# SAMM: A Spontaneous Micro-Facial Movement Dataset

Adrian K. Davison, *Student Member, IEEE*, Cliff Lansley,
Nicholas Costen, *Member, IEEE*, Kevin Tan, and Moi Hoon Yap, *Member, IEEE*

**Abstract**—Micro-facial expressions are spontaneous, involuntary movements of the face when a person experiences an emotion but attempts to hide their facial expression, most likely in a high-stakes environment. Recently, research in this field has grown in popularity, however publicly available datasets of micro-expressions have limitations due to the difficulty of naturally inducing spontaneous micro-expressions. Other issues include lighting, low resolution and low participant diversity. We present a newly developed spontaneous micro-facial movement dataset with diverse participants and coded using the Facial Action Coding System. The experimental protocol addresses the limitations of previous datasets, including eliciting emotional responses from stimuli tailored to each participant. Dataset evaluation was completed by running preliminary experiments to classify micro-movements from non-movements. Results were obtained using a selection of spatio-temporal descriptors and machine learning. We further evaluate the dataset on emerging methods of feature difference analysis and propose an Adaptive Baseline Threshold that uses individualised neutral expression to improve the performance of micro-movement detection. In contrast to machine learning approaches, we outperform the state of the art with a recall of 0.91. The outcomes show the dataset can become a new standard for micro-movement data, with future work expanding on data representation and analysis.

**Index Terms**—Micro-movements, micro-expressions, facial analysis, facial action coding system, baseline

## 1 Introduction

Micro-facial expressions occur when a person attempts to conceal their true emotions [1], [2]. When they consciously realise that a facial expression is occurring, the person may try to suppress the facial expression because showing the emotion may not be appropriate or due to a cultural display rule [3]. Once suppression has occurred, the person may mask over the original facial expression and cause a micro-facial expression. In a high-stakes environment, these expressions tend to become more likely as there is more risk to showing the emotion.

The duration of a micro-expression is very short and is considered the main feature that distinguishes them from a facial expression [4], with the general standard being no more than 500 ms [5]. Other definitions of speed show micro-expressions to last less than 250 ms [6], less than 330 ms [7] and less than half a second [8]. From an anatomical point of view, the facial muscles are made up of fast moving fibres that can contract and relax in less than 20 ms including a latency period where the muscle has to receive instruction from the central nervous system [9].

Experiments by Matsumoto and Hwang [10] summarise a micro-expression to be less than half a second, and discuss whether training humans in detecting micro-facial expressions was effective. The findings showed that training improved the ability to read micro-expressions and was retained a few weeks after the initial training. Training humans can be time consuming and expensive, so looking into ways of aiding a person to detect subtle movements would make training more accessible.

Understanding context and how powerful human emotions are, is fundamental to developing an effective detection system using a computer. Studies into people posing facial expressions have found that regions of the brain that are associated with enjoyment activate when a person voluntarily smiles [11] and more recently experiments with a large number of participants (170) [12] found that voluntarily or involuntarily smiling under stressful situations helped reduce heart rates compared with participants who kept a neutral expression.

The main contribution of this paper is the creation of a new spontaneous micro-facial movement dataset with the largest amount of different ethnicities, resolution and age distribution of any similar dataset currently publicly available. By introducing a dataset with a diverse demographic, the data collection is more representative of a population and the micro-expressions induced from many different people can be investigated, as would be the case in a non-lab controlled environment. It is also the first high resolution dataset with seven basic emotion inducement categories recorded at 200 fps. As part of the experimental design, it was proposed to tailor each video stimuli to each participant, rather than obtaining self-reports after the experiment. This allowed for a better choice of video to show to participants for optimal inducement potential.

- *A.K. Davison, N. Costen, K. Tan, and M.H. Yap are with the Informatics Research Centre, School of Computing, Mathematics and Digital Technology, Manchester Metropolitan University, Manchester M15 6BH, United Kingdom. E-mail: {a.davison, n.costen, k.tan, M.Yap}@mmu.ac.uk.*
- *C. Lansley is with the Emotional Intelligence Academy, Walkden M28 7BQ, United Kingdom. E-mail: cliff@eiacademy.co.uk.*

*Manuscript received 13 Nov. 2015; revised 13 May 2016; accepted 21 May 2016. Date of publication 14 June 2016; date of current version 7 Mar. 2018.*
*Recommended for acceptance by S. Scherer.*
*For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.*
*Digital Object Identifier no. 10.1109/TAFFC.2016.2573832*

The second contribution is the performance assessment of the state-of-the-art spatio-temporal methods on our newly established dataset, SAMM, in terms of the feature representation and accuracy in micro-expression classification (described in Section 5). Finally, we propose an individualised baseline temporal difference method, to improve the performance of the state-of-the-art micro-movement detection methods.

## 2 Related Work

Micro-facial expressions are difficult for humans to spot, and usually requires considerable specialist training. Recent research has resulted in different methods for extracting micro-expressions from the face [18], [19], [20] in an attempt to support human decisions. Existing micro-expression datasets are limited in number, with the current state of the art dependent on these datasets to test their methods. Creating more datasets with a broader participant group and stimuli is required to ensure the field of spontaneous micro-expression detection and recognition can expand.

### 2.1 Current Datasets

One of the first micro-expression datasets was created by Polikovsky et al. [13]. The participants were 10 students in a laboratory setting and recorded at 200 fps with a resolution of $640 \times 480$. The demographic was reasonably spread but limited in number with five Asians, four Caucasians and one Indian participant. The laboratory setting ensured lighting was even and a uniform background was used. The micro-expressions in this dataset were posed by participants whom were asked to perform the seven basic emotions. Posed facial expressions have been found to have significant differences to spontaneous expressions [21], therefore the micro-expressions in this dataset are not representative of natural human behaviour and highlights the requirement for expressions induced naturally.

A similar dataset, USF-HD [15], includes 100 posed micro-facial expressions recorded at 29.7 fps. Posed micro-expressions do not re-create a real-world scenario and a low frame rate can risk losing important information about the micro-expressions. In addition, this dataset defined the micro-expressions as no higher than 660 ms, which is longer than the previously accepted definitions of micro-expressions. Moreover, the categories for micro-expressions are smile, surprise, anger and sad, which is reduced from the seven universal expressions by missing out disgust, fear and contempt.

