Quantification of advanced dementia patients' engagement in therapeutic sessions: an automatic video based approach using computer vision and machine learning

Quantification of Advanced Dementia Patients' Engagement in Therapeutic Sessions: An Automatic Video Based Approach using Computer Vision and Machine Learning

Liangfei Zhang[1], Ognjen Arandjelović[1], Sonia Dewar[2], Arlene Astell[3], Gayle Doherty[2], and Maggie Ellis[2]

Abstract— **Most individuals with advanced dementia lose the ability to communicate with the outside world through speech. This limits their ability to participate in social activities crucial to their well-being and quality of life. However, there is mounting evidence that individuals with advanced dementia can still communicate non-verbally and benefit greatly from these interactions. A major problem in facilitating the advancement of this research is of a practical and methodical nature: assessing the success of treatment is currently done by humans, prone to subjective bias and inconsistency, and it involves laborious and time consuming effort. The present work is the first attempt at exploring if automatic (artificial intelligence based) quantification of the degree of patient engagement in Adaptive Interaction sessions, a highly promising intervention developed to improve the quality of life of nonverbal individuals with advanced dementia. Hence we describe a framework which uses computer vision and machine learning as a potential first step towards answering this question. Using a real-world data set of videos of therapeutic sessions, not acquired specifically for the purposes of the present work, we demonstrate highly promising results.**

I. Introduction

A. Background and motivation

People living with dementia experience a gradual decline in a broad range of cognitive abilities [1]. In the late stages of the condition, it is common to exhibit a complete loss of speech ability [1]. This loss of speech can be misinterpreted as a loss of ability and desire to communicate altogether. Consequently, these individuals are viewed as socially unreachable [2] with little or no effort made to engage them socially. Recent evidence indicates that individuals with advanced dementia who cannot speak spend more than 68% of their time "detached from their environment" [3] with social interactions limited to conduct of care tasks such as bathing and dressing. Therefore, people with advanced dementia who cannot speak find themselves socially isolated [4] which is extremely harmful to their well-being. Moreover, loss of speech negatively impacts family members and friends who struggle to maintain contact and keep connected to individuals with very advanced dementia. Therefore, the development and implementation of techniques which improve the quantity and quality of communication of non-verbal dementia patients are crucial.

Several decades of research indicate that our main pathway for expressing social signals is through nonverbal means [5]. For example, Mehrabian and Epstein [6] proposed that people convey up to 93% of information about their emotions through nonverbal means (e.g. tone of voice and body language). Imitation in particular is a key non-verbal communication process that has been found to facilitate human connection [7].

The communicative potential offered by the natural capacity to imitate has been investigated in a number of clinical populations. Intensive Interaction is a technique that uses imitation to facilitate communication for individuals with Profound and Multiple Learning Disabilities [8]. Research on Intensive Interaction highlights that repetition, predictability and contingency in the adult's responses to pupils enables connection and understanding similar to that seen in early parent-child interaction [9]. Research into Intensive Interaction as a tool for connecting with individuals nonverbal from birth, demonstrates the importance of repetition, predictability and contingency in developing communication with caregivers that resemble early parent-infant interaction [9]. People with dementia retain the nonverbal fundamental elements of communication including imitation, after speech is lost, suggesting that these could be used to foster interactions [2].

Studies highlight that individuals with late stage dementia retain the capacity to use a repertoire of behaviours, such as gestures, sounds and bodily movements to express their feelings and willingness to engage [2]. This communicative repertoire can be used by a facilitator to encourage communication which is within the capabilities of the person with dementia, allowing them to initiate and turn take within an interaction which is not possible through spoken language [3]. This approach, termed Adaptive Interaction, uses imitation to make a connection with each individual and gradually learn their personal communicative repertoire. Work to date shows that Adaptive Interaction encourages a higher frequency of communicative behaviours, turn taking and improved interaction with caregivers [10].

