Expression Recognition for Severely Demented Patients in Music Reminiscence-Therapy

Antitza Dantcheva, Piotr Bilinski, Hung Thanh Nguyen, Jean-Claude Broutart, Francois Bremond

Expression Recognition for Severely Demented Patients in Music Reminiscence-Therapy

Antitza Dantcheva[1]*, Piotr Bilinski[2]*, Hung Thanh Nguyen[1], Jean-Claude Broutart[3], Francois Bremond[1]
[1] Inria, Sophia Antipolis, France, Email: {antitza.dantcheva, hung.nguyen, francois.bremond}@inria.fr
[2] University of Oxford, Oxford, United Kingdom, Email: piotrb@robots.ox.ac.uk
[3] GSF Noisiez, Biot, France, Email: jc.broutart@free.fr

***Abstract*—Recognizing expressions in severely demented Alzheimer's disease (AD) patients is essential, since such patients have lost a substantial amount of their cognitive capacity, and some even their verbal communication ability (*e.g.*, aphasia). This leaves patients dependent on clinical staff to assess their verbal and non-verbal language, in order to communicate important messages, as of the discomfort associated to potential complications of the AD. Such assessment classically requires the patients' presence in a clinic, and time consuming examination involving medical personnel. Thus, expression monitoring is costly and logistically inconvenient for patients and clinical staff, which hinders among others large-scale monitoring. In this work we present a novel approach for automated recognition of facial activities and expressions of severely demented patients, where we distinguish between four activity and expression states, namely *talking*, *singing*, *neutral* and *smiling*. Our approach caters to the challenging setting of current medical recordings of music-therapy sessions, which include continuous pose variations, occlusions, camera-movements, camera-artifacts, as well as changing illumination. Additionally and importantly, the (elderly) patients exhibit generally less profound facial activities and expressions in a range of intensities and predominantly occurring in combinations (*e.g.*, talking and smiling). Our proposed approach is based on the extension of the Improved Fisher Vectors (IFV) for videos, representing a video-sequence using both, local, as well as the related spatio-temporal features. We test our algorithm on a dataset of over 229 video sequences, acquired from 10 AD patients, with promising results, which have sparked substantial interest in the medical community. The proposed approach can play a key role in assessment of different therapy treatments, as well as in remote large-scale healthcare-frameworks.**

I. Introduction

The elderly population has been growing at an accelerated pace[1] and it has been reported that around half of the current population of over 75 year old subjects suffers from physical and / or mental impairments and as a result are in need of high level of care[2]. Dementia is one of the major challenges affecting the quality of life of the elderly and their caregivers, with Alzheimer's disease (AD) being the most common form of dementia, associated with loss of memory and other intellectual abilities, interfering with daily life[3]. Currently 35 million people are living with dementia worldwide (associated estimated worldwide cost of dementia was 530 billion euro in 2010[4]), needing a long-term care, predominantly focused on ensuring patients' comfort and treating discomfort [1]. Except for palliative care, there is no known cure for AD. At the same time, different methods have been explored to alleviate the effects of AD. One such approach, which has gained high interest in the last years is **music therapy**. When selected carefully and used appropriately, music can improve social behavior [2], reduce agitation [3] and facilitate cognitive function [4], and is hence instrumental to improve the quality of life for AD patients [5]. Furthermore, music can help elicit autobiographical memories by promoting positive emotional memories, which is referred to as *music reminiscence - therapy* that is based on mnemotherapy [6].

Although scientists have agreed that music therapy is an effective intervention for dealing with the symptoms of dementia [7], there is no consensus on assessment methods, which has precluded the finding of effective music therapy practices [7]. Classically the assessment of musical therapy sessions involves the subjective and error-prone manual observation and counting of occurrence of a set of behaviors. It has been shown that such testing has limitations concerning the ecological validity, reproducibility and objectivity [8]. Towards overcoming these limitations, information and communication technology (ICT) can provide a set of advanced tools, exploiting advanced methods, such as automated face detection, tracking, as well as human behavior analysis. Related technologies carry the promise to relieve healthcare systems from excessive workload, while decreasing the cost, as well as potentially increasing the performance of classical healthcare in a non-invasive and efficient manner.