The SMIC dataset [14] consists of 164 spontaneous micro-expressions filmed at 100 fps and was one of the first to include spontaneous micro-expressions obtained through emotional inducement experiments. However, this dataset was not coded using the Facial Action Coding System (FACS) [22] and gives no information on neutral sequences (the participant's face not moving before onset). The protocols for the inducement experiment consisted of showing participants videos to react to and asking them to suppress their emotions, however with no FACS coding the categorisation of emotion labels was left to participant's own self-reporting. Leaving the categorisation to participants allows for subjectivity on the emotional stimuli to be introduced. The recording quality was also decreased due to flickering of light and the facial area was $190 \times 230$ pixels. The SMIC included a wider demographic of participants with six being female and 14 male. Ethnicity was more diverse than previous datasets with ten Asians, nine Caucasians and one African participant, however this still only includes three ethnicities and does not provide a good overview of a population.

To address the low number of micro-expressions in previous datasets, the CASME dataset [16] captured 195 spontaneous micro-expressions at 60 fps, but the facial area was lower than SMIC at $150 \times 190$ pixels. Further, all expressions are FACS coded and included the onset, apex and offset frame number. The duration of any micro-expression did not exceed 500 ms unless the onset duration was less than 250 ms because fast-onset facial expressions were classed as micro [5]. However, 60 fps does not well represent micro-expressions as the movements could be easily missed when recording. The categories for classifying a labelled emotion have been selected based on the video content, self-report of participants and universal emotion theory. Moreover, the dataset uses repression and tense as new additions aside from the universal emotion theory and leaves out contempt and anger.

Shortly after, CASME II [17] was created as an extension of the original CASME dataset. The frame rate increased to 200 fps and 247 newly FACS coded micro-expressions from 26 participants were obtained. The facial area used for analysis was the larger than CASME and SMIC at $280 \times 340$ pixels. However, as with the previous version, this dataset includes only Chinese participants and categorises in the same way. Both CASME and CASME II used 35 participants, mostly students with a mean age of 22.03 (SD = 1.60). Along with only using one ethnicity, both datasets use young participants only, restricting the dataset to analysing similar looking participants (based on age features).

Based on the findings from previous benchmark datasets, much more can be done to address the limitations such as consistent lighting and a wide demographic, however the lack of datasets for micro-movements induced spontaneously motivates the creation of this dataset. A summary of all the current micro-expression datasets can be found in Table 1.

### 2.2 FACS Coding

FACS was introduced by Ekman and Friesen [22] as an anatomically-based technique that objectively measures all observable facial movements. A trained 'coder' views facial movements and expressions, usually in video sequences, and decomposes each individual muscle movements as Action Units (AUs). FACS was designed to provide information about emotion, however the way the information is gathered is in descriptive and behavioural terms, rather than making inferences about the emotional context. This approach works well for the SAMM dataset, as the movements are all objective with no assumptions made about the emotion after each experimental stimulus.

The inter-coder reliability of the FACS codes within the dataset is 0.82, and was calculated by using a slightly modified version of the inter-reliability formula found in the FACS Investigator's Guide [23]

TABLE 1
Summary of Publicly Available Datasets Containing Micro-Facial Expressions

| | Polikovsky et al. [13] | SMIC [14] | USF-HD [15] | CASME [16] | CASME II [17] | SAMM |
|---|---|---|---|---|---|---|
| Micro-Movements | 42 | 164 | 100 | 195 | 247 | 159 |
| Participants | 10 | 16 | N/A | 35 | 35 | 32 |
| Resolution | 640 × 480 | 640 × 480 | 720 × 1,280 | 640 × 480/720 × 1,280 | 640 × 480 | 2,040 × 1,088 |
| Facial Resolution | N/A | 190 × 230 | N/A | 150 × 190 | 280 × 340 | 400 × 400 |
| FPS | 200 | 100 | 29.7 | 60 | 200 | 200 |
| Spontaneous/Posed | Posed | Spontaneous | Posed | Spontaneous | Spontaneous | Spontaneous |
| FACS Coded | No | No | No | Yes | Yes | Yes |
| Emotion Classes | 6 | 3 | 6 | 7 | 5 | 7 |
| Mean Age (SD) | N/A | 26.7 (N/A) | N/A | 22.03 (SD = 1.60) | 22.03 (SD = 1.60) | 33.24 (SD = 11.32) |
| Ethnicities | 3 | 3 | N/A | 1 | 1 | 13 |

$$Re = \frac{3(AU(C_1, C_2, C_3))}{All\_AU}, \tag{1}$$

where $Re$ is the reliability score, $AU(C_1, C_2, C_3)$ is the number of AUs where all coders agreed and $All\_AU$ is the total number of AUs scored by both coders. In contrast, other FACS coded datasets usually have two FACS coders to code each video, however to increase reliability and ensure accurate ground truth, three coders were used in this dataset.

### 2.3 State-of-the-Art Detection Methods

Recent work on the recognition of micro-expressions have provided promising results on successful detection of these difficult movements, however there is room for improvement.

Histogram of Oriented Gradients (HOG) [24] was originally created for human detection in 2D images and used the pixel orientation, weighted by its magnitude, to calculate features for describing a human as an object. Polikovsky et al. [13], [25] then extended this to a temporal descriptor that attempted to model micro-expressions. The recognition stage used $k$-means clustering to cluster particular AUs within the defined facial cube regions. The results were compared with ground truth 'Transition Tags' of muscle activation stages: neutral, onset, apex and offset. The classification rate for onset, apex and offset were 78.13 percent (80.02 percent with Transition Tags), 68.34 percent (70.99 percent) and 79.48 percent (81.85 percent) respectively. It should be noted that the dataset used for analysis contained posed micro-expressions and are not a good representation of naturally induced micro-expressions.

Pfister et al. [18] used temporal interpolation with multiple kernel learning and Random Forest (RF) classifiers on their own SMIC dataset [14]. The authors classify a micro-expression into positive or negative categories depending on two annotators labelling based on subjects' self reported emotions achieving and accuracy of 74.3 percent.

Further improvement in recognition was made by Shreve et al. [20] by using optical flow to calculate the optical strain that occurs from the non-rigid motion of facial skin. A 78 percent true positive rate and .3 percent false positive rate was achieved for detecting micro-expressions. Further, this method is able to plot the strain and visualise a micro-expression occurring across time.

Wang et al. [19] developed a method that uses a tensor independent color space (TICS) model to show performance of micro-expression recognition in a different colour space compared with RGB and grey-scale when using LBP-TOP features. Results were low at 56.09 and 56.91 percent for RGB and grey-scale respectively, with only a slight increase to 58.64 percent for TICS.

