Quantifying dynamic facial expressions under naturalistic conditions

Authors: Jayson Jeganathan MBBS[1,2], Megan Campbell PhD[1,2], Matthew Hyett[3], Gordon Parker[4], Michael Breakspear PhD[1,2,5]

1. School of Psychology, College of Engineering, Science and the Environment, University of Newcastle, Newcastle, NSW, Australia
2. Hunter Medical Research Institute, Newcastle, NSW, Australia
3. School of Psychological Sciences, University of Western Australia, Crawley, WA, Australia
4. School of Psychiatry, University of New South Wales, Kensington, NSW, Australia
5. School of Medicine and Public Health, College of Medicine, Health and Wellbeing, University of Newcastle, Newcastle, NSW, Australia

Contact details

Dr Jayson Jeganathan (corresponding author)
HMRI Imaging, Hunter Medical Research Institute
Lot 1 Kookaburra Circuit, New Lambton Heights, NSW 2305, Australia
Jayson.jeganathan@gmail.com

Dr Megan Campbell
HMRI Imaging, Hunter Medical Research Institute
Lot 1 Kookaburra Circuit, New Lambton Heights, NSW 2305, Australia
megan.ej.campbell@gmail.com

Matthew Hyett
School of Psychological Sciences
35 Stirling Highway
University of Western Australia
Perth, WA 6009, Australia
matthewhyett@gmail.com

Professor Gordon Parker

School of Psychiatry
Level 1, AGSM Building
University of New South Wales
Kensington, NSW 2052, Australia
g.parker@unsw.edu.au

Professor Michael Breakspear
HMRI Imaging, Hunter Medical Research Institute
Lot 1 Kookaburra Circuit, New Lambton Heights, NSW 2305, Australia
mjbreaks@gmail.com

Author Contributions

JJ, MC and MB conceptualised the study design. MH and GB collected the melancholia dataset. JJ performed data analysis and drafted the manuscript. JJ, MC, MH, GP, and MB edited the manuscript and approved the final version for submission.

Competing Interests statement

We declare no competing interests.

Key words:

Facial expression, facial affect, naturalistic, FACS, hidden Markov model, major depressive disorder, melancholic depression

Abstract

Facial affect is expressed dynamically – a giggle, grimace, or an agitated frown. However, the characterization of human affect has relied almost exclusively on static images. This approach cannot capture the nuances of human communication or support the naturalistic assessment of affective disorders. Using the latest in machine vision and systems modelling, we studied dynamic facial expressions of people viewing emotionally salient film clips. We found that the apparent complexity of dynamic facial expressions can be captured by a small number of simple spatiotemporal states - composites of distinct facial actions, each expressed with a unique spectral fingerprint. Sequential expression of these states is common across individuals viewing the same film stimuli but varies in those with the melancholic subtype of major depressive disorder. This approach provides a platform for translational research, capturing dynamic facial expressions under naturalistic conditions and enabling new quantitative tools for the study of affective disorders and related mental illnesses.

Introduction

Facial expressions are critical to interpersonal communication and offer a nuanced, dynamic and context-dependent insight into internal mental states. Humans use facial affect to infer personality, intentions and emotions, and it is an important component of the clinical assessment of psychiatric illness. For these reasons, there has been significant interest in the objective analysis of facial affect(1–3). However, decisive techniques for quantifying facial affect under naturalistic conditions remain elusive.

A traditional approach is to count the occurrences of a discrete list of "universal basic emotions"(4). While the most commonly used system designates six basic emotions, there is disagreement about the number and nature of such affective "natural forms" (5,6). Quantifying facial affect using the Facial Action Coding System (FACS), has become the dominant technique to operationalise facial expressions (7). Action units, each corresponding to an anatomical facial muscle group, are rated on a quantitative scale. Traditional emotion labels are associated with the co-occurrence of a specific set of action units – for example a "happy" facial expression corresponds to action units "Cheek Raiser" and "Lip Corner Puller"(8). However, due to the time-intensive nature of manually coding every frame in a video, FACS has traditionally been applied to the analysis of static pictures rather than videos of human faces.

Recent developments in machine learning have automated the identification of basic emotions and facial action units from images and videos of human faces. Feature extraction for images include local textures (9,10) and 3D geometry(11,12), while video analysis benefits from temporal features such as optical flow(13). Supervised learning algorithms classifying facial expressions based on feature values have achieved impressive accuracies benchmarked to manually coded datasets (see 14 for a review).

Videos of faces can now be reliably transformed into action unit time series which capture the rich temporal dynamics of facial expressions(11). This is important because human faces express emotional states dynamically – such as in a giggle or a sob. However, the rich potential of these temporal dynamics has not yet been fully exploited in the psychological and

behavioural sciences. For example, some psychological studies and databases have asked responders to pose discrete emotions such as happiness or sadness(15). This strategy suits the needs of a classic factorial experimental design but fails to produce the natural dynamics of real-world facial expressions. To evoke dynamic emotion, clinical interviews have been used(16), or participants have been asked to narrate an emotive story or been shown emotive pictures rather than videos(17). Such pictures can be grouped into distinct categories and presented repetitively in a trial structure, but their ecological validity is unclear. Consequently, there is an expanding interest in naturalistic video stimuli such as movie clips(18–20). These are more ecologically valid, have greater test-retest reliability than interviews, evoke stronger facial expressions than static pictures, and produce stronger cortical responses during functional neuroimaging(3,21,22). However, interpreting the facial expressions resulting from naturalistic stimulus viewing poses challenges, because each time point is unique. There is currently no obvious way to parse the stimulus video into discrete temporal segments. Naïve attempts at dimensionality reduction – for example, averaging action unit activations across time – omit temporal dynamics and so fail to capture the complexity of natural responses.

Disturbances in facial affect occur across a range of mental health disorders, including major depressive disorder, schizophrenia, and dementia. Capturing the nuances of facial affect is a crucial skill in clinical psychiatry but in the absence of quantitative tests this remains dependent on clinical opinion. Supervised learning has shown promise in distinguishing people with major depression from controls, using input features such as facial action units coded manually(23) or automatically(24), or model-agnostic representations of facial movements such as the ‘Bag of Words’ approach(25,26). Studies documenting action unit occurrence during the course of a naturalistic stimulus(27), a short speech(28) or a clinical interview(29), have demonstrated that depression is associated with reduced frequency of emotional expressions, particularly expressions with positive valence. Unfortunately, by averaging action unit occurrence over time, these methods poorly operationalise the clinician’s gestalt sense of affective reactivity, which derive from a patient’s facial responses across a range of contexts.

Here, we present a novel pipeline for processing facial expression data recorded while participants view a dynamic naturalistic stimulus. The approach is data-driven, eschewing the need to preselect emotion categories or segments of the stimulus video. We derive a time-

frequency representation of facial movement information, on the basis that facial movements in vivo are fundamentally dynamic and multiscale. These time-frequency representations are then divided into discrete packets with a hidden Markov model (HMM), a method for inferring hidden states and their transitions from noisy observations. We find dynamic patterns of facial behaviour which are expressed sequentially and localised to specific action units and frequency bands. These patterns are context-dependent, consistent across participants, and correspond to intuitive concepts such as giggling and grimacing. We first demonstrate the validity of this approach on an open-source dataset of facial responses of healthy adults watching naturalistic stimuli. We then test this approach on facial videos of participants with melancholic depression, a severe mood disorder characterised by psychomotor changes(30). We show that dynamic facial patterns reveal specific changes in melancholia, including reduced facial activity in response to emotional stimuli, anomalous facial responses inconsistent with the affective context, and a tendency to get “stuck” in negatively valenced states. Moreover, using these decoded patterns improves accuracy in classifying patients from healthy controls.

Results

We first analysed dynamic facial expressions from video recordings of 27 participants viewing short emotive clips of 4 minute duration, covering a breadth of basic emotions (the DISFA dataset(19), detailed in Supplementary Table 1). Frame-by-frame action unit activations were extracted with OpenFace software(31) (see Supplementary Table 2 for action unit descriptions).

