Comparing methods for assessment of facial dynamics in patients with major neurocognitive disorders

Yaohui Wang[1], Antitza Dantcheva[1,3], Jean-Claude Broutart[2], Philippe Robert[3], Francois Bremond[1,3], and Piotr Bilinski[4]

[1] INRIA Sophia-Antipolis, STARS, France
{yaohui.wang, antitza.dantcheva, francois.bremond}@inria.fr
[2] GSF Noisiez, France
jc.broutart@free.fr
[3] EA CoBTeK-University Cote d'Azur, France
probert@unice.fr
[4] University of Oxford, UK
piotr.bilinski@eng.ox.ac.uk

Abstract. Assessing facial dynamics in patients with major neurocognitive disorders and specifically with Alzheimers disease (AD) has shown to be highly challenging. Classically such assessment is performed by clinical staff, evaluating verbal and non-verbal language of AD-patients, since they have lost a substantial amount of their cognitive capacity, and hence communication ability. In addition, patients need to communicate important messages, such as discomfort or pain. Automated methods would support the current healthcare system by allowing for telemedicine, *i.e.,* lesser costly and logistically inconvenient examination. In this work we compare methods for assessing facial dynamics such as talking, singing, neutral and smiling in AD-patients, captured during music mnemotherapy sessions. Specifically, we compare 3D ConvNets, Very Deep Neural Network based Two-Stream ConvNets, as well as Improved Dense Trajectories. We have adapted these methods from prominent action recognition methods and our promising results suggest that the methods generalize well to the context of facial dynamics. The Two-Stream ConvNets in combination with ResNet-152 obtains the best performance on our dataset, capturing well even minor facial dynamics and has thus sparked high interest in the medical community.

Keywords: Facial Dynamics · Facial Expressions · Neurocognitive Disorders · Alzheimer's Disease

1 Introduction

Major neurocognitive disorder (NCD), as introduced by the American Psychiatric Association (APA), known previously as dementia, is a decline in mental ability, threatening the independence of a large fraction of the elderly population. Alzheimer's disease (AD) is the most common form of major NCD, associated with loss of short-term-memory, problems with language, disorientation and other intellectual abilities, severely affecting daily life[5]. Worldwide, currently 35 Million people have been diagnosed with major neurocognitive disorder, which has been associated with 530 billion Euros in 2010[6]), tendency increasing[7]. While there is no palliative care, musical therapies have been proposed as efficient therapeutic means, acting as a powerful catalyst for precipitating memories, shown in a number of studies [36, 33, 22]. Specifically *mnemotherapy* can help elicit autobiographical memories by promoting positive emotional memories [2]. This and other therapies can improve the quality of life in AD patients [23, 1]. However, the assessment of such therapies requires comprehensive manual observation by experienced clinicians [32, 10]. Towards overcoming this limitation, computer vision based methods can offer objective assessment by *analyzing affect and expression behaviors, directly related to the effectiveness of therapies.*

While *expression recognition* has attracted significant research attention [28, 19, 38], facial behavior analysis from naturalistic videos, associated to illumination changes, partial occlusions, pose variation, as well as low-intensity expressions pose challenges for current existing methods. In addition, while many areas of computer vision have experienced significant advancements with deep neural networks, analysis of facial dynamics has only recently benefited from deep convolutional networks [34, 14, 43, 26].

Naturally, the accuracy of facial dynamics classification depends on features, as well as architecture used for assessment. Given the plethora of existing algorithms, exploring different types of features and architectures is necessary to devise a robust solution.

Motivated by the above, in this work we explore and compare computer vision methods, introduced in the context of *action recognition* in our challenging setting, namely in the context of *assessing naturalistic facial dynamics*. Specifically we have (a) 3D Convolutional Neural Network (C3D) [35], (b) Very Deep Two-Stream Convolutional Network (with VGG-16 and ResNet-152 [30, 42]) and (c) Improved Dense Trajectories (iDT) [39], as well as combinations and variations thereof. Given a video sequence, we firstly detect the face and proceed to extract features pertaining to the respective method. The obtained feature set is then classified in one of four facial dynamics categories, namely *neutral*, *smiling*, *talking* and *singing*. The automatic detection of named facial dynamics indicates the involvement of the patients during mnemotherapy and hence can support the assessment of a therapy session. Specifically, given that AD-patients in later stages of the disease often suffer from apathy, the (frequent) occurrence of smiling, talking and singing indicates that the therapy is effective. Experiments are conducted on a challenging medical unconstrained dataset containing 322 video sequences of 16 AD-patients including continuous pose-changes, oc-

[5] www.alz.org/alzheimers_disease_what_is_alzheimers.asp

[6] http://www.alz.org/news_and_events_20608.asp

[7] http://www.alz.org/

clusions, camera-movements and artifacts, as well as illumination changes. In addition, the dataset depicts naturalistic facial dynamics of predominantly elderly subjects, which vary in (generally less profound) intensity and occur jointly (*e.g.*, simultaneous talking and smiling). Moreover, we observe a high level of inter- and intra person variability (*e.g.*, expressive and apathetic AD-patients). We note that, despite that, it is imperative to work with such data, as it is representative for current (vast amount of) video-documentation of medical doctors, requiring automated analysis.