A. Contribution

Motivated by the above, we propose a novel framework for recognition of facial activities and expressions in music reminiscence - therapy sessions that can be directly applied in an automated therapy assessment tool. Given a video-sequence of a facial activity / expression of an AD-patient, the proposed method (a) detects the face, (b) proceeds to extract improved dense trajectories (IDT), and (c) subsequently extracts local spatio-temporal features around the trajectories, and (d) proceeds to encode these by a novel addition -

*Authors contributed equally. The second author worked on this project, while he was at Inria.

[1] http://www.un.org/esa/population/publications/worldageing19502050/pdf/80chapterii.pdf

[2] https://www.agingcare.com/

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

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

spatio-temporal Improved Fisher Vectors. We have adopted the approach from Bilinski and Bremond [9], where it has been used for human violence recognition. Based on these, the four most frequently occurring activities and expressions in musical therapy (*talking*, *singing*, *neutral* and *smiling*) are (e) classified by support vector machines (SVM). We test and validate our approach on video-data, which we have collected during musical reminiscence - therapy at an AD-clinic, where even patients suffering from severe apathy exhibit a number of facial activities and expressions.

Challenging for the algorithm has been a set of factors related to the highly unconstrained real-life setting of music-therapy sessions, that includes continuous pose-variations, occlusions, camera-movements, camera-artifacts, as well as changing illumination. Additionally and importantly, facial activities and expressions were not triggered or posed, therefore occurring naturally in a range of intensities and often in combination (*e.g.*, talking and smiling). Moreover the observed expressions were inflicted with high inter- and intra person variability (*e.g.*, introvert and extrovert patients vs. different expression-intensity within single patients). A final challenge comprised the analysis of elderly subjects, which generally brings to the fore the exhibition of less profound expressions than younger subjects. 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.

B. Related work

Existing assisted living technologies based on ICT have focused on the assessment of daily activities [8], tests of cognitive functioning [10], as well as identification of patients [11].

Deviating from the above, in this work, we propose an approach to distinguish between facial activities and expressions in severely demented patients. Recognizing such facial expressions, as well as body motion, postures, gestures and communication through vocalizations in severely demented patients is of essential character, since AD-patients have lost a big amount of their cognitive capacity, and some even their verbal communication ability (*e.g.*, aphasia) and are dependent on clinical staff to assess their verbal / non-verbal language, in order to communicate important messages, as of the discomfort associated to potential complications of AD, such as falls, aspiration and infection [12].

Facial expression recognition is a prominent topic of emotion analysis and is a key factor in applications such as human-computer-interaction, crowd analytics, as well as surveillance. Generally, expressions are classified in one of 7 categories (neutral, happy, surprised, fearful, angry, sad, disgusted), as proposed by Ekman [13]. Expressing emotions is a highly individual task, which challenges automated expression recognition systems significantly [14]. At the same time, human emotion-expression is a key factor in human interaction. Mehrabian [15] made the interesting observation that 7% of communication information is transferred by linguistic language, 38% by paralanguage, and 55% by facial expressions.

Automated facial expression systems can be categorized as model-based and image-based approaches.

In *model-based approaches* facial landmarks and their movements are tracked and utilized for recognizing facial expressions. While such approaches are usually low-dimensional, they often require robust facial feature tracking techniques that can be challenged with variation in pose, illumination and expression, as well as varying image quality.

Image-based approaches on the other hand are holistic, hence features are extracted from the whole image, which allows for fast and simple computation, at the price of high dimensionality [14]. Apart from feature extraction, *feature classification* is a key ingredient in expression recognition. Prominent classifiers in this context include SVM [16], [17], [18], as well as radial-basis function neural networks (RBFNN) [14].

II. Proposed Method

We propose a holistic method for facial activity and expression recognition (see Fig. 1), which given a video-sequence first detects the face, proceeds to extract improved dense trajectories (IDT) (dense sampling and tracking of extracted interest points using dense optical flow), and subsequently extracts local spatio-temporal features around the trajectories. Extracted features are tailored towards characterizing (a) shape of a trajectory, (b) appearance (Histogram of Oriented Gradients (HOG) and Video Covariance Matrix Logarithm descriptor (VCML)) and (c) motion (Histogram of Optical Flow (HOF) and Motion Boundary Histogram (MBH)), which jointly generally represent a comprehensive description of a video [19]. The features are then encoded by spatio-temporal Improved Fisher Vectors. Finally based on the encoded features, four facial activities and expressions, namely *talking*, *singing*, *neutral*, and *smiling* are classified in each video-sequence by support vector machines (SVM).