However, at present there are major practical obstacles in advancing this direction of research. Firstly, the quantification of patient engagement in session requires a trained practitioner (and training is both costly and time consuming). Moreover, it introduces a degree of subjectivity and thus both inter- and intra-labeller variability, which poses a fundamental methodological issues and limits replicability

[1] School of Computer Science, University of St Andrews, UK

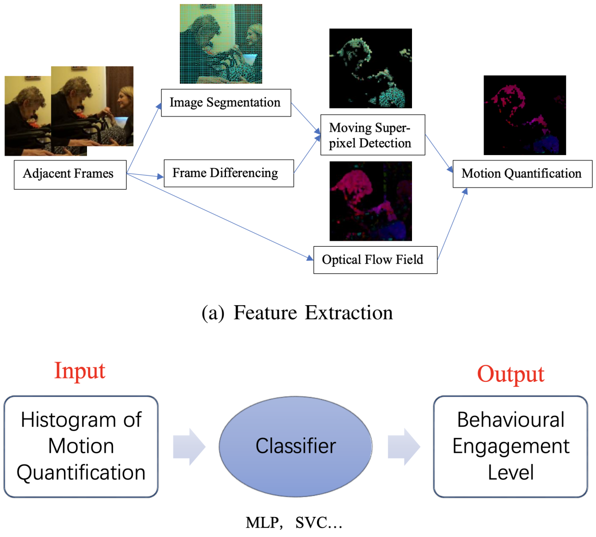
[2] School of Psychology and Neuroscience, University of St Andrews, UK

[3] School of Psychology & Clinical Language Sciences, University of Reading, UK

and credibility of reported findings. Secondly, the process of labelling of videos of therapeutic sessions is highly laborious as it has to be done on a fine temporal scale. This further increases financial and time burden. Thus, considering the indisputable successes of artificial intelligence in health care and medicine, the overarching aim of the present work is to investigate if this process can be automated. In particular, we propose a framework which employs, in the first stages, computer vision techniques for the extraction of meaningful information from video sequences of aforementioned sessions, and in the second stage, machine learning as a means of learning how to use this information to quantify patient engagement in them in a therapeutically useful manner.

II. Proposed Method

The overarching aim of the present work is to develop a framework which combines computer vision and machine learning, that can infer and quantify the degree of engagement of advanced dementia patients in interactive therapeutic sessions. In order for the aforementioned framework to be practically useful, it has to be robust enough to be able to deal with videos acquired in practically unconstrained environments with much clutter, such as care home rooms, and without strong *a priori* assumptions on the relative camera position (or, equivalently, location and pose of the patient) or the patient's clothing. The overall structure of information flow and extraction is summarized in Figure 1; each step is explained in detail next.



(a) Feature Extraction

(b) Classification

Fig. 1. The main building blocks of the proposed method.

A. Image over-segmentation

The concept of superpixels in computer vision refers to contiguous image regions with relatively homogeneous appearance characteristics [11]. What 'homogeneous' means is at least somewhat application dependent and is usually defined with reference to local texture, colour, or brightness. The use of superpixels not only dramatically reduces the dimensionality of the feature space, thereby conferring numerous computational advantages, but it also provides semantically more meaningful elementary primitives to work with.

Simple Linear Iterative Clustering (SLIC) [12] is a popular superpixel algorithm which has demonstrated successful performance in a range of different applications [13]. Hence, it is adopted in the present work too. For full detail the reader should consult the original paper but succinctly put SLIC generates superpixels by clustering pixels based on their similarity in colour and geometric proximity (image distance). Each pixel represented by a five-dimensional vector $[l, a, b, x, y]^T$, where l, a, b are the components in the CIELab colour space, and x, y are the coordinates of the pixel. The distance D_S between two pixels, denoted by subscript indices i and k, is then quantified as follows:

$$D_S = d_{lab} + \frac{m}{S} d_{xy} \tag{1}$$

$$d_{lab} = \sqrt{(l_k - l_i)^2 + (a_k - a_i)^2 + (b_k - b_i)^2} \tag{2}$$

$$d_{xy} = \sqrt{(x_k - x_i)^2 + (y_k - y_i)^2} \tag{3}$$

Where m is the parameter used to control the super-pixel compactness, generally set to a value between 1 and 20. The k-means algorithm is employed for clustering.