We note that we tested existing methods in expression recognition, such as smile detectors[8], [4] on the ADP - dataset, as well as facial-landmark based expression recognition algorithms without success, since already the first incorporated step of face detection failed throughout.

2 Related Work

Existing approaches for the *analysis of facial dynamics* are inspired by cognitive, psychological and neuroscientific findings. The most frequent way to describe facial dynamics is based on the Facial Action Coding System (FACS) proposed by Ekman *et al.* [7], representing movements of facial muscles in different terms of action units (AUs). Hence, classical methods analyze sequences of images containing the neutral face and the expression apex [19]. More recent methods involve linear deterministic and probabilistic methodologies including general or special Linear Dynamical Systems (LDS), as well as various extensions of deterministic Slow Feature Analysis (SFA) [43]. In addition, HMMs [27] have been used to capture the temporal segments of facial behaviour.

More recently, learning facial features in supervised and unsupervised manner using deep neural networks has attracted considerable attention. We proceed to provide such notable work, analyzing both, images and video sequences.

Recognizing facial dynamics in images Liu *et al.* [18] propose a Boosted Deep Belief Network, integrating three separate training stages for expression recognition in images. Han and Meng [11] present the incremental boosting of CNN for AUs recognition. Zhao *et al.* [44] combine region learning, as well as multiple label learning to detect AUs.

Recognizing facial dynamics in video sequences When analyzing facial behavior in videos, many works usually focus on spatial-temporal feature extraction. Jung *et al.* [14] use two neural networks separately to extract temporal appearance features, as well as temporal geometric features for expression recognition. Zafeiriou *et al.* [43] propose a slow-feature-auto-encoder for both supervised and semi-supervised learning of facial behavioural dynamics analysis. Hasani and Mahoor [12] combine a 3D Inception-ResNet with a Long short-term memory (LSTM) network in order to extract both, spatial and temporal features from videos. Li *et al.* [17] combine VGG with ROI and LSTM together towards detection of AUs. Most recently, combining Variational Autoencoder (VAEs)

[8] https://ibug.doc.ic.ac.uk/resources/smile-detectors/

and Generative Adversarial Networks (GANs) has allowed for learning a powerful latent representation utilized for facial behavior analysis in audience [26].

Previous and recent **computer vision work related to healthcare** have focused on, among others the assessment of: cognitive health in smart home environment [5], daily activities [15], AD symptoms [25], depression [45, 6], assistive technologies [16], as well as pain [24].

The rest of the paper is organized as follows. Section 3 describes the methods we compare. Section 4 introduces our dataset, assembled for the purpose of medical patient recording. Section 5 presents experiments, validating the effectiveness of the evaluated methods in assessing facial dynamics. Finally, Section 5.3 discusses and Section 6 concludes the paper.

3 Evaluated Methods

Firstly, we utilize face detection, based on which we crop the faces and proceed to extract facial features using Improved Dense Trajectories, as well as two deep neural network models. The latter have been pre-trained on a *large-scale* human action dataset UCF101 [31]. We inherit the weights in the neural network models and proceed to extract features of our dataset. Finally, we employ Support Vector Machine (SVM) to classify video sequences into four facial dynamics: *smiling*, *talking*, *neutral* and *singing*.

3.1 Face detection

There exist a large number of face detection algorithms, based on a large number of features and implementations. We compared a number of pre-trained algorithms including VGG [9], OpenCV [37], and Doppia [20] with our *ADP-dataset* (see Section 4). The latter performed the best and was hence included in the pre-processing step.

3.2 3D ConvNets

3D ConvNets (C3D) has a simple architecture (see Fig. 1) and has shown a performance of 85.2% accuracy on the UCF101 dataset. We adapt the C3D architecture and extract spatial-temporal features towards categorization of AD patients' facial dynamics.