From these data, we used the continuous wavelet transform to extract a time-frequency representation of individual action unit time series in each participant. To test whether this time-frequency representation captures high frequency dynamic content, we first compared the group average of these individual time-frequency representations with the time-frequency representation of the group mean time series. We selected the activity of action unit 12 "Lip Corner Puller" during a positive valence video clip (a 'talking dog'), as this action unit is conventionally associated with happy affect, and its high frequency activity denotes smiling or laughing. Compared to the group mean time series, the time series of individuals had significantly greater amplitude (Figure 1a), particularly at higher frequencies (Figure 1e). This demonstrates that the time-frequency representations of individual participants capture high frequency dynamics that are obscured by characterising group-averaged time courses. This is because stimulus-evoked facial action unit responses have asynchronous alignment across participants, hence cancelling when superimposed. This problem is avoided in the group-level time-frequency representation, whereby the amplitude is first extracted at the individual level, prior to group averaging. Comparable results occurred in all action units (Supplementary Figure 2).

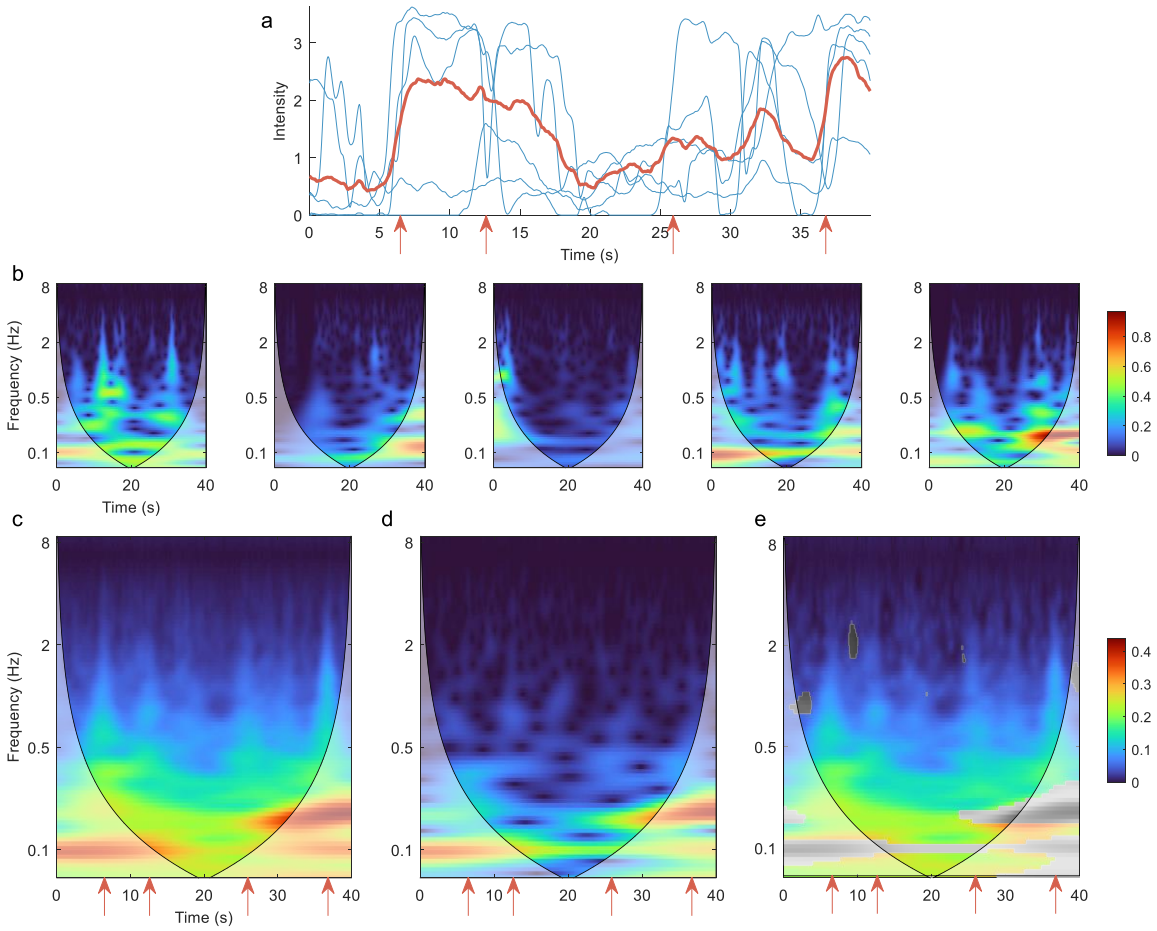

Figure 1. Time-frequency representation of action unit 12 "Lip Corner Puller" during positive valence video stimulus reveals high frequency dynamics. (a) Action unit time series for 5 example participants (blue). The group mean time course across all participants is shown in red. Red arrows indicate funny moments in the stimulus, evoking sudden facial changes in individual participants. These changes are less prominent in the group mean time course. (b) Time-frequency representation for the same 5 participants, calculated as the amplitude of the continuous wavelet transform. Shading indicates the cone of influence – the region contaminated by edge effects. (c) Mean of all participants' time-frequency representations. (d) Time-frequency representation of the group mean time course. Red arrows correspond to time points with marked high frequency activity above 1 Hz. (e) Difference between (c) and (d). Non-significant differences ($p > 0.05$) are shown in greyscale. Common colour scale is used for (c)-(e).

Having shown how the time-frequency representation captures dynamic content, we next sought to quantify the joint dynamics of facial action units. To this end, a hidden Markov model (HMM) was inferred from the time course of all facial action units' time-frequency representations. A HMM infers a set of distinct states from noisy observations, with each state expressed sequentially in time according to state-to-state transition probabilities. Each state has a distinct mapping onto the input space, here the space of frequency bands and action units (Figure 2a). Examples of participants in each state are provided in the Supplementary Videos. Their occurrence corresponded strongly with annotated video clip valence (Figure 2b). We found that inferred state sequences had high between-subject consistency, exceeding chance level across the vast majority of time points and reaching 93% during specific movie events (Figure 2d). States were frequency-localised and comprised intuitive combinations of action units which reflected not only distinct emotion categories as defined in previous literature(8), but also stimulus properties such as mixed emotions. State transition probabilities appeared clustered by valence rather than frequency, such that frequent transitions between low and high frequency oscillations of the same facial action units were more likely than transitions between different emotions (Figure 2e).

- States 1 and 2 were active during stimuli annotated as "happy". They activated two action units typically associated with happiness, action unit 6 "Cheek Raiser" and 12 "Lip Corner Puller", but also action unit 25 "Lips Part". State 2 likely represents laughing or giggling as it encompassed high frequency oscillations in positive valence action units, in comparison to the low frequency content of State 1.
- States 3 and 4 were active during videos evoking fear and disgust – for example of a man eating a beetle larva. They encompassed mixtures of action units conventionally implicated in disgust and fear, at low and high frequency bands respectively. State 3 recruited action units 4 "Brow Lowerer" and 9 "Nose Wrinkler", while state 4 involved these action units as well as action units 15 "Lip Corner Depressor", 17 "Chin Raiser", and 20 "Lip Stretcher".
- States 5 and 6 occurred predominantly during negatively valenced clips, and deactivated oscillatory activity in most action units, with sparing of action units typically associated with sadness, 4 "Brow Lowerer" and 15 "Lip Corner Depressor".