In the present work, SLIC is applied on individual frames of a video recording. Thereafter, as suggested earlier, superpixels are used as the elementary primitives for further processing.

B. Motion based detection of superpixels of interest

Frame differencing is one the simplest ways of detecting motion in video [14]. This method considers pixel-wise changes between successive frames and deems those that exhibit sufficient change (thresholding) as moving. In the present work, we adapt this broad idea to deal with superpixels.

The change at the locus p at time n, $D_n(p)$, can be computed from the values of pixels at the locus in frames with indexes n and $n+1$. If the difference is greater than a threshold T_p, the pixel is considered a moving one ($M_n(p) = 1$):

$$D_n(p) = |I_{n+1}(p) - I_n(p)| \tag{4}$$

$$M_n(p) = \text{sign}\,[D_n(p) - T_p] \tag{5}$$

To ameliorate the effects of noise (e.g. due to illumination), which is further amplified by the aforementioned binarization, we apply the morphological operations of opening and closing on the resulting output.

Following this post-processing step, if the proportion of moving pixels withing a superpixel S_j is greater than the threshold T_s, the superpixel is deemed a moving one ($\mu_n(j) = 1$) at time step n:

$$\mu_n(j) = \text{sign}\left[\sum_{p \in S_j} M_n(p) - T_s\right] \tag{6}$$

C. *From moving superpixels to spatially coherent motion*

To recap succinctly, thus far we have segmented the video into superpixels and detected those that exhibit significant motion by looking at the *magnitude* of the optic flow field in the corresponding image regions. The next step goes one level higher up the spatial ladder by making use of *motion coherence*. In particular, we now use the direction of optic flow too as a means of coherently moving image regions, as expected when looking at the motion of human body which is spatially continuous, albeit articulated. Formalizing this, the total pseudo-energy of patient motion is calculated as the average energy of coherent per-superpixel motion energies. It is important to note that it is crucial to perform normalization using absolute time (e.g. seconds) rather than relative (i.e. between successive frames) so as to achieve invariance to frame rate in processed videos.

D. *From frame characteristics to video level features*

Recall that our aim is to make inference on the level of an entire video recording of a therapeutic session. Working on a frame by frame basis is inadequate as people with advanced dementia are prone to sudden and transient reflexive motions and twitches – what is needed is a representation of motion across time from which the characteristics of genuine response to the therapist can be learnt. In the present work we achieve this by representing motion across a video using a compact representation in the form of a histogram of motion energy across the entirety of the session. To achieve invariance to session duration the resulting histogram is normalized to unit sum (i.e. unit L_1 norm). Thus produced histograms are used as input to the learning module (a classifier, as explained in the next section) in our framework.

III. Evaluation

In this section, we evaluate the proposed method experimentally on real-world data (not acquired specifically for the purposes of the present work) and discuss our findings.

A. *Data labelling and description*

Trained experts manually annotated the video recorded sessions on Observer XT$^{\text{TM}}$, utilizing the Video Coding-Incorporating Observed Emotion (VC-IOE) protocol, quantifying engagement levels of each participant [15]. The annotation scheme consists of six dimensions of engagement including agitation, facial expressions, verbal, visual, collective and behavioural engagement, the last of which was used in the present work. The resulting data set delineates the engagement duration (in ms) as the proportion of time the participant spent engaged in communication throughout the entire session. The total engagement score for each session was normalized to the uniform scale of 0 to 100.

In our data set the 30 participants' scores ranged from 0.0 to 31.94. Hence, we semantically stratified the behavioural engagement level as 'low' ($score = 0$), 'mid' ($0 < score \leq 5$), and 'high' ($5 < score$). The number of videos in the three classes are 13, 8, and 9, respectively. These classes are used as classifier targets used to learn how the features extracted from video data map into this semantically meaningful space, see TABLE I.