In the following, we provide details on each step.

A. Face Detection

There exist a number of face detection algorithms, based on a large number of features and implementations. We tested and evaluated four such pre-trained algorithms, *i.e.*, Dlib[5], VGG [20], OpenCV [21], and Doopia [22] with the ALZH-dataset, the latter performing best and used in the proposed algorithm.

B. Dense Trajectories

Dense trajectories [23] are based on densely sampled feature points (considering multiple spatial scales) and proceeded tracking using dense optical flow. Our approach firstly extracts dense trajectories and subsequently extracts local spatio-temporal video volumes around the detected trajectories.

Trajectory Shape, VCML, HOG, HOF, and MBH descriptors: We align five features with the extracted trajectories, in order to characterize shape of trajectories, appearance

[5]http://dlib.net/

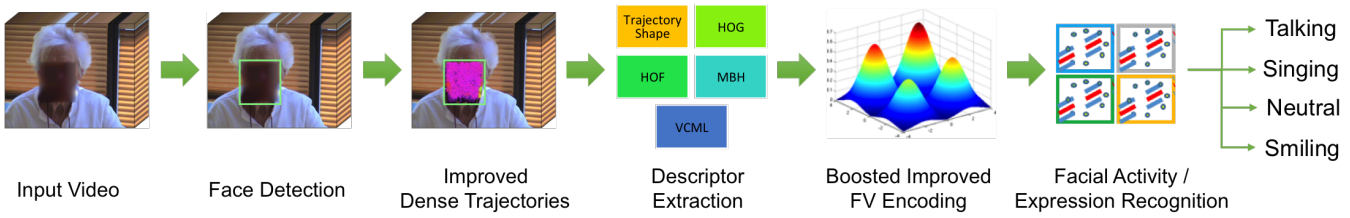

Fig. 1. Overview of proposed method for facial activity and expression recognition.

and motion. *Trajectory Shape* [23] encodes a sequence of displacement vectors, normalized by the sum of displacement vector magnitudes. *VCML, HOG, HOF,* and *MBH* are computed within a space-time volume around a trajectory. To embed structure information, each local volume is subdivided into a grid of $n_x \times n_y \times n_t$ spatio-temporal cells, where for each cell of the grid, a histogram descriptor is computed. Then, the histograms are normalized with the L_2 norm, and the normalized histograms from cells are concatenated into the final descriptors. The VCML descriptor [19] is based on a covariance matrix representation, modeling relationships between low-level features, such as intensity and gradient. For the HOG and HOF descriptors edge and optical flow orientations are quantized into 8 bins using full orientations, with an additional zero bin for the HOF descriptor. MBH divides the optical flow field $I_w = (I_x, I_y)$ into x and y components, spatial derivatives are computed separately for the horizontal and vertical components of the optical flow, and orientation information is quantized into histograms, similarly to the HOG descriptor. The MBH descriptor encodes the relative motion between pixels. Constant motion information is suppressed and only information related to changes in the flow field (*i.e.*, motion boundaries) is retained.

In the following, we utilize the parameters of spatial size of the volume 32×32, and n_x, n_y = 3, and $n_t = 2$, as in [19].

C. Improved Fisher Vectors (IFV)

We apply the IFV encoding to represent video-sequences of facial activities and expressions using the extracted motion trajectories and their corresponding descriptors. IFV [24] describes local features by their deviation from the "universal" generative Gaussian Mixture Model (GMM). The IFV has shown to achieve excellent results in activity and event recognition [19], [25], significantly outperforming the standard bag-of-features approach. IFV compute global statistics of local features only and disregard in the process the actual spatiotemporal positions of the features. This leads to the undesirable limitation - the loss of pertinent spatial information, a limitation that can be circumvented by the use of spatio-temporal grids [26], as well as multi-scale pyramids [27]. Both mentioned circumvention-methods provide only a coarse representation. Further related methods [28], [29] were proposed in the context of image categorization and object recognition, however cannot be directly transferred into the video-domain. We here propose to additionally exploit spatio-temporal positions of features, while keeping the representation more compact.