The original C3D network has 8 convolutional layers, 5 max-pooling layers, 2 fully connected layers and a softmax loss layer. For 5 convolutional layers from 1 to 5, the number of convolutional kernels are 64, 128, 256, 256, 256 respectively. Because all kernels have 3 dimensions, the additional parameter d indicates the kernel temporal depth. Tran *et al.* [35], report for d=3 the best among all experiments, which we also use in the present work. In C3D network, all convolutional kernels are $3 \times 3 \times 3$ with stride 1 in both, spatial and temporal directions, in order to ensure a 3 dimensional output. Such an architecture allows

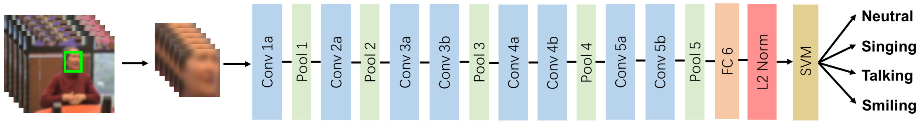


Fig. 1: **C3D based facial dynamics detection:** For each video sequence, faces are detected and the face sequences are passed into a pre-trained C3D network to extract a 4096-dim feature vector for each video. Finally a SVM classifier is trained to predict the final classification result. We have blurred the faces of the subject in this figure, in order to preserve the patient's privacy.

for preserving the temporal information between neighboring frames in a video-clip. With the exception of the first pooling layer, all pooling layers are max-pooling layers, with kernel size 2 × 2 × 2 with stride 1, attaining that the size of the feature map of each layer is reduced by a factor of 8 as opposed to the input(see Figure 2). The kernel size of the first pooling layer is 1 × 2 × 2, which ensures that early merging of the temporal signals is avoided. Since we only need features from the FC6 activation layer in our context, we remove the FC7 and last soft max layers from the original model.

We note that the C3D network was pre-trained on the UCF101 dataset, which contains 13320 videos from 101 action categories including single and multiple person actions. Specifically, each video has been divided into video clips of 16-frames-length, with an 8-frame overlap between two video clips by a sliding window method. All these video clips serve as training-input for the C3D network. The computed C3D-feature of a single video sequence is the average of all these clip FC6 activations followed by an *L2*-normalization.

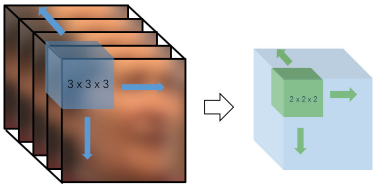


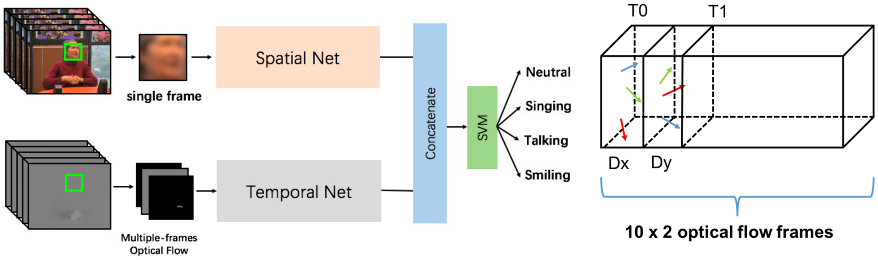
Fig. 2: **3D Convolutional kernel and 3D Max-pooling kernel:** In each convolutional layer except the first one, the kernel size is 3 × 3 × 3 and in each max-pooling layer the kernel size is 2 × 2 × 2. This 3 dimensional design can preserve both spatial and temporal information. We have blurred the faces of the subject in this figure, in order to preserve the patient's privacy.

3.3 Very Deep Two-Stream ConvNets

The second method, which we explore is Two-Stream ConvNets [30](see Fig. 3a), which has reportedly achieved 88% accuracy on the UCF101 action recognition dataset. It extracts features based on RGB frames, as well as based on optical flow from a video sequence. As reported by Wang *et al.* [42] and Feichtenhofer *et al.* [8], one successor network, namely the Very Deep Two-Stream ConvNets outperformed the original Two-Stream ConvNets (by 3% on UCF101).

Two-Stream ConvNets incorporates a *spatial ConvNet*, accepting as input single frame with dimension $224 \times 224 \times 3$, as well as a separate stream - a *temporal ConvNet*, accepting as input stacked optical flow fields, with dimension $224 \times 224 \times 20$. Specifically the optical flow field is composed of horizontal and vertical components D_x and D_y. A stack of D_x and D_y of 10 frames together are fed into the following ConvNet. Hence, while the first stream is based on RGB based features, the second stream is based on complementary motion between video frames, resulting to an increased accuracy over each of the streams.

We test two variations of Very Deep Two-Stream ConvNets, the first one including VGG-16 in both streams, the second one ResNet-152 in both streams. We note that for both, VGG-16 and ResNet-152, we remove the last fully connected layer and follow a *L2*-normalization step after the activations.