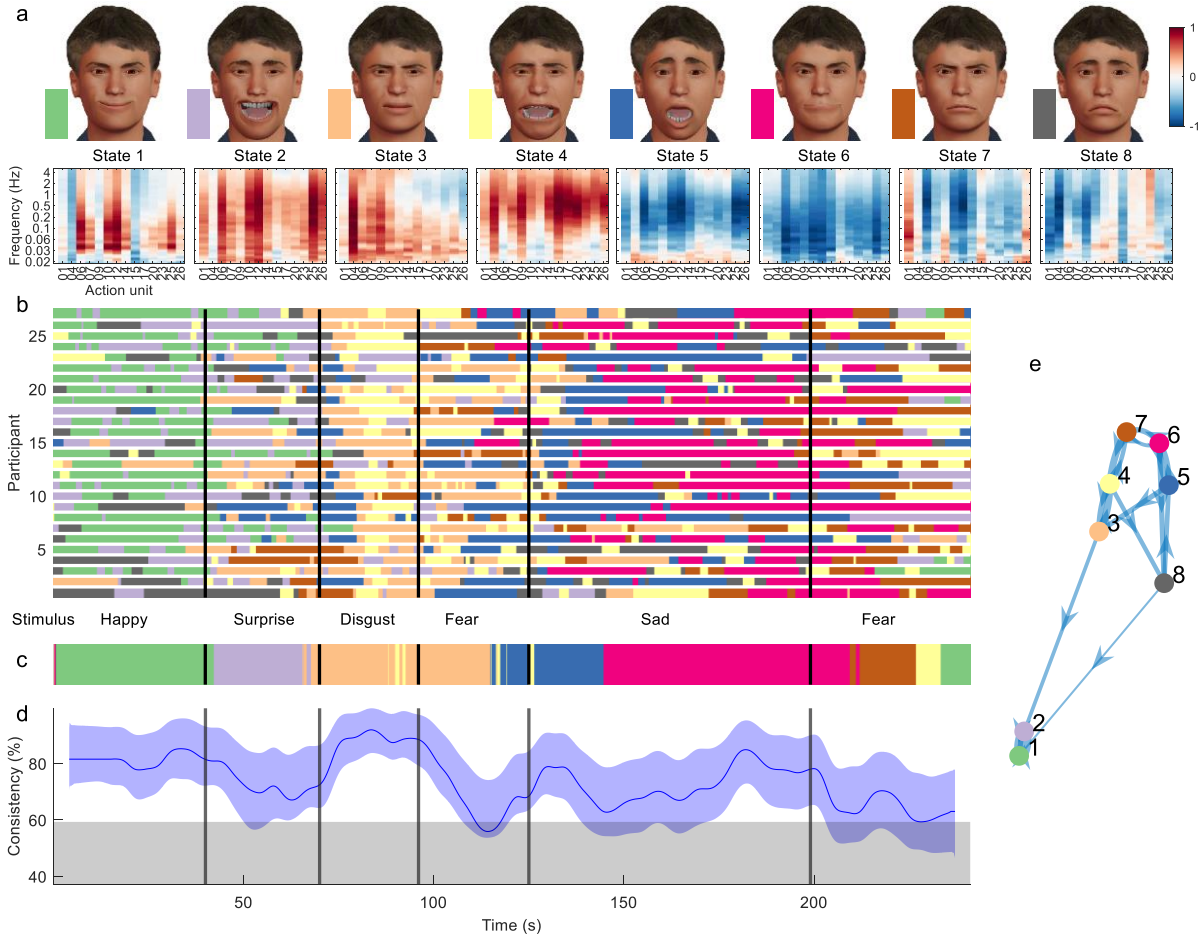

Figure 2. Dynamic facial states inferred from time-frequency representation of DISFA dataset. (a) Mean of the observation model for each state, showing their mapping onto action units and frequency bands. Avatar faces (top row) for each state show the relative contribution of each action unit, whereas their spectral projection (bottom row) shows their corresponding dynamic content. (b) Sequence of most likely states for each participant at each time point. Vertical lines demarcate transition between stimulus clips with different affective annotations. (c) Most common states across participants, using a 4s sliding temporal window. (d) Proportion of participants expressing the most common state. Blue shading indicates 5% - 95% bootstrap confidence bands for the estimate. Grey shading indicates the 95th percentile for the null distribution, estimated using time-shifted surrogate data. (e) Transition probabilities displayed as a weighted graph. Each node corresponds to a state. Arrow thickness indicates the transition probability between states. For visualization clarity, only the top 20% of transition probabilities are shown. States are positioned according to a force-directed layout where edge length is the inverse of the transition probability.

<u>Facial affect in melancholia</u>

We next analysed facial video recordings from a cohort of participants with melancholic depression and healthy controls who watched three video clips consecutively – a stand-up

comedy, a sad movie clip and an amusing video (weather report) in a non-English language (German). These three stimuli were chosen from a database of independently rated videos of high salience(32). The stand-up comedy comprises episodic jokes with a deadpan delivery and audience laughter, whereas the weather report depicts someone laughing uncontrollably and spontaneously. Clinical participants with melancholia were recruited from a tertiary mood disorders clinic and met melancholia criteria including psychomotor changes, anhedonia and diurnal mood variation (see Methods). We conducted analyses based firstly on group-averaged time courses, and then on the time-frequency representation.

Group time courses in melancholia

Facial action unit time courses showed clear group differences (see Figure 3 for action units typically implicated in expressing happiness and sadness, and Supplementary Figure 5 for all action units). For each action unit in each participant, we calculated the median action unit activation across each stimulus video. These were compared with a 3-way ANOVA, with factors for clinical group, stimulus, and the facial valence. We considered two stimulus videos, one with positive and one with negative valence, and two facial valence states, happiness and sadness, calculated as sums of positively and negatively valenced action unit activations respectively(8). A significant 3-way interaction was found between clinical group, stimulus, and facial valence (p=0.003). Post-hoc comparisons with Tukey's honestly significant difference criterion (Supplementary Figure 4) quantified that during stand-up comedy, participants with melancholia had reduced activation of action unit 12 "Lip Corner Puller" (p<0.0001) and increased activation of action unit 4 "Brow Lowerer" (p<0.0001). Interestingly, facial responses of participants with melancholia during stand-up comedy, were similar to those of controls during the sad movie ($p > 0.05$ for both action units).

To move away from individual action units, we next extracted the first principal component across all action units. The time course of this composite component closely followed joke punch lines during stand-up comedy (Figure 3b). This responsivity of this component to movie events was substantially diminished in the melancholia cohort

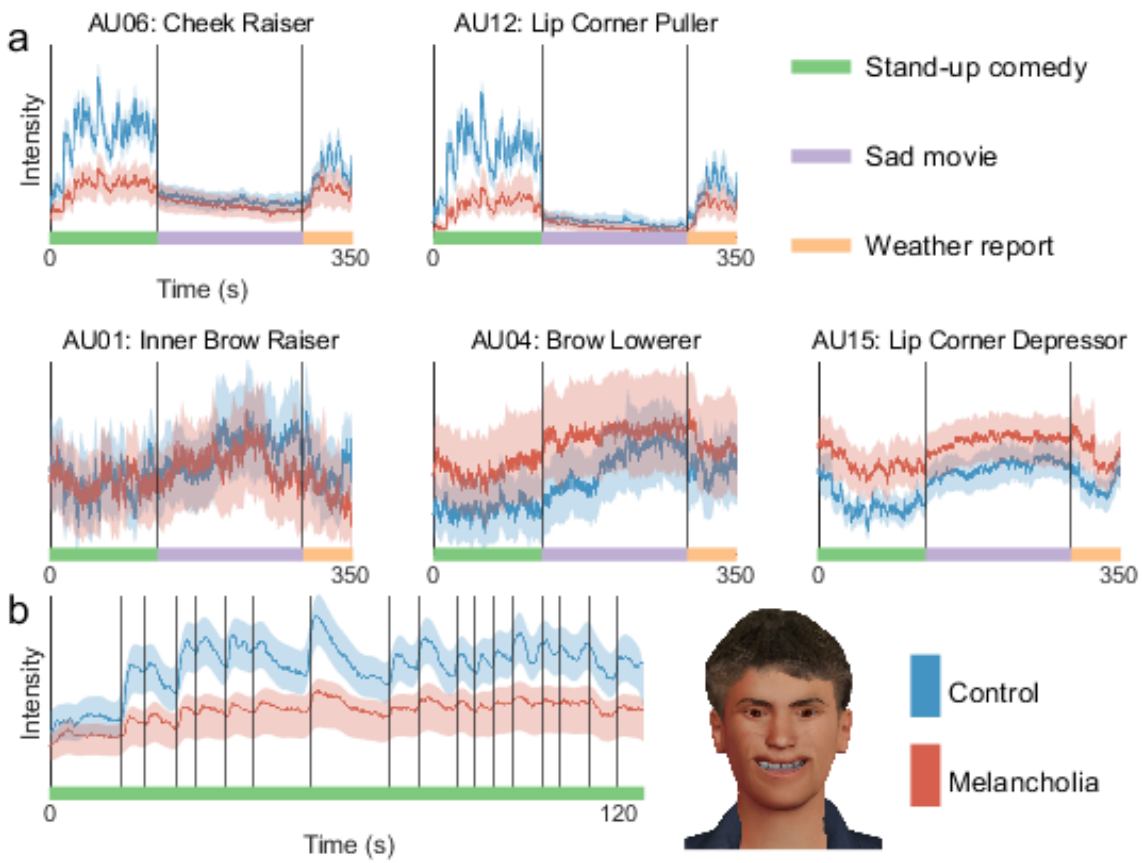

Figure 3. At each time point, mean intensity across participants of facial action unit activation in controls (blue) and melancholia (red). Shading indicates 5% and 95% confidence bands based on a bootstrap sample (n=100). (a) Action units commonly implicated in happiness (top row) and sadness (bottom row). Participants watched stand-up comedy, a sad video, and a funny video in sequence. Vertical lines demarcate transitions between video clips. (b) First principal component of action units, shown during stand-up comedy alone. Vertical lines indicate joke annotations. Avatar face shows the relative contribution of each action unit to this component.