Moilanen et al. [26] use an appearance-based feature difference analysis method that incorporates chi-squared ($\chi^2$) distance and peak detection to determine when a movement crosses a threshold and can be classed as a movement. This follows a more objective method that does not require machine learning. The datasets used are the CASME-A and B [16] and the original data from SMIC (not currently publicly accessible). For CASME-A the spotting accuracy (true positive rate) was 52 percent with 30 false positives (FP), CASME-B had 66 percent with 32 FP and SMIC-VIS-E achieved 71 percent with 23 FP. The threshold value for peak detection was set semi-automatically, with a percentage value between [0,1] being manually set for each dataset. Only spatial appearance is used for descriptor calculation, therefore leaving out temporal planes associated with video volumes.

A newly proposed feature, Main Directional Mean Optical-flow (MDMO), has been developed by Liu et al. [27] for micro-expression recognition using Support Vector Machines (SVM) as a classifier. The method of detection also uses 36 regions on the face to isolate local areas for analysis, but keeping the feature vector small for computational efficiency. The best result on the CASME II dataset was 67.37 percent using leave-one-subject-out(LOSO) cross validation, which performed better than the LBP-TOP and Histogram of Optical Flow (HOOF) features. Further, as the MDMO feature is used with machine learning, the vector must be normalised and therefore loses the frame-based temporal attributes that would be useful for detecting onset, apex and offset frames.

Patel et al. [28] introduced a method using optical flow motion vectors, calculated within small region of interest built around facial landmarks, to detect the onset, apex and offset of a micro-expression. The method can also remove head movements, eye blinks and eye gaze changes, common reasons for false positives in micro-movement detection methods, by the use of thresholding. A peak frame is considered true if all the points of an AU group have a motion greater than a certain threshold. An attempt is made to get this system to perform in real-time, however many of the computational time are in seconds, including the facial landmark detection and optical flow calculation. The method also only uses the SMIC dataset at 25 fps, which means the micro-movements are not FACS coded and has a limited temporal resolution for finding subtle motions. The computational times also take a long time at this frame rate, and so higher frame rates for this method would be even higher. The results detailed an area under curve of 95 percent, but produced a high number of FP.

## 3 Method: Experiment Protocols

The experiment comprised of seven stimuli that attempt to induce emotion in the participants, they were told to suppress their emotions so that micro-expressions may occur. To increase the chance of this happening, a prize of £50 was offered to the participant that could hide their emotion the best, therefore introducing a high-stakes situation [1], [2]. Both the goal of hiding their emotions and the monetary prize were advertised to potential participants before taking part. Each participant completed a questionnaire prior to the experiment so that the stimuli could be tailored to each individual to increase the chances of emotional arousal.

To obtain a wide variety of emotional responses, the dataset was required to be as diverse as possible. A total of 32 participants were recruited for the experiment from within the university with a mean age of 33.24 years (SD: 11.32, ages between 19 and 57). The ethnicities of the participants are 17 White British, three Chinese, two Arab, two Malay and one each: African, Afro-Caribbean, Black British, White British/Arab, Indian, Nepalese, Pakistani and Spanish. An even gender split was also achieved, with 16 male and 16 female participants.

### 3.1 Emotion Inducement Procedure

Participants were first introduced to the experiment, and each were asked if they have read the participant information. A release agreement was signed and the participant was shown to their seat. The observer let the participant know that they can stop at any time, due to the potential for the stimuli to over stimulate their emotions. Participants were also reminded that they are to suppress their true emotions and keep a neutral face with the aim of winning £50. The observer went to the observation room and the experiment began.

Each stimuli was shown and the participants were asked after every one if they were happy to continue, this ensured participants fully offset from any emotion they were feeling. Participants were only recorded when the stimulus was shown. After recording the observer returned to the experiment room and thanked the participant.

A formal ethical clearing procedure took place to safeguard participant's when experiencing stimuli that provokes an emotional response. Every person was free to stop the experiment at any time and a full information sheet and release agreement on how the data would be used was issued.

Suppression of emotions is inherently a social act, and keeping the observer a short distance away in an observation room may seem opposite to this. However, due to the lab setting that participants are within, participants may not be fully relaxed if they are constantly aware of someone watching. The observer is kept out of sight to maximise the chances of natural suppression by making participants as comfortable as possible.

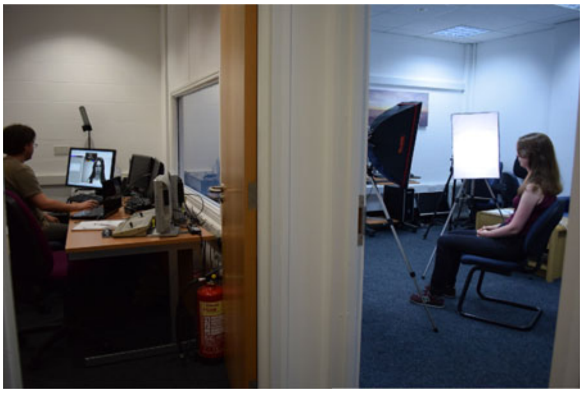

Fig. 1. The left hand room shows the observation room and the right side shows where the participant completes the experiment.

### 3.2 Equipment and Experimental Set-Up

The experiment was set-up in a room that helped keep interaction with the participant to minimum and allow them to become comfortable in the laboratory surroundings. The observer controlled the equipment in one room and the participant had emotional stimuli shown to them to induce particular emotions (see Fig. 1). A one-way mirror allows observation of the participants and an intercom system was used to communicate with participants, if necessary, without physically entering the room to keep interruption to a minimum. Participants watch all emotional stimuli on a 32 inch flat-screen television.

The experiment room contained all the equipment required for capturing the high-speed videos. To set up the environment, the camera and participant chair stayed in the same position for every person, however the lights required to be adjusted based on a person's height to ensure an even lighting on the face. The camera is connected to a system that is able to continuously capture high-speed video data for up to 20 minutes.

#### 3.2.1 Camera

The camera used was a Basler Ace acA2000-340km, with a grey-scale sensor, set to record at 200 fps and resolution set to $2{,}040 \times 1{,}088$ pixels. To the best of our knowledge this is currently the highest resolution available for this type of dataset.

#### 3.2.2 Lighting

Lighting can be problematic for high-speed cameras as many lighting systems use alternating current that refreshes regularly at a usual frequency of 50 Hz. Recording at 200 fps, the camera can pick-up the lights refreshing and this shows as flickering on the resulting images. To counter this, two lights that contain an array of LEDs was used and incorporated direct current to avoid flickering. Light diffusers were placed around the lights to soften and even out the light on participant's faces.