(a) Two-Stream Architecture

(b) Stacked Optical Flow Field volume

Fig. 3: a) While the *spatial ConvNet* accepts a single *RGB* frame as input, the *temporal ConvNet*'s input is the D_x and D_y of 10 consecutive frames, namely 20 input channels. Both described inputs are fed into the Two-stream ConvNets, respectively. We use in this work two variations of Very Deep Two Stream ConvNets, incorporating VGG-16 [29] ResNet-152 [13] for both streams respectively. (b) The optical flow of each frame has two components, namely D_x and D_y. We stack 10 times D_y after D_x for each frame to form a 20 frames length input volume.

Input configuration: Given a video sequence of T frames, we extract N *RGB* frames (*spatial ConvNet*) and N optical flow fields (*temporal ConvNet*).

The step of sampling in *spatial ConvNet* is $\left\lfloor \frac{T-1}{N-1} \right\rfloor$. If we stack dense optical flow of 10 sequential frames to form a 20 input volume (see Figure 3b, both horizontal and vertical components times 10), the sampling step for *temporal ConvNet* is $\left\lfloor \frac{T-10+1}{N} \right\rfloor$. For each optical flow field volume I, we have $I_{2t} = D_x, \quad I_{2t+1} = D_y, \quad t \in 10$.

The pre-trained *spatial* and *temporal* ConvNets extract two respective feature vectors, which concatenated serve as input for our classifier that we describe below.

3.4 Improved Dense Trajectories

Despite the prevalence of DNN, iDT as introduced by Wang *et al.* [40] constitute one of the best hand-crafted feature-based approaches. We employ iDT for their good coverage of foreground motion and high performance in action recognition (competitive to DNN). In addition, it is complementary to DNN and hence a fusion of iDT and DNN has shown to provide improved accuracy. iDT extracts local spatio-temporal video trajectories by applying dense sampling of feature points on multiple spatial scales with subsequent tracking of detected feature points using dense optical flow. We extract dense trajectories and proceed to extract local spatio-temporal video volumes around the detected trajectories. We extract 5 types of features aligned with the trajectories to characterize shape (point shifts), appearance (Histogram of Oriented Gradients (HOG) and motion (Histogram of Optical Flow (HOF) and Motion Boundary Histogram (MBHx, MBHy)). We encode the iDT features with bag-of-features (BOF) in order to represent video sequence using the extracted motion trajectories and their corresponding descriptors.

3.5 Classifier

For classification, we train a multi-class SVM classifier for the tested methods. We combine the methods *Grid-Search* and *Cross-validation* in order to obtain the best parameters.

4 Dataset

For this study, we created the Alzheimer's disease patients (ADP) - dataset, comprising of 322 video sequences including 16 female patients, with 5 or more takes of each facial dynamics class. The length of the video sequences ranged from 1.44 seconds to 33.08 seconds. All videos have been recorded on 25fps, with resolution of 576 × 720. Two patients with aphasia endued only video sequences of *neutral*, *smiling* and *singing*. Interestingly, while these two patients were not able to speak, they performed singing-like facial movements, which we labeled as *singing*. For this study we manually segmented and annotated the data, which was challenging, due to high intra- and inter-class variability of patients, as well as facial dynamics. In terms of annotation, two researchers annotated the data (one working in the area of computer vision, the other one in clinical experiments), overlapping in $> 85\%$. We note that facial dynamics appeared jointly (*e.g.*, singing and smiling), due to the unintrusive nature of the setting. Currently, such overlaps were not considered in the annotation, classes were annotated mutually exclusively.

The patients participated in individual mnemotherapy sessions, located in a small auditorium. Videos of these sessions were acquired with a camescope Sony Handycam DCR-SR 32, placed sideways of patient and clinician, capturing predominantly non-frontal and highly unconstrained videos of the patient.

We identified the most frequent occurring facial dynamics as *neutral*, *smiling*, *talking* and *singing / singing-like movements*. We note that even “neutral” there are still facial movements (*e.g.* blinking), hand or head movements.

5 Experimental Results

5.1 Implementation Details

We conduct our experiments on a single GTX Titan X GPU for both face detection and feature extraction, with emphasis on the use of C3D Network and Two-Stream ConvNets.

Face detection is performed by Doppia [20]. Due to the challenging dataset including among others variations of illumination, patient pose, as well as camera-movements, some faces are not detected in single video frames (constituting to false negatives). In such cases, we remove the concerning frames. In case of prevalence of undetected faces, we exclude the concerning video sequences from the analysis.

To compute optical flow, we follow the work of Wang *et al.* [41] and use the TVL1 algorithm implemented in OpenCV. In our experiments, we set $N = 25$, for both, RGB and optical flow, following the works [42, 41]. Each detected face (RGB-frame) is rescaled to $224 \times 224 \times 3$ and optical flow is rescaled to $224 \times 224 \times 20$ in Two-Stream Networks. For C3D we rescale each detected face to $112 \times 112 \times 3$.