<u>Time-frequency representation in melancholia</u>

Time-frequency representations were calculated for all action units in all participants. For each action unit, the mean time-frequency representation for the control group was subtracted from the participants with melancholia (see Supplementary Figure 6 for the mean of the controls). Significant group differences ($p < 0.05$) were found by comparison to a null distribution composed of 100 resampled surrogate datasets (see Methods). Participants with melancholia had a complex pattern of reduced activity encompassing a broad range of frequencies (Figure 4a). The most prominent differences were in positive valence action units during positive valence stimuli, but significant reductions were seen in most action units.

Differences in high frequency bands occurred during specific movie events such as jokes (Figure 4b). There were sporadic instances of increased activity in melancholia participants during the sad movie involving mainly action units 15 “Lip corner depressor” and 20 “Lip stretcher”.

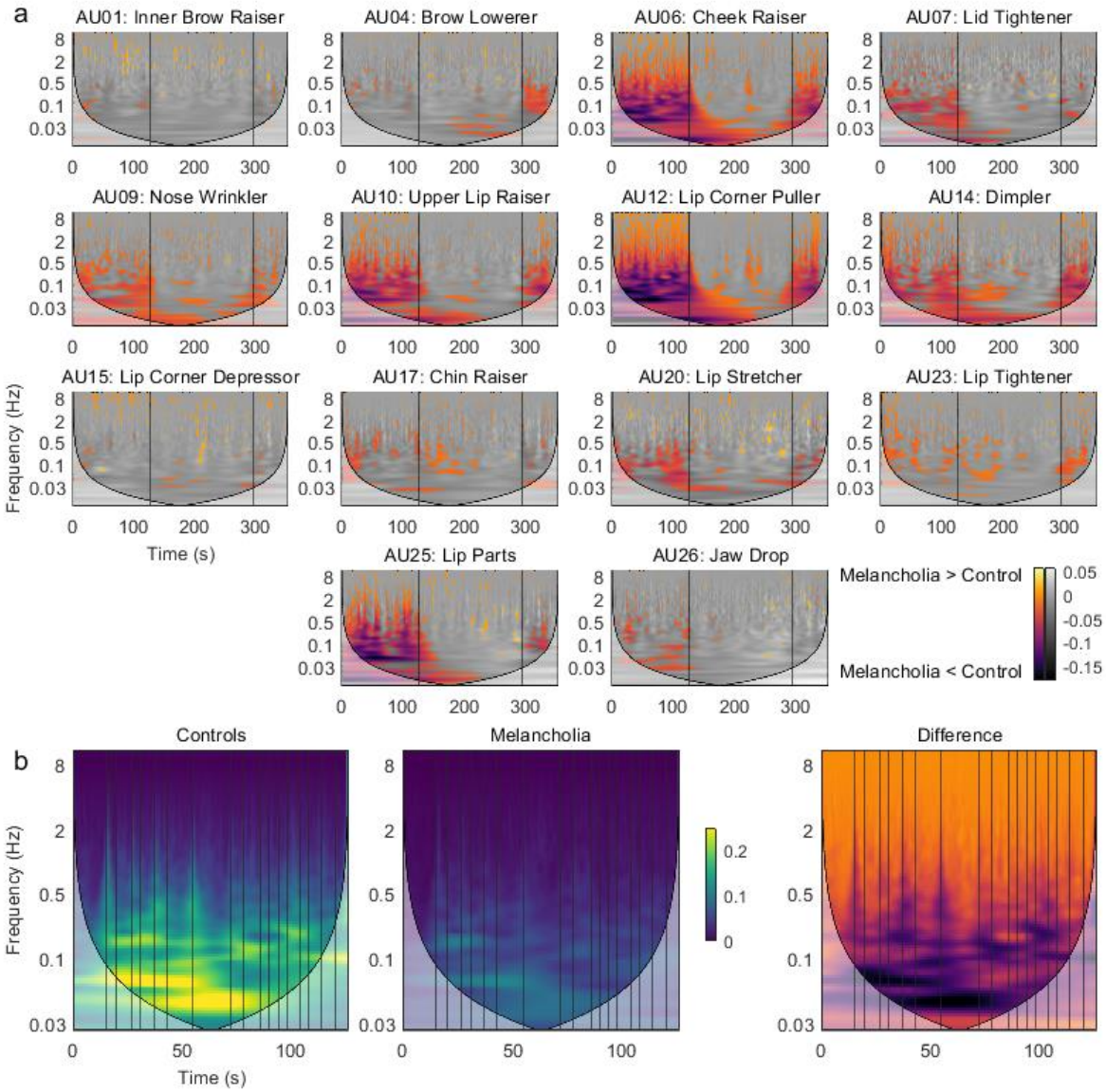

Figure 4. (a) Mean time-frequency activity in melancholia benchmarked to the control group. Negative colour values (red-purple) indicate melancholia < controls ($p < 0.05$). Non-significant group differences ($p > 0.05$) are indicated in greyscale. Vertical lines demarcate stimulus videos. (b) Action unit 12 “Lip Corner Puller” during stand-up comedy in controls, participants with melancholia, and difference between groups. Vertical lines indicate joke annotations.

We next pursued whether the additional time-frequency information would improve the classification accuracy of differentiating participants with melancholia from controls. A support vector machine, using as inputs the mean action unit activation for each stimulus video, achieved 63% accuracy with 5-fold cross-validation. In contrast, using as inputs the mean time-frequency amplitude in discrete frequency bands within 0 – 5 Hz, improved average cross-validation accuracy to 71%. As a control for the additional number of input features, we tested a third set of models which naively modelled temporal dynamics using mean action unit activations within shorter time blocks. These models had 63 – 64% accuracy despite having a greater number of input features than the time-frequency representation (Supplementary Table 3).

Sequential affective states in melancholia

Inverting a HMM from the time-frequency representations of facial action units yielded the sequential expression of 8 states across participants (Figure 5).

- States 1 and 2 activated positive valence action units, each in distinct frequency bands, and were dominant through the stand-up comedy for most participants (Figure 5B). State 2 comprised high frequency oscillations in positive valence action units, corresponding to laughing or giggling.
- The sad movie was associated with early involvement of state 3, which deactivated high-frequency activity, followed by states 4 and 5, which also deactivated oscillatory activity, but with more specificity for lower frequencies and positive valence action units.
- State 6 comprised action units 4 “Brow Lowerer”, 9 “Nose Wrinkler”, 17 “Chin Raiser”, and 23 “Lip Tightener”, traditionally associated with anger, disgust, or concern. State 7 can be associated with “gasping”, with very high frequency activation of most mouth-associated action units including 25 “Lips Part”. These states occurred sporadically through the weather report.
- State 8 predominantly activated action unit 1 “Inner Brow Raiser”, commonly associated with negative valence.