We proceed to revisit briefly the Improved Fisher Vectors, firstly introduced by Perronnin et al. [24].

Let $\mathbf{X} = \{\mathbf{x}_t, t = 1 \dots T\}$ be a set of T local features extracted from a video, where each local feature is of dimension D, $\mathbf{x}_t \in \mathbb{R}^D$. Let $\boldsymbol{\lambda} = \{w_i, \mu_i, \boldsymbol{\Sigma}_i, i = 1 \dots K\}$ be parameters of a GMM: $u_{\boldsymbol{\lambda}}(\mathbf{x}) = \sum_{i=1}^{K} w_i u_i(\mathbf{x})$ fitting the distribution of local features, where $w_i \in \mathbb{R}$, $\mu_i \in \mathbb{R}^D$ and $\boldsymbol{\Sigma}_i \in \mathbb{R}^{D \times D}$ are respectively the mixture weight, mean vector and covariance matrix of the i-th Gaussian u_i. We assume that the covariance matrices are diagonal and we denote the variance vector by σ_i^2, *i.e.* $\boldsymbol{\Sigma}_i = \mathrm{diag}(\sigma_i^2)$, $\sigma_i^2 \in \mathbb{R}^D$.

Moreover, let $\gamma_t(i)$ be the soft assignment of a descriptor $\mathbf{x}_t$ to a Gaussian i:

$$\gamma_t(i) = \frac{w_i u_i(\mathbf{x}_t)}{\sum_{j=1}^{K} w_j u_j(\mathbf{x}_t)}, \tag{1}$$

and let $\boldsymbol{\mathcal{G}}^{\mathbf{X}}_{\mu,i}$ (resp. $\boldsymbol{\mathcal{G}}^{\mathbf{X}}_{\sigma,i}$) be the gradient *w.r.t.* the mean μ_i (resp. standard deviation σ_i) of a Gaussian i:

$$\boldsymbol{\mathcal{G}}^{\mathbf{X}}_{\mu,i} = \frac{1}{T\sqrt{w_i}} \sum_{t=1}^{T} \gamma_t(i) \left(\frac{\mathbf{x}_t - \mu_i}{\sigma_i} \right), \tag{2}$$

$$\boldsymbol{\mathcal{G}}^{\mathbf{X}}_{\sigma,i} = \frac{1}{T\sqrt{2w_i}} \sum_{t=1}^{T} \gamma_t(i) \left[\frac{(\mathbf{x}_t - \mu_i)^2}{\sigma_i^2} - 1 \right], \tag{3}$$

where the division between vectors is as a term-by-term operation. Then, the gradient vector $\boldsymbol{\mathcal{G}}^{\mathbf{X}}_{\boldsymbol{\lambda}}$ is the concatenation of all K gradient vectors $\boldsymbol{\mathcal{G}}^{\mathbf{X}}_{\mu,i} \in \mathbb{R}^D$ and all the K gradient vectors $\boldsymbol{\mathcal{G}}^{\mathbf{X}}_{\sigma,i} \in \mathbb{R}^D$, $i = 1 \dots K$.

Finally, the IFV-representation $\boldsymbol{\Phi}^{\mathbf{X}}_{\boldsymbol{\lambda}}$, $\boldsymbol{\Phi}^{\mathbf{X}}_{\boldsymbol{\lambda}} \in \mathbb{R}^{2DK}$, is the gradient vector $\boldsymbol{\mathcal{G}}^{\mathbf{X}}_{\boldsymbol{\lambda}}$ normalized by the power normalization and then the L2 norm.

D. Augmenting IFV with spatio-temporal information

IFV-encoding attends to the simplistic assumption of conditional independence of spatial and temporal domains (see above). Hence, only global statistics of local features are considered, disregarding the actual spatio-temporal positions of the features. We propose an extension, the incorporation of spatio-temporal positions of features as follows.