For classification we use the scikit-learn library [21]. We employ a Stratified 10-Folds cross-validation scheme, which preserves the ratio of samples for each category for train and test set (see Table 2). The dataset is divided into 10-folds, 9 folds are used for training and the remaining fold is used for testing. This is repeated 10 times and reported results are the average thereof. We note that video sequences in the test set are not present in the training set. Per split, we compute mean accuracy (MA) (mean accuracy of 10 splits) and we report the MA over all splits.

We test SVM classifier with linear and radial basis function (RBF) kernels. Our experiments show that *e.g.,* for Two-Stream (ResNet-152) the RBF kernel performs best (with C = 25, gamma = 2).

5.2 Results

In Table 1 the performance of the facial dynamics-classification is presented as mean accuracy (MA) pertained to C3D, Two-Stream ConvNets (based on VGG-16 and ResNet-152), as well as separately to each spatial and temporal net. We observe that while the spatial and temporal net of Two-Stream ConvNets (ResNet-152) substantially outperform the VGG-16 counterparts, the overall Two-Stream ConvNets (VGG-16) and Two-Stream ConvNets (ResNet-152) perform comparably well. The best performance was obtained by the ResNet-152 based Two-Stream ConvNets, namely $MA = 76.40\%$, marginally outperforming VGG-16 based Two-Stream ConvNets (by 0.4%), and substantially outperforming C3D network (by 9%). In addition, we show performance of C3D and Two-Stream ConvNets fused with iDT. When fusing iDT with the other algorithms, the classification rate consistently increases, up to 79.5% for Two-Stream (ResNet-152).

Table 1: Classification accuracies of C3D, Very Deep Two-Stream ConvNets, iDT, as well as fusion thereof on the presented ADP-dataset. We report the Mean Accuracy (MA) associated to the compared methods. Abbreviations used: SN...Spatial Net, TN...Temporal Net.

Method	**MA (%)**
C3D	67.4
SN of Two-Stream ConvNets (VGG-16)	65.2
TN of Two-Stream ConvNets (VGG-16)	69.9
Two-Stream ConvNets (VGG-16)	76.1
SN of Two-Stream ConvNets (ResNet-152)	69.6
TN of Two-Stream ConvNets (ResNet-152)	75.8
Two-Stream ConvNets (ResNet-152)	76.4
iDT	61.2
C3D + iDT	71.1
Two-Stream ConvNets (VGG-16) + iDT	78.9
Two-Stream ConvNets (ResNet-152) + iDT	**79.5**

The classification rates for each split of the 10-fold cross-validation are reported in Table 2.

In Fig. 4 we present the overall confusion matrix, associated to the best performing algorithm – Two-Stream ConvNets (ResNet-152). The related results indicate that the highest confusion rates are observed between the dynamics *smiling*, *singing* and *talking*. This may be explainable by the co-occurrence of facial dynamics, as well as with the general low-intensity facial dynamics exhibited by the elderly patients. In some cases, the categories *neutral* and *smiling* have

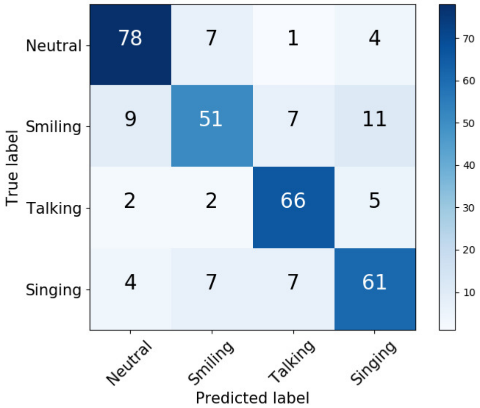


Fig. 4: **Confusion matrix for categorized facial dynamics of Two-Stream ResNet-152 + iDT (best performing method) on the ADP-Dataset.**

been confused. In Table 2 we show accuracy for each facial dynamics-category associated to Fig. 4. We see that the facial dynamic with highest classification rate is *neutral*, which is intuitive due to the discriminative low-motion. *Smiling* on the other hand is classified with the biggest error, which might be due to the low-intensity expressions exhibited by the elderly patients (see Figure 5). We here note that annotation was performed utilizing audio and video.

Table 2: Classification accuracy of Two-Stream ConvNets (ResNet-152) + iDT on the ADP dataset. The numbers in parentheses indicate the number of "neutral", "smiling", "talking" and "singing" samples in each split.