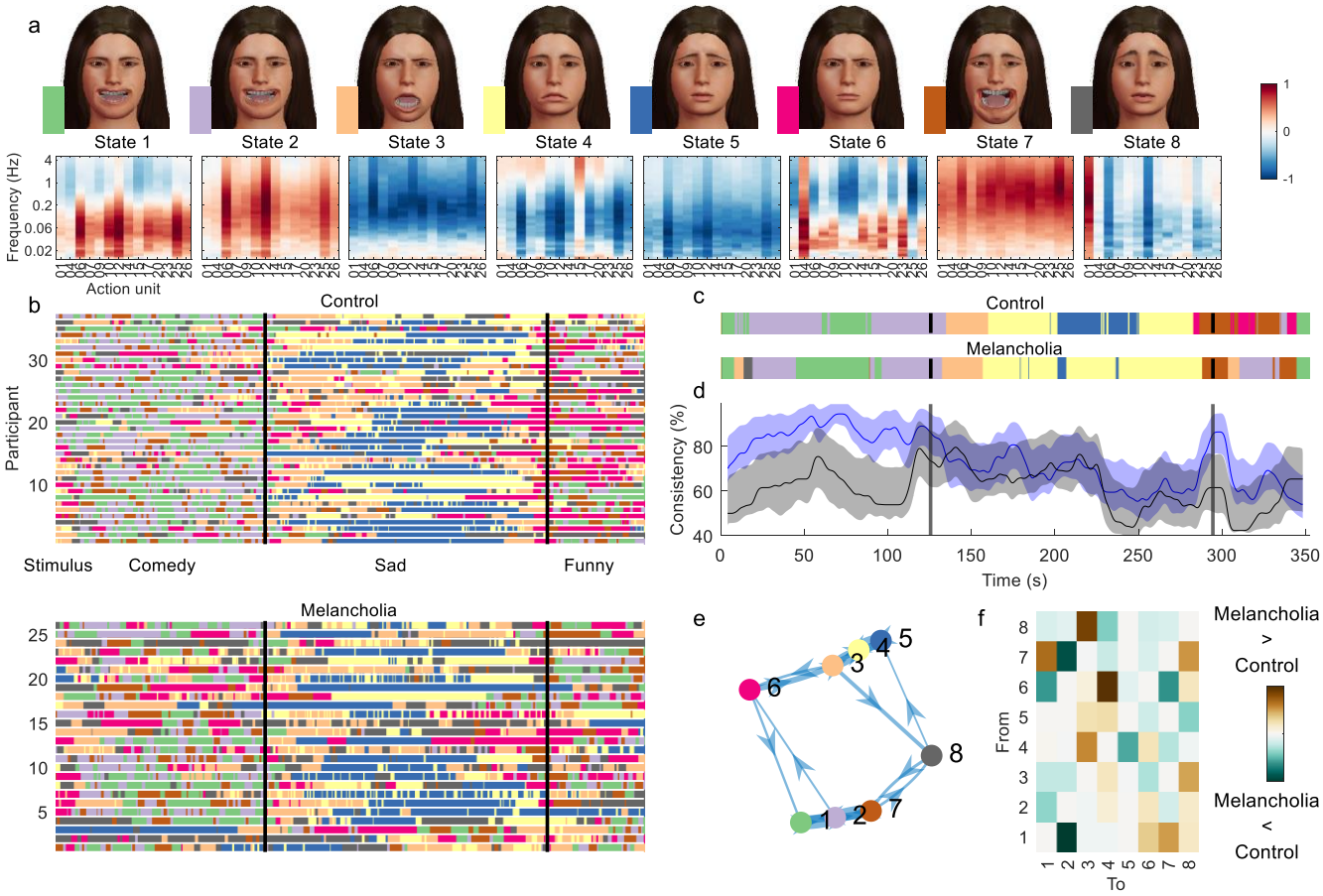

Figure 5. Hidden Markov model inferred from time-frequency representation of melancholia dataset. (a) Contribution of action units and their spectral expression to each state. Avatar faces for each state show the relative contribution of each action unit. (b) State sequence for each participant at each time point, for controls (top) and participants with melancholia (bottom). Vertical lines demarcate stimulus clips. (c) Most common state across participants, using a 4s sliding temporal window. (d) Proportion of participants expressing the most common state for control (blue) and melancholia cohorts (black). Shading indicates 5% and 95% bootstrap confidence bands. (e) Transition probabilities displayed as a weighted graph, with the top 20% of transition probabilities shown. States are positioned according to a force-directed layout where edge length is the inverse of transition probability. (f) Differences in mean transition probabilities between participants with melancholia and controls. Each row/column represents a HMM state. Colours indicate (melancholia – controls) values

The temporal sequence of most common states was similar across groups (Figure 5C), but the between-subjects consistency was markedly reduced in the melancholic participants during both funny videos (Figure 5D). Some participants with melancholia - for example participants 2 and 3 (Figure 5B) - had highly anomalous state sequences compared to other participants.

Fractional occupancy – the proportion of time spent by participants in each state – was significantly different between groups for the positive valence states - state 1 (Melancholia < Controls, p_{FDR}=0.03) and state 2 (Melancholia < Controls, p_{FDR}=0.004) – as well as for negatively valenced state 8 (Melancholia > Controls p_{FDR}=0.03). We then asked whether group differences in the time spent in each state were attributable to changes in the likelihood of switching in to, or out of, specific facial states. Participants with melancholia were significantly less likely to switch from a low-frequency positive valence state (1, smiling) to high-frequency positive valence oscillations (state 2, giggling), but were more likely to switch to states associated with any other emotion (states 4, 5, 6, 7, and 8). From the high frequency positive valence state, they were more likely to switch to the deactivating "ennui" state 4 (all $p_{FDR} < 0.05$).

Discussion

Facial expressions played a crucial role in the evolution of social intelligence in primates(2) and continue to mediate human interactions. Observations of facial affect, its range and reactivity play a central role in clinical settings. Quantitative analysis of facial expression has accelerated of late, driven by methods to automatically factorise expressions into action units(7) and the availability of large datasets of posed emotions(15). The dynamics of facial expression mediate emotional reciprocity, but have received less attention(3). Naturalistic stimuli offer distinct advantages to affective research for their ability to evoke these dynamic responses(33), but their incompressibility has made analysis problematic. By leveraging techniques in computer vision, we developed a pipeline to characterise facial dynamics during naturalistic video stimuli. Analysis of healthy adults watching emotionally salient videos showed that facial expression dynamics can be captured by a small number of spatiotemporal states. These states co-activate facial muscle groups with a distinct spectral fingerprint, and transition dynamically with the emotional context. Application of this approach to melancholia showed that the clinical gestalt of facial non-reactivity in melancholia(34) can be objectively identified not just with restrictions in spectral content, but also with anomalous facial responses, more frequent occurrence of an ennui affect, and more frequent state switching from transiently positive facial expressions to neutral and negative states. This approach provides a unique perspective on how facial affect is generated by the interplay between inner affective states and the sensorium.

Our pipeline first comprises automatic action unit extraction, then spectral wavelet-based analysis of the ensuing feature dynamics. Wavelet energy at a given time corresponds to the occurrence of a specific facial "event", while energy in a given frequency reflects the associated facial dynamics, like laughing. Unlike temporal averaging methods, which require an arbitrary timescale, wavelets cover a range of timescales. The spectral approach also allows participant facial responses to be pooled, without the limitations of averaging responses whose phases are misaligned. We then inferred a hidden Markov model, identifying spatially and spectrally resolved modes of dynamic facial activity which occur sequentially with high consistency across participants viewing the same stimulus. States transitions aligned with intuitive notions of affective transitions – for example, the common transition between the low frequency and high frequency positive valence state, reflected transitions between smiling and laughing.

Our method builds on the Emotion Facial Action Coding System (EMFACS)(8), where each canonical emotion label (happy, angry, etc) is defined on the basis of a sparse set of minimally necessary action units. The sparsity of this coding allows manual raters to find the minimal necessary combinations of action units in a facial video to reflect an emotion label, but may not include all action units that are involved in each affective state. Affective states inferred from our HMM reflected prototypical action unit combinations from EMFACS, but also provide a richer mapping across a broader range of action units. For example, while happiness has been previously associated with just two action units, "Cheek Raiser" and "Lip Corner Puller", such sparse activations are rare, particularly during intense emotional displays. We demonstrated that laughing during stand-up comedy activated eyebrow-related action units, some of which are traditionally associated with sadness. Conversely, negatively valenced stimuli dampened facial movements, with a relative sparing of those action units typically associated with sadness.

Ensembles of HMMs have previously been used to improve emotion classification accuracy when benchmarked against manually coded datasets. In these studies, one HMM models the temporal dynamics of one action unit(35,36) or one universal basic emotion(13,37), with HMM states corresponding to expression onset/offset. Given a video frame, the HMM with the greatest evidence determines the decoded expression. Nested HMMs have also been employed, with a second level HMM predicted transitions between the basic emotions(38). In

contrast, the present method uses a single HMM to describe facial expressions without prior emotion categories, capturing the dynamic co-occurrence of facial actions that together comprise distinct affective states. By taking the spectral activity of action units as input features into the HMM, our approach uniquely captures the spatiotemporal texture of naturally occurring facial affect. This enables, for example, the disambiguation of a smile from a giggle. The importance of the spectral characterization is highlighted by our finding that in melancholia, smile states were more likely to transition to ennui, and less likely to the laughter state. Our use of dynamic spectra as inputs into a HMM is similar to their recent use in neuroimaging research(39). Using the raw time series is also possible – hence additionally capturing phase relationships, although this comes with an additional computational burden and reduced interpretability of states(40).