TABLE I

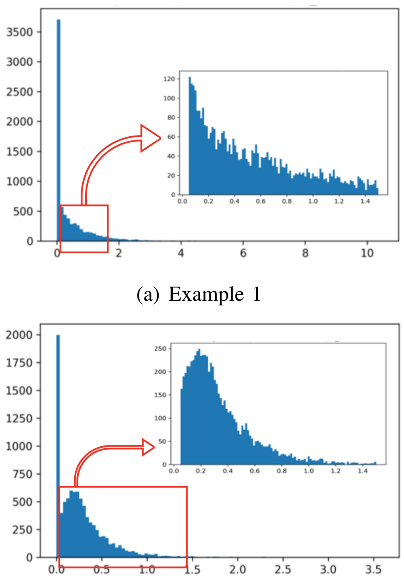
Summary statistics of experimental data.

Engagement (qualitative)	Engagement (quantitative)	Participants
Low	$score = 0$	13
Mid	$0 < score \leq 5$	8
High	$5 < score$	9

B. *Empirical evaluation*

In empirical evaluation of the proposed method, we used the features extracted from videos of therapeutic sessions as described in the previous section, in combination with well-known and well-understood classifiers, with two primary aims. The first of these concerns insight into the effectiveness of the introduced methodology. The second one concerns the broader question of feasibility of automatic means of addressing the underlying task.

We adopted stratified random sampling to split our video data set into training and validation, and test subsets, according to the proportion 0.75:0.25. The usual 5-fold verification protocol was employed for model parameter selection.



(a) Example 1

(b) Example 2

Fig. 2. Examples of our video level representation in the form of a histogram of frame-wise salient motion energy.

Motivated by prior work and the overarching successes of these on a diverse set of tasks [16], we adopt a neural network based approach, in the form of a multi-layer perceptron (MLP) with a single 60-neuron hidden layer and the rectified linear unit activation function, and a support vector based classifier (SVC).

C. *Result*

The key results are summarized in TABLE II. To comment on the performance on the coarse level first, we can see that both classifiers achieved highly promising results, thus confirming both the distal premise of the present work, namely that the task is question can be adequately addressed by automatic methods, and the proximal one, that is that the proposed feature extraction procedure is sensible and capable of capturing salient information in videos of interest.

Examining the results in more detail reveals additional interesting insights. Firstly, from the embedded confusion matrix we observe that both MLP and SVC based methods performed well across all levels of patient engagement i.e. the high overall performance was not unduly affected by outstanding recognition in some cases, and much worse in others. Continuing with the analysis, we observe that both classifiers never failed to identify a lack of engagement correctly. Input with mid and high patient engagement levels were also handled very well though now with some errors. This is only to be expected, given that there is a degree of subjectivity (or the drawing of proverbial line in the sand) in where the transition from the former to the latter is (note that high engagement was never confused with low but rather only with mid, further corroborating the claim that our feature extraction is meaningful); indeed, the error rate is no higher than the variability observed across different trained human annotators.

TABLE II

SUMMARY OF KEY RESULTS.

Classifier	Class	Confusion matrix			F1-score	Accuracy
		Low	Mid	High		
SVC	Low	1.00	0.00	0.00	0.86	0.90
	Mid	0.12	0.88	0.00	0.00	
	High	0.00	0.11	0.89	0.67	
MLP	Low	1.00	0.00	0.00	1.00	**0.97**
	Mid	0.00	1.00	0.00	0.67	
	High	0.00	0.11	0.89	0.67	

IV. SUMMARY AND FUTURE WORK

In this work we introduced the first algorithm for the automatic quantification of the degree of patient engagement in Adaptive Interaction sessions, a highly promising type of intervention aimed at individuals with advanced dementia who are non-verbal. The first part of the framework comprises a series of computer vision based techniques which ultimately extract meaningful features from video sequences of aforementioned sessions, and represent them in a compact and homogeneous manner. The pipelines can be broadly described as being bottom-up: local, low-level features in the form of frame differences and optic flow are aggregated to the level of superpixels, which are then further robustly integrated to describe frame level motion. Finally, video level characterization is achieved by considering frame level motion across the entire sequence. The second part of the framework is machine learning based – training and validation data sets are used to learn how our representation can be used to predict semantic labels provided by trained clinicians.