Split	**Train**	**Test**	**Accuracy (%)**
1	289 (81, 70, 67, 71)	33 (9, 8, 8, 8)	78.8
2	289 (81, 70, 67, 71)	33 (9, 8, 8, 8)	75.8
3	289 (81, 70, 67, 71)	33 (9, 8, 8, 8)	78.8
4	289 (81, 70, 67, 71)	33 (9, 8, 8, 8)	87.9
5	289 (81, 70, 67, 71)	33 (9, 8, 8, 8)	78.8
6	290 (81, 70, 68, 71)	32 (9, 8, 7, 8)	78.1
7	290 (81, 70, 68, 71)	32 (9, 8, 7, 8)	84.4
8	290 (81, 70, 68, 71)	32 (9, 8, 7, 8)	71.9
9	291 (81, 71, 68, 71)	31 (9, 7, 7, 8)	83.9
10	292 (81, 71, 68, 72)	30 (9, 7, 7, 7)	76.7
Average			79.5

Table 3: Mean Accuracy (%) of Two-Stream ConvNets (ResNet-152) + iDT on the ADP dataset. Assessment by category.

Method	**Neutral**	**Smiling**	**Talking**	**Singing**
iDT	75.6	39.8	68.0	59.5
C3D + iDT	75.6	53.8	81.3	64.4
Two-Stream (ResNet-152)+iDT	86.7	65.4	88.0	77.2

In a similar healthcare setting, an algorithm distinguishing between similar facial expressions and activities, based on spatio-temporal Dense Trajectories and improved Fisher Vectors has been proposed [3], which we outperform by 16.1% utmost.

5.3 Observations and Future Work

In this section we summarize the main findings of this research.

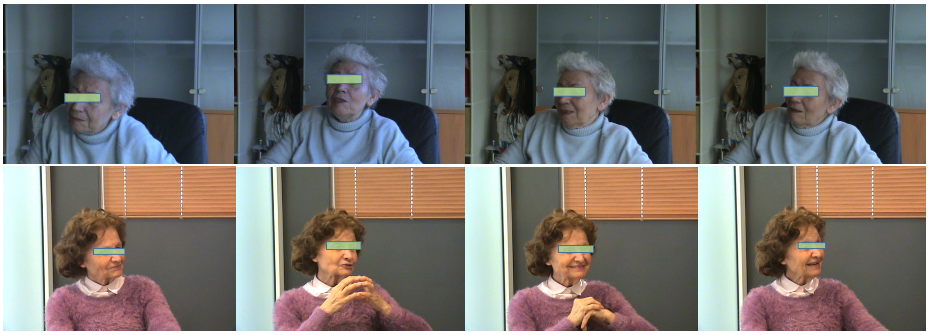


Fig. 5: Example images of two subjects from the ADP-dataset. From left to right we depict the classes “neutral”, “talk”, “smile” and “sing”. Low-intensity expressions exhibited by elderly patients impede correct classification in some cases.

- Based on our experiments with iDT and the DNN-architectures, we observe that DNNs contribute highly in obtaining very promising classification rates, despite the small size of our dataset. We note that while the presented results significantly outperform previous methods based on handcrafted features (*i.e.,* [3]), when fusing handcrafted features (*e.g.,* iDT) with DNN-based approaches, accuracy increases consistently.
- The methods adapted from action recognition generalize well to classification of facial dynamics. We observe this by the good classification rates, as

well as by the facts that (a) *temporal ConvNets* performs better than *spatial ConvNets* (cf. [8] [30]), (b) fusion with iDT consistently improves the performance of DNNs (cf. [42] [8]), (c) Two-Stream ConvNets outperforms C3D (cf. [8]).

- Due to the limited size of our dataset, an end-to-end training from scratch of a DNN architecture is not feasible. Large action recognition datasets such as the UCF101 human action dataset offer suitable training alternatives for DNN-architectures.
- The accuracy of facial dynamics classification depends on the features, as well as architecture used for assessment. Given the myriad of existing algorithms, exploring different types of features and architectures will be necessary to devise a robust solution.
- High inter- and intra-variance of subjects and facial dynamics contribute to the remaining error rates. Further challenges include the low-intensity of facial dynamics exhibited by elderly patients, the unconstrained setting, allowing for facial dynamics to occur jointly, as well as ambiguous human annotation.

However, more work is necessary in this regard. Future work will involve fine-tuning of the involved methods. In addition, we intend to explore personalized facial dynamics assessment, where we will train algorithms on video sequences related to each patient, individually. Finally, we will design specific neural network models for facial dynamics assessment, placing emphasis on a single end-to-end model, incorporating face detection, facial feature extraction, as well as classification.