Dynamic facial patterns were influenced by the affective properties of the stimulus video. For the DISFA dataset, the HMM inferred two disgust-associated states, in low and high frequency bands respectively. These states occurred predominantly during two disgusting video clips. For the melancholia dataset, the inferred HMM states over-represented happiness and sadness, and under-represented disgust. This is ostensibly because the stimulus had prominent positive and negatively valenced sections without disgusting content. The co-occurrence of the states and the state transitions across participants speaks to the influence of the video content on affective responses and hence, more broadly, the dynamic exchange between facial affect and the social environment.

We found that participants with melancholia exhibited broad reductions in facial activity, as well as specific reductions in high frequency activity in response to specific events such as joke punchlines, reflecting the clinical gestalt of impaired affective reactivity(30). Viewing affect as a dynamic process provided two further insights into facial responses in melancholia. First, decreased between-subject consistency and more anomalous facial responses suggest that their facial activity is less likely to be driven by a common external stimulus. Ambiguous facial responses are also seen in schizophrenia(41), suggesting the possibility of a common underlying mechanism with melancholia. Second, participants with melancholia were less likely to enter high frequency positive valence states like laughing, and once there, transitioned out quickly to the “ennui” state. This reflects the clinical impression that positive mood states persist in healthy controls, but such states are fleeting in those with melancholia, who tend to get “stuck” in negative mood states instead. The results are

commensurate with the proposal that depressed states relate to persistent firing in non-reward functional areas mediated by attractor dynamics(42). Additionally, these findings accord with neurobiological models of melancholia whereby dysfunctional cortical-basal ganglia circuitry underlie the disturbances in volition and psychomotor activity that characterise the disorder(30). More generally, the notion of affect as a sequence of spatiotemporal states aligns with the proposal that instabilities in brain network activity generate adaptive fluctuations in mood and affect, with these being either over- or under-damped in affective disorders(43). Our paradigm also raises clinical questions predicated on dynamics – for example, do biological or psychological treatments for melancholia work by increasing the probability of entering positive affective states, or reducing the probability of exiting such states?

Several caveats bear mention. First, a small number of participants with constant zero activation in one or more action units were excluded from analysis, because this produces an ill-defined spectral transform. Excluded participants, of whom 1 was a control and 4 had melancholia, may have had the greatest impairments in facial affect. This issue could be addressed with a lower detectable limit of action unit activation. Second, time-frequency maps were standardised in mean and variance before HMM inference. This ensures that states occur sequentially across time, but reduces the differences in state sequences across groups. Omitting this standardisation step yields states that are biased towards group differences rather than temporal differences (see Supplementary Figure 7). Future work could consider methods that are less susceptible to this trade-off. Finally, the utility of our approach is likely to be improved by multimodal fusion of facial, head pose, vocal and body language behaviour, each of which independently improve classification(44–47).

Human emotion and affect are inherently dynamic. Our work demonstrates that momentary affective responses, such as laughing or grimacing, traditionally viewed from a qualitative standpoint, can be understood within a quantitative framework. These tools provide a translational platform for mental health research to understand the dynamics of facial affect - for example in clinical states such as melancholia with its distinctive sign of psychomotor disturbance, the masked facies of Parkinson’s disease, emotional incongruence and affective blunting in schizophrenia, and emotional lability integral to bipolar disorder.

Materials and Methods

Data

The Denver Intensity of Spontaneous Facial Action (DISFA) dataset contains facial videos recorded at 20 frames per second from 27 participants who viewed a 4 minute video consisting of short emotive clips from Youtube(19) (Supplementary Table 1).

The melancholia dataset comprises 30 participants with major depressive disorder, who were recruited from the specialist depression clinic at the Black Dog Institute in Sydney, Australia. These participants met criteria for a current major depressive episode, were diagnosed as having the melancholic subtype by previously detailed criteria(48), and did not have lifetime (hypo)mania or psychosis (Table 1). 38 matched healthy controls were recruited from the community. All participants were screened for psychotic and mood conditions with the Mini International Neuropsychiatric Interview (MINI). Exclusion criteria were current or past substance dependence, recent electroconvulsive therapy, neurological disorder, brain injury, invasive neurosurgery, or an estimated full scale IQ score (WAIS-III) below 80. Participants provided informed consent for the study. Participants watched 3 video clips consecutively – stand-up comedy (120 seconds), a sad movie clip (152 seconds), and a German weather report video depicting a weather reporter laughing uncontrollably (56 seconds). Facial video was recorded at a resolution of 800 x 600 pixels at 25 frames per second using an AVT Pike F-100 FireWire camera. The camera was mounted on a tripod, which was placed behind the monitor so as to record the front of the face. The height of the camera was adjusted with respect to the participant's height when seated.

Table 1. Demographics and clinical characteristics

	Healthy controls	Melancholia	Group comparison, *t* or χ^2, p-value
Number of participants	38	30	-
Age, mean (SD)	46.5 (20.0)	46.2 (15.5)	0.95

Sex (M:F)	13:19	17:13	0.21
	Medication, % yes (n)		
Any psychiatric medication	7% (1)	85% (23)	-
Nil medication	93% (13)	15% (4)	-
Selective serotonin reuptake inhibitor	7% (1)	15% (4)	-
Dual-action antidepressant[a]	0% (0)	48% (13)	-
Tricyclic or monoamine oxidase inhibitor	0% (0)	19% (5)	-
Mood stabilizer[b]	0% (0)	11% (3)	-
Antipsychotic	0% (0)	33% (9)	-

[a] For example, serotonin noradrenaline reuptake inhibitor
[b] For example, lithium or valproate

Facial action units

For the melancholia dataset, facial video recordings of different participants were aligned with FaceSync(49). For both datasets, facial action unit intensities were extracted with OpenFace(31). OpenFace uses a convolutional neural network architecture, Convolutional Experts Constrained Local Model (CE-CLM), to detect and track facial landmark points. After face images are aligned to a common 112 x 112 pixel image, histogram of oriented gradients features are extracted. A linear kernel support vector machine was then trained on 6 facial expression datasets with manually coded action unit occurrence times.

Action unit time series from OpenFace for each participant were not normalised, as we were interested in between-subjects differences. Recordings with more than 0.5% missing frames were excluded, and any remaining missing frames were linearly interpolated. Action unit 45 “Blink” was not used as it is not directly relevant to emotion. Action units 2 “Outer Brow Raiser” and 5 “Upper Lid Raiser” were not used as they had constant zero value throughout the recording for most participants. Participants with any other action units with zero value through the recording were also excluded, as the time-frequency representation is undefined for these time series. This comprised 1 control and 4 participants with melancholia.

Time-frequency representation

For each participant, each facial action unit time series was transformed into a time-frequency representation, using the amplitude of the continuous wavelet transform. An analytic Morse wavelet was used with symmetry parameter 3, time-bandwidth product 60, and 12 voices per octave. Mean time-frequency maps were visualised with a cone of influence – outside which edge effects produce artefact (Supplementary Figure 2 for DISFA, Supplementary Figure 6 for melancholia dataset). To determine information lost by averaging raw time series across participants, the amplitude of the continuous wavelet transform for the group mean time series was calculated. At each point in time-frequency space, the distribution of individual participants' amplitude was compared with the amplitude of the group mean, with a two-sided t-test (p=0.05) (Figure 1).

Hidden Markov model

A Hidden Markov model (HMM), implemented in the HMM-MAR MATLAB toolbox (https://github.com/OHBA-analysis/HMM-MAR)(50), was used to identify states corresponding to oscillatory activity localised to specific action units and frequency bands. A HMM specifies state switching probabilities which arise from a time-invariant transition matrix. Each state is described by a multivariate Gaussian observation model with distinct mean and covariance in (action unit x frequency) space. Input data were 110 frequency bins in 0-5Hz, for each of 14 facial action units. Individual participants' time series were standardised to zero mean and unit variance before temporal concatenation to form a single time series. This time series was downsampled to 10Hz, and the top 10 principal components were used (for DISFA). Other HMM parameters are listed in Supplementary Table 4.