Let $\mathbf{P} = \{\mathbf{p}_t, t = 1 \dots T\}$ be a set of T trajectories extracted from a video sequence and $\mathbf{p}_t = ((a_{t,1}, b_{t,1}, c_{t,1}), \dots, (a_{t,n_t}, b_{t,n_t}, c_{t,n_t}))$ is a sample trajectory, where a feature point detected at a spatial position

$(a_{t,1}, b_{t,1})$ in a frame $c_{t,1}$ is tracked in $n_t \geq 1$ subsequent frames until a spatial position (a_{t,n_t}, b_{t,n_t}) in a frame c_{t,n_t}. We define the video normalized position $\hat{\mathbf{p}}_t$ of a center of a trajectory $\mathbf{p}_t$ as:

$$\hat{\mathbf{p}}_t = \left[\frac{1}{v_w n_t} \sum_{i=1}^{n_t} a_{t,i}, \frac{1}{v_h n_t} \sum_{i=1}^{n_t} b_{t,i}, \frac{1}{v_l n_t} \sum_{i=1}^{n_t} c_{t,i} \right]', \quad (4)$$

where v_w is the video width (with the units in pixels), v_h is the video height (in pixels), and v_l is the video length (number of frames). We normalize the position of a center of a trajectory, so that the video size does not significantly change the magnitude of the feature position vector.

Once positions of local features are represented in a video normalized manner, we incorporate the normalized positions of local features into the Improved Fisher Vectors model, so that videos are represented using both local descriptors and their spatio-temporal positions.

Let $\mathbf{Y} = \{\mathbf{y}_t = [\tilde{\mathbf{p}}_t, \mathbf{x}_t], t = 1 \dots T\}$ be a set of local features, where $\mathbf{x}_t \in \mathbb{R}^D$ is a local feature descriptor and $\tilde{\mathbf{p}}_t \in \mathbb{R}^E$ is its corresponding normalized position, typically $E = 3$, calculated as above. Let $\tilde{\boldsymbol{\lambda}} = \{\tilde{w}_i, \tilde{\mu}_i, \tilde{\boldsymbol{\Sigma}}_i, i = 1 \dots K\}$ be parameters of a GMM $u_{\tilde{\boldsymbol{\lambda}}}(\mathbf{y}) = \sum_{i=1}^{K} \tilde{w}_i u_i(\mathbf{y})$ fitting the distribution of local features, where $\tilde{w}_i \in \mathbb{R}$, $\tilde{\mu}_i \in \mathbb{R}^{D+E}$ and $\tilde{\boldsymbol{\Sigma}}_i \in \mathbb{R}^{(D+E)\times(D+E)}$ are respectively the mixture weight, mean vector and covariance matrix of the i-th Gaussian. As before, we assume that the covariance matrices are diagonal and we denote by $\tilde{\sigma}_i^2$ the variance vector, *i.e.* $\tilde{\boldsymbol{\Sigma}}_i = \mathrm{diag}(\tilde{\sigma}_i^2)$, $\tilde{\sigma}_i^2 \in \mathbb{R}^{D+E}$. We calculate $\boldsymbol{\mathcal{G}}^{\mathbf{Y}}_{\tilde{\mu},i}$ (Eq. 2) and $\boldsymbol{\mathcal{G}}^{\mathbf{Y}}_{\tilde{\sigma},i}$ (Eq. 3) for all K Gaussian components, and concatenate all the gradient vectors into a vector $\boldsymbol{\mathcal{G}}^{\mathbf{Y}}_{\tilde{\boldsymbol{\lambda}}}$. Finally, the new Improved Fisher Vectors representation is the gradient vector $\boldsymbol{\mathcal{G}}^{\mathbf{Y}}_{\tilde{\boldsymbol{\lambda}}}$ normalized by the power normalization and then the L2 norm.

E. Classification: Support Vector Machines

For classification we use linear SVM [30], which has shown very good results with high-dimensional data (e.g., with Fisher vectors [24]).

III. Experiments

A. ALZH-Dataset

The musical therapy sessions took place in a small auditorium at the Noisiez GSF Foundation in Biot, France, where AD-patients underwent individual music sessions. The video data was acquired with a camescope Sony Handycam DCR-SR 32, placed on the side of patient and clinician, capturing predominantly non-frontal and highly unconstrained videos of the patient.