#### 3.2.3 High-Speed Data Capture

Images were captured using a frame grabber and a RAID array of solid state drives to ensure no dropped frames occur. The software used was IO Industries Streams 7 that

TABLE 2
Tailored Stimuli Used to Induce Emotions

| Video Stimuli Description | Duration | Emotion Link |
|---|---|---|
| Westboro baptist church | 0′50″ | Contempt |
| Lancing a boil | 0′21″ | Disgust |
| Snake attacks camera | 0′17″ | Fear |
| Angrily barking dog through a fence | 0′24″ | Fear |
| Jumping spider | 0′25″ | Fear |
| Attacking crab | 0′19″ | Fear |
| Scary puppets | 0′17″ | Fear |
| Large spider chase | 0′17″ | Fear |
| First person bungee jump | 0′16″ | Fear |
| Moth flying around | 0′21″ | Fear |
| Racist woman | 0′25″ | Anger |
| A dog being kicked | 0′26″ | Anger |
| Bullying | 0′35″ | Anger |
| Unruly teenagers | 0′28″ | Anger |
| Movie death (Champ) | 0′46″ | Sadness |
| Bullying | 0′35″ | Sadness |
| A dog being kicked | 0′26″ | Sadness |
| Twin towers collapsing | 0′49″ | Sadness |
| Baby laughing | 0′26″ | Happiness |
| Flight of the Conchords song | 0′22″ | Happiness |
| Dog playing | 0′14″ | Happiness |
| Presentation with Participant’s face | N/A | Surprise |

allows for recording and analysis of the data. As the software initially records to a proprietary format, the original can be used to export various formats as required.

### 3.3 Inducement Stimuli

The majority of the emotional inducement stimuli were video clips from the Internet. If a participant was fearful of heights, a first-person video of someone bungee jumping was shown. Further information on the tailored videos are discussed in the questionnaire section and a description of the video clips used is shown in Table 2 along with the emotion linked to the inducement.

For surprise, a presentation was used and shown as the last stimulus. The presentation appeared to be boring slides that used a lot of text, however within the slides was an image of the participant. This enabled an unexpected event, without the risk of startling the participant.

### 3.4 Questionnaire

Some datasets [14], [16], [17] assign an emotion label to videos based on self-reports completed by participants. Therefore participants wait until a stimulus has been experienced and then record what emotion they felt during each stimulus.

TABLE 3
Questionnaire Participants Filled In Before the Experiment

| Question No. | Question |
|---|---|
| 1 | What are your current fears and phobias? |
| 2 | Highlight the principles or moral standpoints that you hold |
| 3 | How is your view of another person/group affected when they have opposite moral standpoints to your own? |
| 4 | Outline the beliefs and values that you hold |
| 5 | Describe what makes you angry |
| 6 | Describe what makes you sad |
| 7 | Describe what makes you disgusted |
| 8 | Describe what makes you happy |

In contrast, our experiment required each person to fill in a questionnaire before turning up to the experiment so that each emotional stimuli video could be tailored to what each person found to induce emotion in themselves. All of the questions can be found in Table 3.

Ground truth was obtained for every movement using FACS and inconsistencies between coders eliminated by mutual cross-checks to establish a consensus. This is especially relevant for micros, as coding is done objectively based on the muscle movements, and does not try to interpret an emotion [22]. For this dataset, an objective micro-expression is named a micro-movement and coded to have an onset, apex and offset frame. Each lasts 100 frames or less, translating to 500 ms at 200 fps. Any movements that were coded to be longer than 100 frames in duration would be classed as a macro-facial expression. An example micro-movement that has been FACS coded from the SAMM dataset can be seen in Fig. 2.

### 3.5 Independent Video Stimuli Ratings

To help understand the emotional response that might be exhibited after watching the video stimuli in this experiment, an independent rating of all video was completed. The independence of these ratings refers to using people who were not participants in the dataset. It should be noted that the inducement of surprise was made using a presentation and participant’s individual faces, so these cannot be independently rated.

#### 3.5.1 Rating Quantification

To quantify the emotional response, the ratings are made using Self-Assessment Manikins (SAM) first devised by Lang [29]. They can be used as an inexpensive and easy method for quickly assessing reports of affective response

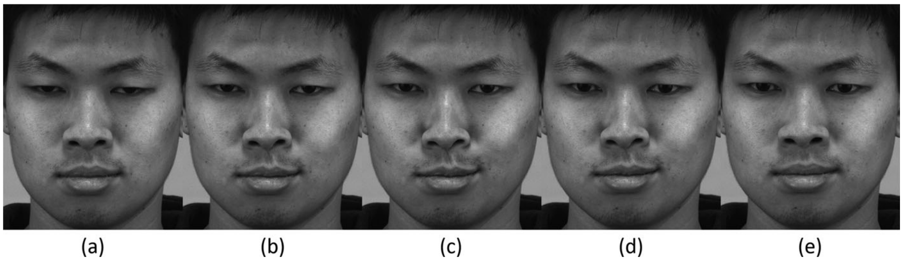

Fig. 2. An example of a coded micro-movement. The movement shows AU 13 and AU 39, which is the sharp lid puller and nostrils compressing. Image (a) is the onset frame, (c) is the apex where the mouth curls upwards sharply and the nostrils move inwards. Finally, (e) is the offset frame.

TABLE 4
The Independent Ratings of Emotional Response for Each Video Stimulus Used in the Experiment

| Video Stimuli Description | Valence | Dominance | Arousal | Emotion | Emotion Rating |
|---|---|---|---|---|---|
| Westboro baptist church | 3.07 | 4.20 | 5.73 | Contempt | 3.00 |
| Lancing a boil | 3.00 | 3.73 | 6.3 | Disgust | 3.23 |
| Snake attacks camera | 4.00 | 3.70 | 5.83 | Surprise | 2.73 |
| Angry dog barking through a fence | 5.13 | 4.56 | 5.3 | Interest | 2.20 |
| Jumping spider | 4.10 | 3.36 | 6.40 | Fear | 3.03 |
| Attacking crab | 3.70 | 4.56 | 4.73 | Fear | 2.10 |
| Scary puppets | 5.63 | 6.40 | 3.96 | Interest | 2.33 |
| Large spider chase | 3.43 | 4.30 | 4.83 | Fear | 1.93 |
| First person bungee jump | 3.06 | 4.63 | 4.70 | Fear | 2.23 |
| Moth flying around | 4.23 | 5.9 | 3.70 | Interest | 1.86 |
| Racist woman | 1.70 | 2.86 | 7.3 | Anger | 3.80 |
| A dog being kicked | 2.66 | 3.83 | 6.23 | Anger | 3.30 |
| Bullying | 3.43 | 4.93 | 4.56 | Contempt | 2.40 |
| Unruly teenagers | 2.63 | 4.70 | 6.13 | Disgust | 3.16 |
| Movie death (Champ) | 2.50 | 3.50 | 5.70 | Sadness | 3.10 |
| Twin towers collapsing | 1.86 | 2.70 | 6.70 | Sadness | 3.46 |
| Baby laughing | 7.66 | 6.76 | 6.70 | Joy | 3.36 |
| Flight of the Conchords song | 5.23 | 5.13 | 6.00 | Surprise | 2.96 |
| Dog playing | 6.13 | 6.36 | 4.66 | Joy | 1.90 |