6 Conclusions

In this work we have compared three methods for assessment of facial dynamics exhibited by AD-patients in mnemotherapy. The three tested methods include Improved Dense Trajectories, 3D ConvNets and Two-Stream ConvNets, which we have adapted from action recognition. Despite the pre-training of mentioned methods on an action recognition dataset, the methods have generalized very well to facial dynamics. Experiments conducted on an assembled dataset of Alzheimer's disease patients have resulted in a true classification rate of up to 79.5% for the fusion of Two-Stream ConvNets (ResNet-152) and iDT.

Acknowledgement

Y. Wang and A. Dantcheva were supported by the French Government (National Research Agency, ANR) under grant agreement ANR-17-CE39-0002.

References

1. Ashida, S.: The effect of reminiscence music therapy sessions on changes in depressive symptoms in elderly persons with dementia. Journal of Music Therapy **37**(3), 170–182 (2000)

2. Broutart, J.C., Robert, P., Balas, D., Broutart, N., Cahors, J.: Démence et perte cognitive: Prise en charge du patient et de sa famille, chap. Mnémothérapie, reviviscence et maladie d'Alzheimer. De Boeck Superieur (March 2017)
3. Dantcheva, A., Bilinski, P., Nguyen, H.T., Broutart, J.C., Bremond, F.: Expression Recognition for Severely Demented Patients in Music Reminiscence-Therapy. In: EUSIPCO (2017)
4. Dantcheva, A., Bremond, F.: Gender estimation based on smile-dynamics. IEEE Transactions on Information Forensics and Security (TIFS) **12**(3), 719–729 (2017)
5. Dawadi, P.N., Cook, D.J., Schmitter-Edgecombe, M., Parsey, C.: Automated assessment of cognitive health using smart home technologies. Technology and health care **21**(4), 323–343 (2013)
6. Dibeklioglu, H., Hammal, Z., Cohn, J.F.: Dynamic multimodal measurement of depression severity using deep autoencoding. IEEE Journal of Biomedical and Health Informatics **PP**(99), 1–1 (2017)
7. Ekman, P., Friesen, W.: Facial action coding system: a technique for the measurement of facial movement. Palo Alto: Consulting Psychologists (1978)
8. Feichtenhofer, C., Pinz, A., Zisserman, A.: Convolutional two-stream network fusion for video action recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (2016)
9. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. IEEE transactions on pattern analysis and machine intelligence **32**(9), 1627–1645 (2010)
10. Folstein, M.F., Folstein, S.E., McHugh, P.R.: "Mini-mental state": a practical method for grading the cognitive state of patients for the clinician. Journal of psychiatric research **12**(3), 189–198 (1975)
11. Han, S., Meng, Z., KHAN, A.S., Tong, Y.: Incremental boosting convolutional neural network for facial action unit recognition. In: Lee, D.D., Sugiyama, M., Luxburg, U.V., Guyon, I., Garnett, R. (eds.) Advances in Neural Information Processing Systems 29, pp. 109–117 (2016)
12. Hasani, B., Mahoor, M.H.: Facial expression recognition using enhanced deep 3d convolutional neural networks. In: Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on. pp. 2278–2288. IEEE (2017)
13. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. arXiv preprint arXiv:1512.03385 (2015)
14. Jung, H., Lee, S., Yim, J., Park, S., Kim, J.: Joint fine-tuning in deep neural networks for facial expression recognition. In: Computer Vision (ICCV), 2015 IEEE International Conference on. pp. 2983–2991. IEEE (2015)
15. König, A., Crispim Junior, C.F., Derreumaux, A., Bensadoun, G., Petit, P.D., Bremond, F., David, R., Verhey, F., Aalten, P., Robert, P.: Validation of an automatic video monitoring system for the detection of instrumental activities of daily living in dementia patients. Journal of Alzheimer's Disease **44**(2), 675–685 (2015)
16. Leo, M., Medioni, G., Trivedi, M., Kanade, T., Farinella, G.M.: Computer vision for assistive technologies. Computer Vision and Image Understanding **154**, 1–15 (2017)
17. Li, W., Abtahi, F., Zhu, Z.: Action unit detection with region adaptation, multi-labeling learning and optimal temporal fusing. In: Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on. pp. 6766–6775. IEEE (2017)
18. Liu, P., Han, S., Meng, Z., Tong, Y.: Facial expression recognition via a boosted deep belief network. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). pp. 1805–1812 (2014)