We identified the most frequent occurring facial activities and expressions as follows:

- **Talking**: Patients were talking in the introductory interviews, as well as while sharing their memories during the music-sessions, in case of reminiscence.
- **Singing / Singing-like movements**: In case of recognizing the music, the participants were singing, humming or performing singing-like movements in case of aphasia.
- **Neutral**: The neutral expression is important for its multiple implications. Predominantly the participants took the neutral expression, when they were in apathy. We note that even in a neutral expression, there are still facial movements (*e.g.* blinking), hand or head movements.
- **Smiling**: Smiles were exhibited, when participants recognized the music, when they were pleased by the music, when they experienced reminiscence, as well as during pleasant interaction with the clinician.

For this study, we created the ALZH - dataset, comprising of 229 video sequences including 10 female patients, with 5 or more takes of each activity / expression class. Two patients with aphasia endued only video sequences of *talking*, *singing* and *neutral*. 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 and facial activities and expressions. In addition, classes appeared jointly (*e.g.*, singing and smiling), due to the unintrusive nature of the setting.

B. Implementation Details

To estimate the GMM parameters for the IFV and the Spatio-Temporal IFV, we randomly sample a subset of $100,000$ features from the training set. We test 4 codebook sizes, namely $K = \{32, 64, 128, 256\}$, and we determine the number of Gaussians (*i.e.*, the codebook size) using cross-validation. To increase precision, we initialize the GMM ten times and we keep the codebook with the lowest error.

Towards evaluating the performance of our proposed algorithm, we employ a 10-folds cross-validation scheme. Specifically, the ALZH-dataset is divided into 10 folds, 9 folds are used for training, and the remaining fold is used for testing it. 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 calculate mean class accuracy (MCA) (mean accuracy from each class) and we report the average MCA over all splits.

C. Results

In Table I we present the average MCA of the proposed algorithm and 2 further variations thereof. Specifically, we investigate as a first variation (a) the performance without face detection. The rational is that head and hands movement might contain potentially useful information in expression recognition. However, we observe that, due to a vast amount of camera-artifacts (*i.e.*, the static background containing a seemingly considerable amount of motion), the analysis of the full-frames reduces the performance. Further, we report (b) the performance of the original IFV scheme. Our proposed algorithm significantly outperforms the original IFV-scheme from 50.83% to 53.99% (without face detection), and from 58.4% to 63.37% (with face detection). Additionally, we note that face detection significantly improves the performance of our proposed algorithm, namely from 53.99% to 63.37%.

TABLE I
AVERAGE MEAN CLASS ACCURACY (MCA).

IFV, no face detection	50.83%
IFV, face detection	58.4%
Spatio-temporal IFV, no face detection	53.99%
Spatio-temporal IFV, face detection	63.37%

We proceed to report the overall confusion matrix in Table II, associated to the best performing algorithm (spatio-temporal IFV, face detection). We note that the expression *Smiling* predominantly appeared in combination with talking or singing, which is reflected in the high confusion rates. Additionally, some patients exhibited restrictive emotionality (*i.e.*, the patients with aphasia) and hence intensity of expressions and activities was minor, which contributed to the confusion rates between *Neutral* and other classes.

TABLE II
CONFUSION MATRIX FOR RECOGNIZED EXPRESSIONS.

output / target class	**Talking**	**Singing**	**Neutral**	**Smiling**
Talking	27	8	5	12
Singing	8	38	3	10
Neutral	0	8	52	8
Smiling	12	4	6	28

In an attempt to compare our algorithm to other state-of-the-art algorithms, we tested smile detectors[6] [31] on our dataset, but without success, since already the first step - the contained face detection failed throughout.

IV. CONCLUSIONS

In this work we presented an approach for facial activity and expression recognition exhibited by AD-patients in musical reminiscence - therapy. The presented approach is holistic and has the advantage of being robust to real life settings that includes covariates, such as illumination, patient pose, as well as camera-movements. The proposed algorithm utilizes improved dense trajectories represented by spatio-temporal IFV-encoding. Our promising results suggest that the approach captures well even minor facial activities and expressions and has thus sparked substantial interest in the medical community. Future work involves the use of an increased number of trajectories, as well as the analysis of individualized expression recognition.

REFERENCES

[1] M. Maas. Management of patients with alzheimer's disease. *Nursing Clinics of North America*, 1(12):57–68, 1988.

[2] H. M. Ridder. 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.