in a variety of contexts. Bradley and Lang [30] later compared the SAM, which typically use a 9-point scale for rating, against a Semantic Differential scale devised by Mehrabian and Russell [31] that requires an 18-point scale. Results showed that SAM may better track the personal response to an affective stimulus.

The SAM require three types of judgements in rating emotional response. The first is the valence, or pleasure, which rates the feeling of positivity (higher up the scale) or negativity (lower down the scale). Each person was asked to choose the rating at which they felt best described the valence they felt during the video stimulus. Next is the dominance rating, which rates how much a person feels in control of a situation. The lowest rating, or smallest manikin, means that the person feels they have no power to handle the situation. The opposite is true for the highest rating, or the biggest manikin. Finally is the arousal rating, which is the emotional excitation level based on how they feel when watching the videos. The range goes from sleepiness or boredom at the lowest rating score up to excitation in the highest rating.

After all the rating for the SAM are completed, the final rating people are asked to complete is to select an emotion that best describe how they felt overall. The emotion is then rated from 0-4, with 0 being the weakest feeling of that emotion and four the strongest.

### 3.5.2 *Results*

The ratings for all of the 19 video stimuli were taken by 30 people, 60 percent who were white British and 40 percent who were other ethnicities. There was a split of 14 males and 16 females with a mean age of 34.48 years (SD: 13.73). None of the raters had never seen the videos before and a summary of the results can be seen in Table 4. To calculate the final valence, dominance, arousal and emotion scores, the mean rating across each video from each person is taken. Along with the ratings based on SAMM, an emotion was also chosen by the people rating the videos, with the most common emotion chosen being the one used to describe the video.

The majority of independent ratings observed were inline with emotion categories the stimuli were set to. Further, many of the ratings show a consistent inducement potential in the chosen category. Even though some videos are rated lower, such as the 'Dog Biting' video, this does not take away from the fact the videos were tailored for each participant, and so generalising these videos to a different audience is likely to produce varying results.

## 4 Dataset Analysis

The SAMM dataset contains micro-movements captured at 200 fps. Macro-movements were also coded as to not disregard potential useful movements that may be used at a later date.

Frequency occurrence for all AUs is calculated from two groups of durations

- Up to 100 frames (or half a second).
- From 101 to 166 frames (or two-thirds a second).

Using up to 100 frames allows for comparison against CASME II, which labelled their data to this length. Additional statistics for the second group can be used for when the duration of the movement is defined slightly higher than usual. Table 5 outlines the frequency occurrences for well known AUs in these groups. There was a total of 222 AUs in the group of up to 100 frames and 116 in the group up to 166 frames. The percentage occurrence of all 338 micro-movements in Table 5 was 45.3 percent, and up to 100 frames was 29.7 percent. These percentages show a large portion of the overall AUs coded in the dataset turned out to be in the micro-movement category.

The FACS coding was completed by three certified coders, who were not otherwise involved in the study, to ensure inter-coder reliability. Coding was performed after the videos have been recorded in accordance with usual FACS coding procedure. At no point did the coders know the context of the videos they were coding, which means no coder was aware of the stimuli used to induce emotion in participants. Every movement was coded, including the

TABLE 5
The Occurrence Frequency for Both Duration Groups Has Been Calculated for the Main Upper and Lower Face AUs

| Upper Face | | | Lower Face | | |
|---|---|---|---|---|---|
| AU | No. Of Occurrences | | AU | No. Of Occurrences | |
| | Up to 100 Frames | 101 to 166 Frames | | Up to 100 Frames | 101 to 166 Frames |
| 1 | 6 | 5 | 9 | 5 | 1 |
| 2 | 16 | 7 | 10 | 5 | 3 |
| 4 | 23 | 14 | 12 | 29 | 13 |
| 5 | 9 | 8 | 14 | 11 | 7 |
| 6 | 5 | 0 | 15 | 4 | 1 |
| 7 | 45 | 14 | 17 | 7 | 6 |
| Other | 9 | 9 | 20 | 7 | 2 |
| | | | 23 | 1 | 3 |
| | | | Other | 40 | 23 |
| **Total** | 113 | 57 | **Total** | 109 | 59 |

*All other rates can be found in the FACS code sheet provided with the dataset.*

macro-movement, with an onset, apex and offset frame recorded to note the duration. Unlike most other datasets, every FACS AU was coded regardless of their relation to emotions. This includes head and eye movement codes. By FACS coding the data comprehensively, the dataset can be used for a much wider purpose when it becomes publicly available.

A chi-square($\chi^2$) test was conducted using all observed facial movements to test the significance of the different AUs invoked by the emotional context. Certain FACS AUs are used within Emotion FACS (EMFACS) [23] to only describe critical AUs related to emotion. Table 6 shows occurrences of these key reliable muscle movements during specific stimuli. Non-reliable muscle movements have been included for statistical analysis to show AUs that are not classed as reliable, but occurred frequently across participants. The reliable muscles for Contempt did not occur during any stimuli, and so this group has not been included. The Surprise reliable group has been included, but has too few results to allow for reliable statistical comparison and has been omitted from calculations. The data for each individual AU has been pooled into categories for the $\chi^2$ test to be acceptable and the significance level was set to $\alpha = 0.05$.

To determine if there is a statistically significant relationship between AU groups and stimuli categories within the experiment, two hypotheses are proposed. The first hypothesis states that there is no association between facial movements and the corresponding stimuli. The alternative hypothesis states there is some association between a participant's facial movement and the stimuli experienced. For only reliable AU groups, the $\chi^2 = 82.28$, with a critical $\chi^2 = 18.49$, $p = 9.25 \times 10^{-7}$ and degrees of freedom (*df*) = 30. When the non-reliable movements are included the $\chi^2 = 136.4$, with a critical $\chi^2 = 23.27, p = 1.34 \times 10^{-13}$ and *df* = 36.