The initialisation algorithm used 10 optimisation cycles per repetition. Variational model inference optimised free energy, a measure of model accuracy penalised by model complexity, and stopped after the relative decrement in free energy dropped below 10^{-5}. Free energy did not reach a minimum even beyond n=30 states (Supplementary Figure 3). Previous studies have chosen between 5 and 12 states(51,52). We chose an 8-state model as done in previous work(39), as visual inspection of the states showed trivial splitting of states beyond this value. However, the analyses were robust to variations in the exact number of states.

HMM state observation models were visualised with FACSHuman(53). The contribution of each action unit to each state was calculated by summing across all frequency bands. For each state, positive contributions were rescaled to the interval [0,1] and visualised on an avatar face (Figure 2a). State sequences for individual subjects were calculated with the Viterbi algorithm (Figure 2). To calculate between-subjects consistency of state sequences over time, we used an 8s sliding window. Within this window, for each state, we counted the number of participants who expressed this state at least once, and found the most commonly expressed state. Uncertainty in this consistency measure at each time point was estimated from the 5 and 95 percentiles of 100 bootstrap samples. The null distribution for consistency was obtained by randomly circular shifting the Viterbi time series for each subject independently (n=100). Consistency values exceeding the 95th percentile (59% consistency) were deemed significant.

Analysis of melancholia dataset

Mean action unit activations were calculated for each group, and uncertainty visualised with the 5th and 95th percentiles of 100 bootstrap samples (Figure 3, Supplementary Figure 5). A 3-way ANOVA for activation was conducted with group, stimulus video, and facial valence as regressors. To avoid redundancy between the two positive valence videos, we limited the ANOVA to two stimulus videos – the stand-up comedy and sad movie clips. In keeping with previous work(8), we defined happiness as the sum of action units 6 “Cheek Raiser” and 12 “Lip Corner Puller”, and sadness as the sum of action units 1 “Inner Brow Raiser”, 4 “Brow Lowerer”, and 15 “Lip Corner Depressor”. Post-hoc comparisons used Tukey’s honestly significant difference criterion (Supplementary Figure 4).

Time-frequency representations were computed as the amplitude of the continuous wavelet transform. Group differences in wavelet power, localised in time and frequency, were calculated by subtracting the mean time-frequency representation of each clinical group (Figure 4). To confirm that these effects were not due to movement-related noise in action unit encoding having different effects depending on the frequency and time window considered, the null distribution of the effect was obtained by resampling 100 surrogate cohorts from the list of all participants. Time-frequency points with effect size inside 2.5 – 97.5 percentile were considered non-significant and excluded from visualisation.

To compare classification accuracy with action unit time series or time-frequency data, a support vector machine with Gaussian kernel was used. All tests used mean accuracy over 5 repetitions of 5-fold cross validation, but varied in the input features. Inputs to the first model were mean action unit activations for each action unit (n=14) and each stimulus video (n=3). For the time-frequency model, inputs were mean wavelet amplitude in each frequency bin (n=10) in each stimulus video, for each action unit. For the third set of models, input features were mean action unit activation within discrete time chunks of 2, 10, and 30 seconds (Supplementary Table 3).

The HMM was inferred as described above (Figure 5). Supplementary Figure 7 shows the results when input data were not standardised. Local transition probabilities were then inferred for each participant separately. Two-sided significance testing for group differences in fractional occupancy was implemented within the HMM-MAR toolbox by permuting between subjects as described previously(54). Next, we considered only those state transitions that could explain the group differences in fractional occupancy and tested these transitions for group differences with t-tests (one-sided in the direction that could explain fractional occupancy findings). Group differences in fractional occupancy and transition probability were corrected to control the false discovery rate(55).

Results were consistent across repetitions of HMM inference with different initial random seeds. In addition, all analyses were repeated with time-frequency amplitudes normalised by the standard deviation of the time series, to ensure that results were not solely due to group differences in variance for each action unit time. This was motivated by previous work showing that the square of wavelet transform amplitude increases with variance for white noise sources(56). Results were consistent with and without normalisation, including differences between clinical groups, the distributions and time courses of HMM states.

Acknowledgements

JJ acknowledges the support of a Health Education & Training Institute Award in Psychiatry and Mental Health, and the Rainbow Foundation. MB acknowledges the support of the National Health and Medical Research Council (1118153, 10371296, 1095227) and the Australian Research Council (CE140100007).

Data availability

The DISFA dataset is publically available at http://mohammadmahoor.com/disfa/. The melancholia dataset is not publically available due to ethical and privacy considerations for patients.

Code availability

Code to replicate the analysis of healthy controls in the DISFA dataset is available at https://github.com/jaysonjeg/FacialDynamicsHMM

References

1. Naumann LP, Vazire S, Rentfrow PJ, Gosling SD. Personality Judgments Based on Physical Appearance. Pers Soc Psychol Bull. 2009 Dec 1;35(12):1661–71.

2. SCHMIDT KL, COHN JF. Human Facial Expressions as Adaptations:Evolutionary Questions in Facial Expression Research. Am J Phys Anthropol. 2001;Suppl 33:3–24.

3. Ambadar Z, Schooler JW, Cohn JF. Deciphering the Enigmatic Face: The Importance of Facial Dynamics in Interpreting Subtle Facial Expressions. Psychol Sci. 2005 May 1;16(5):403–10.

4. Ekman P. Are there basic emotions? Psychol Rev. 1992 Jul;99(3):550–3.

5. Ortony A. Are All 'Basic Emotions' Emotions? A Problem for the (Basic) Emotions Construct. Perspect Psychol Sci J Assoc Psychol Sci. 2022 Jan;17(1):41–61.

6. Keltner D, Sauter D, Tracy J, Cowen A. Emotional Expression: Advances in Basic Emotion Theory. J Nonverbal Behav. 2019 Jun;43(2):133–60.

7. Ekman P, Friesen WV, Friesen WV, Hager J. Facial action coding system: A technique for the measurement of facial movement. 1978 Jan 1 [cited 2020 Apr 15]; Available from: https://www.scienceopen.com/document?vid=759f6f74-7ccd-47b5-904a-25ca0f29ea90

8. Friesen WV, Ekman P. EMFACS-7: Emotional Facial Action Coding System. Version 7. 1983.

9. Feng X. Facial expression recognition based on local binary patterns and coarse-to-fine classification. In: The Fourth International Conference onComputer and Information Technology, 2004 CIT '04. 2004. p. 178–83.

10. Kumar P, Happy SL, Routray A. A real-time robust facial expression recognition system using HOG features. In: 2016 International Conference on Computing, Analytics and Security Trends (CAST). 2016. p. 289–93.

11. Tian Y li, Kanade T, Cohn JF. Recognizing Action Units for Facial Expression Analysis. IEEE Trans Pattern Anal Mach Intell. 2001 Feb;23(2):97–115.

12. Ghimire D, Lee J. Geometric Feature-Based Facial Expression Recognition in Image Sequences Using Multi-Class AdaBoost and Support Vector Machines. Sensors. 2013 Jun;13(6):7714–34.

13. Yeasin M, Bullot B, Sharma R. Recognition of facial expressions and measurement of levels of interest from video. IEEE Trans Multimed. 2006 Jun;8(3):500–8.

14. Ekundayo O, Viriri S. Facial Expression Recognition: A Review of Methods, Performances and Limitations. In: 2019 Conference on Information Communications Technology and Society (ICTAS). 2019. p. 1–6.

15. Lucey P, Cohn JF, Kanade T, Saragih J, Ambadar Z, Matthews I. The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops. 2010. p. 94–101.

16. Darzi A, Provenza NR, Jeni LA, Borton DA, Sheth SA, Goodman WK, et al. Facial Action Units and Head Dynamics in Longitudinal Interviews Reveal OCD and Depression severity and DBS Energy. In: 2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021). 2021. p. 1–6.

17. Lang P, Bradley M, Cuthbert B. International affective picture system (IAPS): Affective ratings of pictures and instruction manual. Technical Report A-8. University of Florida, Gainesville, FL; 2008.