[3] S. Ashida. 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.

[4] M. Brotons and S. M. Koger. The impact of music therapy on language functioning in dementia. *Journal of music therapy*, 37(3):183–195, 2000.

[5] M. P. Lawton. Quality of life in alzheimer disease. *Alzheimer Disease & Associated Disorders*, 8:138–150, 1994.

[6] J.-C. Broutart, P. Robert, D. Balas, N. Broutart, and J. Cahors. *Démence et perte cognitive: Prise en charge du patient et de sa famille*, chapter Mnémothérapie, reviviscence et maladie d'Alzheimer. De Boeck Superieur, March 2017.

[7] S. M. Koger, K. Chapin, and M. Brotons. Is music therapy an effective intervention for dementia? A meta-analytic review of literature. *Journal of Music Therapy*, 36(1):2–15, 1999.

[8] A. König, C. F. Crispim Junior, A. Derreumaux, G. Bensadoun, P.-D. Petit, F. Bremond, R. David, F. Verhey, P. Aalten, and P. Robert. 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.

[9] P. Bilinski and F. Bremond. Human violence recognition and detection in surveillance videos. In *AVSS*, 2016.

[10] S. Okahashi, K. Seki, A. Nagano, Z. Luo, M. Kojima, and T. Futaki. A virtual shopping test for realistic assessment of cognitive function. *Journal of neuroengineering and rehabilitation*, 10(1):1, 2013.

[11] T. Banerjee, J. M. Keller, and M. Skubic. Resident identification using kinect depth image data and fuzzy clustering techniques. In *EMBC*, 2012.

[12] A. C. Hurley, B. J. Volicer, P. A. Hanrahan, S. Houde, and L. Volicer. Assessment of discomfort in advanced Alzheimer patients. *Research in nursing & health*, 15(5):369–377, 1992.

[13] P. Ekman and W. V. Friesen. Facial action coding system. 1977.

[14] C. Y. Chang and Y. C. Huang. Personalized facial expression recognition in indoor environments. In *IJCNN*, 2010.

[15] A. Mehrabian. Communication without words. *Communication Theory,*, pages 193–200, 2008.

[16] G. Littlewort, M. S. Bartlett, I. Fasel, J. Susskind, and J. Movellan. Dynamics of facial expression extracted automatically from video. *Image and Vision Computing*, 24(6):615–625, 2006.

[17] A. Dantcheva, P. Bilinski, J. C. Broutart, P. Robert, and F. Bremond. Emotion facial recognition by the means of automatic video analysis. *Gerontechnology Journal*, 15(suppl):12s, 2016.

[18] P. Bilinski, A. Dantcheva, and F. Bremond. Can a smile reveal your gender? In *BIOSIG*, volume 15, 2016.

[19] P. Bilinski and F. Bremond. Video covariance matrix logarithm for human action recognition in videos. In *IJCAI*, 2015.

[20] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE TPAMI*, 32(9):1627–1645, 2010.

[21] P. Viola and M. J. Jones. Robust real-time face detection. *International journal of computer vision*, 57(2):137–154, 2004.

[22] M. Mathias, R. Benenson, M. Pedersoli, and L. Van Gool. Face detection without bells and whistles. In *ECCV*, 2014.

[23] H. Wang and C. Schmid. Action recognition with improved trajectories. In *ICCV*, 2013.

[24] F. Perronnin, J. Sánchez, and T. Mensink. Improving the Fisher Kernel for large-scale image classification. In *ECCV*. 2010.

[25] P. Bilinski, M. Koperski, S. Bak, and F. Bremond. Representing visual appearance by video brownian covariance descriptor for human action recognition. In *AVSS*, 2014.

[26] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *CVPR*, 2008.

[27] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006.

[28] S. McCann and D. G. Lowe. Spatially local coding for object recognition. In *ACCV*, 2012.

[29] J. Sánchez, F. Perronnin, and T. De Campos. Modeling the spatial layout of images beyond spatial pyramids. *Pattern Recognition Letters*, 33(16):2216–2223, 2012.

[30] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM TIST*, 2:27:1–27:27, 2011.

[31] A. Dantcheva and F. Bremond. Gender estimation based on smile-dynamics. *IEEE TIFS*, 12(3):719–729, 2017.

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