19. Martinez, B., Valstar, M.F., Jiang, B., Pantic, M.: Automatic analysis of facial actions: A survey. IEEE Transactions on Affective Computing (2017)
20. Mathias, M., Benenson, R., Pedersoli, M., Van Gool, L.: Face detection without bells and whistles. In: European conference on computer vision. pp. 720–735. Springer (2014)
21. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011)
22. Raglio, A., Bellelli, G., Mazzola, P., Bellandi, D., Giovagnoli, A., Farina, E., Stramba-Badiale, M., Gentile, S., Gianelli, M., Ubezio, M., et al.: Music, music therapy and dementia: a review of literature and the recommendations of the italian psychogeriatric association. Maturitas **72**(4), 305–310 (2012)
23. Ridder, H.M., Gummesen, E., et al.: The use of extemporizing in music therapy to facilitate communication in a person with dementia: An explorative case study. Australian Journal of Music Therapy **26**, 6 (2015)
24. Rodriguez, P., Cucurull, G., Gonzàlez, J., Gonfaus, J.M., Nasrollahi, K., Moeslund, T.B., Roca, F.X.: Deep pain: Exploiting long short-term memory networks for facial expression classification. IEEE Transactions on Cybernetics (2017)
25. Romdhane, R., Mulin, E., Derreumeaux, A., Zouba, N., Piano, J., Lee, L., Leroi, I., Mallea, P., David, R., Thonnat, M., et al.: Automatic video monitoring system for assessment of Alzheimers disease symptoms. The journal of nutrition, health & aging **16**(3), 213–218 (2012)
26. Saha, S., Navarathna, R., Helminger, L., Weber, R.M.: Unsupervised deep representations for learning audience facial behaviors. arXiv preprint arXiv:1805.04136 (2018)
27. Sandbach, G., Zafeiriou, S., Pantic, M., Rueckert, D.: Recognition of 3d facial expression dynamics. Image and Vision Computing **30**(10), 762–773 (2012)
28. Sariyanidi, E., Gunes, H., Cavallaro, A.: Automatic analysis of facial affect: A survey of registration, representation, and recognition. IEEE transactions on pattern analysis and machine intelligence **37**(6), 1113–1133 (2015)
29. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. CoRR **abs/1409.1556** (2014)
30. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N.D., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems 27, pp. 568–576 (2014)
31. Soomro, K., Roshan Zamir, A., Shah, M.: UCF101: A dataset of 101 human actions classes from videos in the wild. In: CRCV-TR-12-01 (2012)
32. Suzuki, M., Kanamori, M., Watanabe, M., Nagasawa, S., Kojima, E., Ooshiro, H., Nakahara, D.: Behavioral and endocrinological evaluation of music therapy for elderly patients with dementia. Nursing & Health Sciences **6**(1), 11–18 (2004)
33. Svansdottir, H., Snaedal, J.: Music therapy in moderate and severe dementia of Alzheimer's type: a case–control study. International psychogeriatrics **18**(04), 613–621 (2006)
34. Tran, D.L., Walecki, R., Rudovic, O., Eleftheriadis, S., Schuller, B.W., Pantic, M.: Deepcoder: Semi-parametric variational autoencoders for facial action unit intensity estimation. CoRR **abs/1704.02206** (2017)
35. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3d convolutional networks. In: Proceedings of the IEEE international conference on computer vision. pp. 4489–4497 (2015)

36. Vink, A.C., Bruinsma, M.S., Scholten, R.J.: Music therapy for people with dementia. The Cochrane Library (2003)
37. Viola, P., Jones, M.J.: Robust real-time face detection. International journal of computer vision **57**(2), 137–154 (2004)
38. Walecki, R., Rudovic, O., Pavlovic, V., Pantic, M.: Variable-state latent conditional random field models for facial expression analysis. Image and Vision Computing **58**, 25 – 37 (2017)
39. Wang, H., Kläser, A., Schmid, C., Liu, C.L.: Dense trajectories and motion boundary descriptors for action recognition. Research Report RR-8050, INRIA (Aug 2012)
40. Wang, H., Schmid, C.: Action recognition with improved trajectories. In: Proceedings of the IEEE international conference on computer vision. pp. 3551–3558 (2013)
41. Wang, L., Qiao, Y., Tang, X.: Action recognition with trajectory-pooled deep-convolutional descriptors. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2015)
42. Wang, L., Xiong, Y., Wang, Z., Qiao, Y.: Towards good practices for very deep two-stream convnets. CoRR **abs/1507.02159** (2015)
43. Zafeiriou, L., Nikitidis, S., Zafeiriou, S., Pantic, M.: Slow features nonnegative matrix factorization for temporal data decomposition. In: Image Processing (ICIP), 2014 IEEE International Conference on. pp. 1430–1434. IEEE (2014)
44. Zhao, K., Chu, W.S., Zhang, H.: Deep region and multi-label learning for facial action unit detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3391–3399 (2016)
45. Zhu, Y., Shang, Y., Shao, Z., Guo, G.: Automated depression diagnosis based on deep networks to encode facial appearance and dynamics. IEEE Transactions on Affective Computing **PP**(99), 1–1 (2017). https://doi.org/10.1109/TAFFC.2017.2650899