18. Kollias D, Tzirakis P, Nicolaou MA, Papaioannou A, Zhao G, Schuller B, et al. Deep Affect Prediction in-the-Wild: Aff-Wild Database and Challenge, Deep Architectures, and Beyond. Int J Comput Vis. 2019 Jun 1;127(6):907–29.

19. Mavadati SM, Mahoor MH, Bartlett K, Trinh P, Cohn JF. DISFA: A Spontaneous Facial Action Intensity Database. IEEE Trans Affect Comput. 2013 Apr;4(2):151–60.

20. Soleymani M, Lichtenauer J, Pun T, Pantic M. A Multimodal Database for Affect Recognition and Implicit Tagging. IEEE Trans Affect Comput. 2012 Jan;3(1):42–55.

21. Sonkusare S, Breakspear M, Guo C. Naturalistic Stimuli in Neuroscience: Critically Acclaimed. Trends Cogn Sci. 2019 Aug 1;23(8):699–714.

22. Schultz J, Pilz KS. Natural facial motion enhances cortical responses to faces. Exp Brain Res Exp Hirnforsch Exp Cerebrale. 2009 Apr;194(3):465–75.

23. Cohn JF, Kruez TS, Matthews I, Yang Y, Nguyen MH, Padilla MT, et al. Detecting depression from facial actions and vocal prosody. In: 2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops. 2009. p. 1–7.

24. Gavrilescu M, Vizireanu N. Predicting Depression, Anxiety, and Stress Levels from Videos Using the Facial Action Coding System. Sensors. 2019 Jan;19(17):3693.

25. Dibeklioğlu H, Hammal Z, Yang Y, Cohn JF. Multimodal Detection of Depression in Clinical Interviews. Proc ACM Int Conf Multimodal Interact ICMI Conf. 2015 Nov;2015:307–10.

26. Bhatia S, Hayat M, Breakspear M, Parker G, Goecke R. A Video-Based Facial Behaviour Analysis Approach to Melancholia. In: 2017 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017). 2017. p. 754–61.

27. Renneberg B, Heyn K, Gebhard R, Bachmann S. Facial expression of emotions in borderline personality disorder and depression. J Behav Ther Exp Psychiatry. 2005 Sep 1;36(3):183–96.

28. Trémeau F, Malaspina D, Duval F, Corrêa H, Hager-Budny M, Coin-Bariou L, et al. Facial Expressiveness in Patients With Schizophrenia Compared to Depressed Patients and Nonpatient Comparison Subjects. Am J Psychiatry. 2005 Jan 1;162(1):92–101.

29. Girard JM, Cohn JF, Mahoor MH, Mavadati SM, Hammal Z, Rosenwald DP. Nonverbal social withdrawal in depression: Evidence from manual and automatic analyses. Image Vis Comput. 2014 Oct 1;32(10):641–7.

30. Parker G, Hadzi-Pavlovic D, Eyers K, editors. Melancholia: a disorder of movement and mood: a phenomenological and neurobiological review. Cambridge University Press; 1996.

31. Baltrusaitis T, Zadeh A, Lim YC, Morency L. OpenFace 2.0: Facial Behavior Analysis Toolkit. In: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition. Xi'an, China; 2018. p. 59–66.

32. Guo CC, Hyett MP, Nguyen VT, Parker GB, Breakspear MJ. Distinct neurobiological signatures of brain connectivity in depression subtypes during natural viewing of emotionally salient films. Psychol Med. 2016 May;46(7):1535–45.

33. Dhall A, Goecke R, Lucey S, Gedeon T. Collecting Large, Richly Annotated Facial-Expression Databases from Movies. IEEE Multimed. 2012 Jul;19(3):34–41.

34. Parker G. Defining melancholia: the primacy of psychomotor disturbance. Acta Psychiatr Scand Suppl. 2007;(433):21–30.

35. Jiang B, Valstar M, Martinez B, Pantic M. A Dynamic Appearance Descriptor Approach to Facial Actions Temporal Modeling. IEEE Trans Cybern. 2014 Feb;44(2):161–74.

36. Koelstra S, Pantic M, Patras I. A Dynamic Texture-Based Approach to Recognition of Facial Actions and Their Temporal Models. IEEE Trans Pattern Anal Mach Intell. 2010 Nov;32(11):1940–54.

37. Sandbach G, Zafeiriou S, Pantic M, Rueckert D. Recognition of 3D facial expression dynamics. Image Vis Comput. 2012 Oct;30(10):762–73.

38. Cohen I, Sebe N, Garg A, Chen LS, Huang TS. Facial expression recognition from video sequences: temporal and static modeling. Comput Vis Image Underst. 2003 Jul 1;91(1):160–87.

39. Baker AP, Brookes MJ, Rezek IA, Smith SM, Behrens T, Probert Smith PJ, et al. Fast transient networks in spontaneous human brain activity. eLife. 2014 Mar 25;3:e01867.

40. Vidaurre D, Hunt LT, Quinn AJ, Hunt BAE, Brookes MJ, Nobre AC, et al. Spontaneous cortical activity transiently organises into frequency specific phase-coupling networks. Nat Commun. 2018 Jul 30;9(1):2987.

41. Hamm J, Pinkham A, Gur RC, Verma R, Kohler CG. Dimensional information-theoretic measurement of facial emotion expressions in schizophrenia. Schizophr Res Treat. 2014;2014:243907.

42. Rolls ET. A non-reward attractor theory of depression. Neurosci Biobehav Rev. 2016 Sep 1;68:47–58.

43. Perry A, Roberts G, Mitchell PB, Breakspear M. Connectomics of bipolar disorder: a critical review, and evidence for dynamic instabilities within interoceptive networks. Mol Psychiatry. 2019 Sep;24(9):1296–318.

44. Bhatia S, Hayat M, Goecke R. A multimodal system to characterise melancholia: cascaded bag of words approach. In: Proceedings of the 19th ACM International Conference on Multimodal Interaction [Internet]. New York, NY, USA: Association for Computing Machinery; 2017 [cited 2020 Sep 8]. p. 274–80. (ICMI '17). Available from: http://doi.org/10.1145/3136755.3136766

45. Alghowinem S, Goecke R, Wagner M, Parkerx G, Breakspear M. Head Pose and Movement Analysis as an Indicator of Depression. In: 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction. 2013. p. 283–8.

46. Alghowinem S, Goecke R, Wagner M, Epps J, Breakspear M, Parker G. Detecting depression: A comparison between spontaneous and read speech. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. 2013. p. 7547–51.

47. Joshi J, Goecke R, Parker G, Breakspear M. Can body expressions contribute to automatic depression analysis? In: 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG). 2013. p. 1–7.

48. Taylor MA, Fink M. Melancholia: The diagnosis, pathophysiology, and treatment of depressive illness. New York, NY, US: Cambridge University Press; 2006. xiv, 544 p. (Melancholia: The diagnosis, pathophysiology, and treatment of depressive illness).

49. Cheong JH, Brooks S, Chang L. FaceSync: Open source framework for recording facial expressions with head-mounted cameras. 2017.

50. Vidaurre D, Quinn AJ, Baker AP, Dupret D, Tejero-Cantero A, Woolrich MW. Spectrally resolved fast transient brain states in electrophysiological data. NeuroImage. 2016 Feb 1;126:81–95.

51. Vidaurre D, Smith SM, Woolrich MW. Brain network dynamics are hierarchically organized in time. Proc Natl Acad Sci. 2017 Nov 28;114(48):12827–32.

52. Kottaram A, Johnston LA, Cocchi L, Ganella EP, Everall I, Pantelis C, et al. Brain network dynamics in schizophrenia: Reduced dynamism of the default mode network. Hum Brain Mapp. 2019 May;40(7):2212–28.

53. Gilbert M, Demarchi S, Urdapilleta I. FACSHuman, a software program for creating experimental material by modeling 3D facial expressions. Behav Res Methods. 2021 Oct 1;53(5):2252–72.

54. Vidaurre D, Woolrich MW, Winkler AM, Karapanagiotidis T, Smallwood J, Nichols TE. Stable between-subject statistical inference from unstable within-subject functional connectivity estimates. Hum Brain Mapp. 2019;40(4):1234–43.

55. Storey JD. A direct approach to false discovery rates. J R Stat Soc Ser B Stat Methodol. 2002;64(3):479–98.

56. Torrence C, Compo GP. A Practical Guide to Wavelet Analysis. Bull Am Meteorol Soc. 1998;79(1):18.