Our findings are highly promising and confirm two key premises of the present work. The first of these is that the task in question can indeed be addressed by automatic means. The second one concerns the specific framework described which is shown to be extremely successful on a real-world data set, to the best of our knowledge the first one of its type in the literature.

The present pioneering work opens a multitude of avenues for future work. Firstly, we intend to explore the use of more subtle behavioural aspects, such as precise head motion or indeed the motion of hands specifically, as well as, even more subtly, facial expression. Other possible directions for work concern more practical, computational issues, such as real-time processing and feedback, which would be of undoubted use to medical practitioners conducting therapeutic sessions.

REFERENCES

[1] Alzheimer's Association *et al.*, "2018 Alzheimer's disease facts and figures," *Alzheimer's & Dementia*, vol. 14, no. 3, pp. 367–429, 2018.

[2] M. Ellis and A. Astell, "Promoting communication with people with severe dementia. Zeedyk S.(Ed). Techniques for promoting social engagement in individuals with communicative impairments," 2008.

[3] ——, "Communicating with people living with dementia who are nonverbal: The creation of adaptive interaction," *PloS ONE*, vol. 12, no. 8, p. e0180395, 2017.

[4] V. Abad, "Reaching the socially isolated person with Alzheimer's disease through group music therapy-a case report," in *Voices: A World Forum for Music Therapy*, vol. 2, no. 3, 2002.

[5] P. M. Brunet, H. Donnan, G. McKeown, E. Douglas-Cowie, and R. Cowie, "Social signal processing: What are the relevant variables? and in what ways do they relate?" in *IEEE International Conference on Affective Computing and Intelligent Interaction and Workshops*, 2009, pp. 1–6.

[6] A. Mehrabian and N. Epstein, "A measure of emotional empathy." *Journal of Personality*, vol. 40, no. 4, pp. 525–543, 1972.

[7] M. Stel and K. van den Bos, "Mimicry as a tool for understanding the emotions of others," in *Proceedings of Measuring Behavior*, vol. 114, 2010, p. 117.

[8] M. S. Zeedyk, P. Caldwell, and C. E. Davies, "How rapidly does intensive interaction promote social engagement for adults with profound learning disabilities?" *European Journal of Special Needs Education*, vol. 24, no. 2, pp. 119–137, 2009.

[9] J. Watson and A. Fisher, "Research section: evaluating the effectiveness of intensive interactive teaching with pupils with profound and complex learning difficulties," *British Journal of Special Education*, vol. 24, no. 2, pp. 80–87, 1997.

[10] M. Ellis and A. Astell, *Adaptive interaction and dementia: how to communicate without speech.* Jessica Kingsley Publishers, 2017.

[11] X. Ren and J. Malik, "Learning a classification model for segmentation," in *Proceedings Ninth IEEE International Conference on Computer Vision*, 2003, pp. 10–17.

[12] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 11, pp. 2274–2282, 2012.

[13] M. Niemeyer and O. Arandjelović, "Automatic semantic labelling of images by their content using non-parametric Bayesian machine learning and image search using synthetically generated image collages." *International Conference on Data Science and Advanced Analytics*, pp. 160–168, 2018.

[14] K. A. Joshi and D. G. Thakore, "A survey on moving object detection and tracking in video surveillance system," *International Journal of Soft Computing and Engineering*, vol. 2, no. 3, pp. 44–48, 2012.

[15] C. Jones, B. Sung, and W. Moyle, "Assessing engagement in people with dementia: a new approach to assessment using video analysis," *Archives of psychiatric nursing*, vol. 29, no. 6, pp. 377–382, 2015.

[16] O. Arandjelović and R. Cipolla, "A new look at filtering techniques for illumination invariance in automatic face recognition." *In Proc. IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 449–454, 2006.