From this analysis, the hypothesis with no association between facial movements and stimuli can be rejected as there is statistical significance between the two. Further to this conclusion, from the observed values the pooled Happiness reliable AUs and stimuli have the highest frequency and show a correlation between the movement and emotional context. In other groups this is less apparent, however unlike in similar experiments performed by Mammucari et al. [32] the experimental protocols required participants to suppress their true emotion, therefore making it less likely, if the experiment was a success, for participants to show all reliable muscles. For example, some participants showed a single AU rather than combinations of AUs, and the masking of reliable AUs with other movements is a side-effect of asking participants to suppress.

## 5 Micro-Expression Recognition

To validate the movements within the dataset that are up to 100 frames in length, state-of-the-art features used for micro-expression recognition are applied. The movements are split into blocks where the features are applied to each

TABLE 6
Frequency Occurrences of Reliable AUs from EMFACS Pooled Together to Form AU Groups

| Stimuli Category | Reliable AU Groups | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Disgust | Fear | Anger | Sadness | Happiness | Surprise | Non-Reliable Movements | |
| **Contempt** | 5 | 1 | 9 | 4 | 15 | 1 | 54 | **89** |
| **Disgust** | 12 | 2 | 12 | 5 | 17 | 0 | 43 | **91** |
| **Fear** | 7 | 4 | 5 | 10 | 31 | 1 | 41 | **99** |
| **Anger** | 5 | 0 | 8 | 1 | 10 | 0 | 67 | **91** |
| **Sadness** | 1 | 4 | 3 | 10 | 10 | 5 | 53 | **86** |
| **Happiness** | 10 | 4 | 5 | 14 | 69 | 2 | 39 | **143** |
| **Surprise** | 8 | 3 | 4 | 15 | 37 | 4 | 75 | **146** |
| | **48** | **18** | **46** | **59** | **189** | **13** | **372** | **745** |

*The values correspond to how many times a group occurred when a participant experienced a video based on a stimulus category. Also shown is the non-reliable movements that do not relate to emotional context, but occur frequently.*

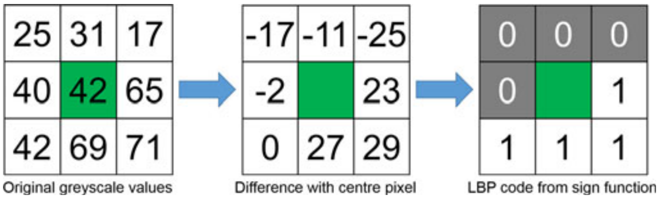
25 31 17
40 42 65
42 69 71
Original greyscale values
-17 -11 -25
-2 23
0 27 29
Difference with centre pixel
0 0 0
0 1
1 1 1
LBP code from sign function

Fig. 3. LBP code calculation by using the difference of the neighbourhood pixels around the centre.

individual video block. Each block is then assigned a ground truth as a movement, indicating a block containing a micro-movement, or a non-movement, indicating a video block with no movement present. Finally, Random Forests are used as a classifier to perform 10-fold cross validation and leave-one-subject-out classification. The original resolution of the images within SAMM $2{,}048 \times 1{,}088$, with the cropped facial resolution of $400 \times 400$ being used for the experiments.

## 5.1 Spatio-Temporal Feature Extraction

Four spatio-temporal methods were used to perform initial tests on the dataset and provide results to compare previous methods that use machine learning classification on the proposed dataset. The first two methods are LBP-TOP based and the others are HOG based.

### *5.1.1 LBP-TOP Based Descriptors*

The first method uses LBP-TOP which was first described as a texture descriptor by Zhao et al. [45] that used XT and YT temporal planes rather than just the two-dimensional XY spatial plane. Yan et al. [17] used this method to report initial findings in the CASME II dataset, and so this method is used on the SAMM dataset to compare results.

LBP-TOP was extended by Davison et al. [33] to include Gaussian derivatives (GDs) that improved on the classification accuracy than on LBP-TOP alone. The Gaussian function is a well-known algorithm and is usually referred to being a normal distribution

$$G_{(xx,yy)}(x,y;\sigma) = \left(\frac{(x^2,y^2)}{\sigma^4} - \frac{1}{\sigma^2}\right)G(x,y;\sigma) \tag{2}$$

$$G_{xy}(x,y;\sigma) = \frac{xy}{\sigma^4}G(x,y;\sigma), \tag{3}$$

where $\sigma$ is the scaling element of the Gaussian derivatives. Ruiz-Hernandez et al. [34] use the second order derivative to extract blobs, bars and corners to eventually use the features to detect faces in a scene (Eqs. (2) and (3)). GDs also provide a powerful feature set with scale and rotation invariant image description. However, when processing higher order derivatives, the feature selection becomes more sensitive to noise, and computationally expensive. This is the reason why the first two derivatives are used.

The features are then summed and LBP-TOP is applied. Each block has the standard LBP operator applied [35] with $\alpha$ being the centre pixel and P being neighbouring pixels with a radius of $R$

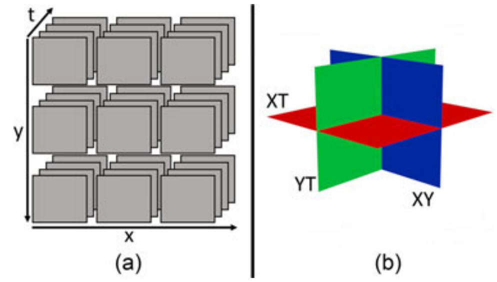

Fig. 4. (a) Visual representation of the spatio-temporal configuration of video frames split into blocks. (b) The XY, XT, YT planes used for feature analysis in LBP-TOP and 3D HOG.

$$LBP_{P,R} = \sum_{p=0}^{P-1} s(g_p - g_\alpha)2^p, \tag{4}$$

where $g_\alpha$ is the grey value of the centre pixel and $g_p$ is the grey value of the $p$-th neighbouring pixel around $R$. $2^p$ defines weights to neighbouring pixel locations and is used to obtain the decimal value. The sign function to determine what binary value is assigned to the pattern is calculated as

$$s(\mathbf{A}) = \begin{cases} 1, & \text{if } \mathbf{A} \geq 0 \\ 0, & \text{if } \mathbf{A} < 0. \end{cases} \tag{5}$$

If the grey value of $P$ is larger than or equal to $C$, then the binary value is 1, otherwise it will be 0. Fig. 3 illustrates the sign function on a neighbourhood of pixels. After the image has been assigned LBPs, the histogram can be calculated by

$$H_i = \sum_{x,y} I\{f_l(x,y) = i\}, i = 0, \ldots, n-1, \tag{6}$$

where $f_l(x,y)$ is the image labelled with LBPs. The parameters set for each method were chosen based on the best results obtained from the respective research. The radii for LBP-TOP only was set to 1, 1, 4 for the X, Y and T planes respectively. When GDs were added in the second method, the radii were set to 3, 3, 3 and the Gaussian sigma value set to 5. The temporal representation of the video blocks and XY, XT and YT planes can be seen in Fig. 4.

