-CAD.

References

[1] T. Baltrušaitis, A. Zadeh, L. Yao Chong, L.-P.; Morency, OpenFace 2.0: Facial Behavior Analysis Toolkit, in: Proceedings of FG 2018, Xi'an, China, May 2018.

[2] I. Bartolini, A. Di Luzio, CAT-CAD: A Computer-Aided Diagnosis Tool for Cataplexy, Computers, 2021, 10(4).

[3] I. Bartolini, M. Patella, C. Romani, SHIATSU: Tagging and Retrieving Videos Without Worries, Multimedia Tools and Applications, 2013, 63(2).

[4] I. Bartolini, M. Patella, G. Stromei, The Windsurf Library for the Efficient Retrieval of Multimedia Hierarchical Data, in: Proceedings of SIGMAP 2011, Seville, Spain, July 2011.

[5] I. Bartolini et al, Automatic Detection of Cataplexy, Sleep Medicine, 2018, 52.

[6] L.M. Bergasa et al, Analysing Driver's Attention Level Using Computer Vision, in: Proceedings of ITSC 2008, Beijing, China, June 2008.

[7] P. Burkert et al, DeXpression: Deep Convolutional Neural Network for Expression Recognition, arXiv, September 2015.

[8] T. Fawcett, An Introduction to ROC Analysis, Pattern Recognition Letters, 2006, 27(8).

[9] J. Jo, H.G. Jung, K. Ryoung, J. Kim, Vision-Based Method for Detecting Driver Drowsiness and Distraction in Driver Monitoring System, Optical Engineering, 2011, 50(12).

[10] L. Lazli, M. Boukadoum, O.A. Mohamed, A Survey on Computer-Aided Diagnosis of Brain Disorders through MRI Based on Machine Learning and Data Mining Methodologies with an Emphasis on Alzheimer Disease Diagnosis and the Contribution of the Multimodal Fusion, Applied Sciences, 2020, 10.

[11] B. Mandal, L. Li, G.S. Wang, J. Lin, Towards Detection of Bus Driver Fatigue Based on Robust Visual Analysis of Eye State, IEEE Transactions on Intelligent Transportation Systems, 2017, 18(3).

[12] F. Pizza et al, Clinical and Polysomnographic Course of Childhood Narcolepsy with Cataplexy, Brain, 2013, 136(12).

[13] G. Plazzi et al, Complex Movement Disorders at Disease Onset in Childhood Narcolepsy with Cataplexy, Brain: A Journal of Neurology, 2011, 134(12).

[14] F. Provost, Machine Learning from Imbalanced Data Sets 101, in: Proceedings of the AAAI'2000 Workshop on Imbalanced Data Sets, 68(2000), Austin, TX, July 2000.

[15] C. Szegedy et al, Going Deeper with Convolutions, in: Proceedings of CVPR 2015, Boston, MA, June 2015.

[16] S. Vandi et al, A Standardized Test to Document Cataplexy, Sleep Medicine, 2019, 53.