### *5.1.2 3D HOG*

The next method is the temporal HOG descriptor used by Polikovsky et al. [13], [25] to cluster AUs that were classed as micro-expressions. Like LBP-TOP it uses three planes to describe the spatial and temporal features. Gradients are calculated in the three dimensions of a video and the pixel orientation and magnitude is calculated for the XY, XT and YT planes. The magnitude values are binned into orientations so that the values are weighted based on the orientation of the gradient

$$Orientation(x,y) = \arctan\left(\frac{\mathbf{G}_y}{\mathbf{G}_x}\right) \tag{7}$$

$$Magnitude(x,y) = \sqrt{(\mathbf{G}_x)^2 + (\mathbf{G}_y)^2}, \tag{8}$$

where $\mathbf{G}_x$ and $\mathbf{G}_y$ are the derivatives of the $x$ and $y$ spatial directions respectively. The original HOG descriptor is then

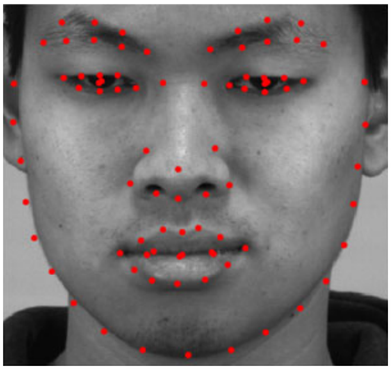

Fig. 5. The 83 points detected using Face++.

applied to each plane using Dollar's Matlab toolbox using the implementation described in [36], [37]. In contrast to the original HOG descriptor, the orientation defined for this experiment will be $2\pi$ instead of $\pi$ as we are interested in detecting movements for all directions. Pixel magnitude (Eq. (8)) and orientation (Eq. (7)) are calculated and the magnitude values are binned into particular orientations so that the values are weighted based on the orientation of the gradient. The histogram bin parameter selection was the same as in the Polikovsky et al. protocols, where XY uses eight bins and XT and YT uses 12 bins with two 'no change' bins. Each plane is then concatenated to form the final 32-bin feature descriptor.

### 5.1.3 Deformable Part Models

The final method is another HOG-based sliding window descriptor by Felzenszwalb et al. [38] and is also implemented using Dollar's Matlab toolbox as a computationally faster implementation but identical results as the original. Deformable Part Models (DPM) use star models defined by HOG as a coarse root filter that covers the entire object and then creates higher resolution part filters to describe local regions of an object (in this case the face). The parts, when described together, form an object only when they are meaningful in their geometrical constraints in the spatial domain originally, and extended into the temporal domain for this method.

## 5.2 Method

Normalisation is applied to all sequences so that all the faces are in the same position by using affine transformation. The points used for alignment are obtained using the Face++ automatic facial point detector [39]. The face of the sequences then needs to be cropped to remove the unnecessary background in each image. An example of the 83 facial points can be seen in Fig. 5, where the eye centre points and the outermost points were used for alignment and cropping the face respectively.

Feature extraction begins by grey-scaling each image sequence and dividing each image into blocks. To test different regions, images were divided into $4 \times 4$ and $5 \times 5$ regions with 16 and 25 video blocks for each movement respectively. Using $5 \times 5$ blocks allows for comparison with the CASME II procedure [17], and the other $4 \times 4$ blocks tests different local regions (see Fig. 6). Each video block then had the temporal descriptor applied as outlined in the previous section.

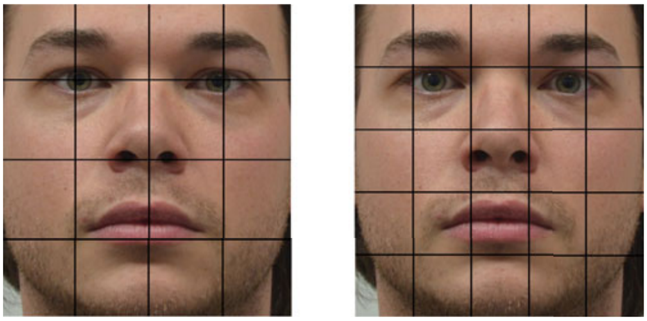

Fig. 6. Images split into $4 \times 4$ blocks (left) and $5 \times 5$ blocks (right).

Different blocking configurations may suit the aligned images better and changing the sizes allowed for testing this hypothesis.

Finally, the image sequences are classified using RF with the default parameters in the machine learning tool, Weka [40]. Binary classification is used with the two classes being movement, referring to the video blocks with micro-movements, and non-movements, referring to the video blocks that contain no movements.

Each video block was assigned a ground truth label, from the FACS coding, and 10-fold cross validation is used to calculate the overall classification accuracy and the F-measure and Matthews Correlation Coefficient (MCC) for the movement class. For further validation, a leave-one-subject-out approach was applied where each subject was left out once for each test and an average result was taken.

## 5.3 Results

We achieve good preliminary results from testing the dataset on existing temporal descriptors and discuss the scope for further investigation into the dataset and the micro-movements it contains. Using binary classification, micro-movement blocks and non-movement blocks, LBP-TOP with a radius of 3, 3, 3 for XY, XT and YT planes respectively and using a $4 \times 4$ block configuration produces the best result of **0.67** when using the F-measure statistic.

The results shown are the accuracy of correctly classified movement and non-movement blocks. This is not representative of the result as a whole. F-measure uses precision and recall to obtain a harmonic mean of the classification accuracy. For binary classification, the MCC introduced by Matthews et al. [41] takes into account the true and false positives (TP and FP) and true and false negatives (TN and FN) to obtain a balanced coefficient measure between $-1$ and 1, where 1 is perfect classification, 0 is random chance and $-1$ is total disagreement.

Table 7 details the performance of the two spatio-temporal descriptors based on Local Binary Patterns and Table 8 shows the performances of 3D HOG and DPM, which are both based on Histogram of Oriented Gradient descriptors. In all cases, the descriptors are attempting to generalise the micro-movements and non-movement blocks across all instances and performs well for this difficult task. Further discussion on the generalisation of micro-movements will be within the next section.

To further expand on the testing results, we employ a method similarly used in [27] to leave each subject out for

TABLE 7
Results Calculated Using 10-Fold Cross Validation

| | | Block Configuration | | | | | |
|---|---|---|---|---|---|---|---|
| | | 4 × 4 | | | 5 × 5 | | |
| Descriptor | Radii | Accuracy (%) | F-Measure | MCC | Accuracy (%) | F-Measure | MCC |
| LBP-TOP | 3,3,3 | **82.70** | **0.67** | **0.56** | 91.32 | 0.53 | 0.52 |
| LBP-TOP | 1,1,4 | 82.31 | 0.66 | 0.55 | **91.52** | **0.53** | **0.53** |
| LBP-TOP & GD | 3,3,3 | 82.07 | 0.65 | 0.54 | 90.80 | 0.47 | 0.47 |
| LBP-TOP & GD | 1,1,4 | 82.35 | 0.66 | 0.55 | 91.04 | 0.50 | 0.49 |
| LBP-TOP & GD (XT Plane) | 3,3,3 | 80.18 | 0.61 | 0.49 | 89.16 | 0.32 | 0.34 |
| LBP-TOP & GD (XT Plane) | 1,1,4 | 80.46 | 0.62 | 0.50 | 88.68 | 0.30 | 0.30 |

*Spatio-temporal methods based on LBP are described using different LBP radii and different block splitting sizes. The final two results only use the XT plane and the best results are highlighted in bold.*

TABLE 8
Results Calculated Using 10-Fold Cross Validation

| | Block Configuration | | | | | |
|---|---|---|---|---|---|---|
| | 4 × 4 | | | 5 × 5 | | |
| **Descriptor** | Accuracy (%) | F-Measure | MCC | Accuracy (%) | F-Measure | MCC |
| 3D HOG | 79.05 | 0.59 | 0.46 | 90.36 | 0.42 | 0.43 |
| DPM | 78.85 | 0.56 | 0.45 | 90.14 | 0.41 | 0.42 |

*Spatio-temporal methods for HOG and DPM are described using different block splitting sizes.*

TABLE 9
Results for the LBP-Based Features Using Leave-One-Subject-Out Tests

| | | Block Configuration | | | | | |
|---|---|---|---|---|---|---|---|
| | | 4 × 4 | | | 5 × 5 | | |
| **Descriptor** | **Radii** | Accuracy (%) | F-Measure | MCC | Accuracy (%) | F-Measure | MCC |
| LBP-TOP | 3,3,3 | 77.86 | 0.49 | 0.39 | 88.99 | 0.26 | 0.26 |
| LBP-TOP | 1,1,4 | 79.11 | 0.55 | 0.45 | 89.66 | 0.32 | 0.33 |
| LBP-TOP & GD | 3,3,3 | 79.59 | 0.52 | 0.44 | 88.93 | 0.27 | 0.28 |
| LBP-TOP & GD | 1,1,4 | 80.06 | 0.55 | 0.46 | 89.96 | 0.37 | 0.38 |
| LBP-TOP & GD (XT Plane) | 3,3,3 | 78.86 | 0.52 | 0.42 | 87.63 | 0.16 | 0.15 |
| LBP-TOP & GD (XT Plane) | 1,1,4 | 78.92 | 0.53 | 0.43 | 87.79 | 0.19 | 0.18 |

testing and use the rest of the data for training. This method is aptly named leave-one-subject-out tests. Table 9 details all the LBP-based features using LOSO, with the highest performing feature being $4 \times 4$ block LBP-TOP with GD, using the radii 1, 1, 4, achieving an accuracy of 80.06 percent. Table 10 shows the 3D HOG and DPM feature with the highest performing being $5 \times 5$ block 3D HOG feature with an accuracy of 88.57 percent.

The performance of LOSO tests shows that the amount of data used for training may still not be enough to represent micro-expressions well. The accuracies are higher due to the majority of non-movement blocks being correctly classified. The more representative result of finding micro-expressions can be seen in the F-measure and MCC values, however these results are promising when compared with similar outcomes of the 10-fold cross validation.

## 6 Micro-Movement Detection Using Baseline Analysis

The key to fully understanding micro-facial expressions is not to immediately associate them with an emotion, or emotional class. Any classification of micro-facial expressions should be done objectively, like in FACS, by focusing on the muscle movements of the face, or AUs. Previously described methods have focused on the classification of micro-expressions, using machine learning techniques, into distinct classes such as happiness, tense, positive and negative. Unlike macro-facial expressions, micro-movements on the face can manifest themselves in such subtle ways that distinct classes are near-impossible to predict every case. This is evident in the relatively low accuracy scores of recent work compared with macro-facial expression recognition. Further, as micro-facial expressions are the result of attempting to hide true emotion, it is likely that people will try to mask over an AU with another. For example, if a person wants to mask a smile (AU 12), then instinctively they may frown (AU 15) to cover up the smile, even though these muscles are separate.

By treating micro-expressions objectively as muscle activation rather than expressions of emotion, it would be more descriptive and less computationally expensive to analyse micro-expressions as temporal differences. Moilanen et al. [26] proposed an objective method that does not

TABLE 10
Results for the 3D HOG and DPM Features Using Leave-One-Subject-Out Tests

| | Block Configuration | | | | | |
|---|---|---|---|---|---|---|
| | $4 \times 4$ | | | $5 \times 5$ | | |
| **Descriptor** | Accuracy (%) | F-Measure | MCC | Accuracy (%) | F-Measure | MCC |
| 3D HOG | 71.28 | 0.31 | 0.19 | 88.57 | 0.22 | 0.21 |
| DPM | 71.17 | 0.30 | 0.17 | 88.31 | 0.19 | 0.18 |

require classification by using appearance-based difference analysis using an LBP feature and $\chi^2$ distance to find micro-expression peaks across a temporal sequence. A threshold value for peak detection is set by taking the difference between the maximum and mean values of a contrasting difference vector and multiplying by a percentage value $p$ that can be in the range of [0,1].

By these calculations, the value will never exceed the maximum peak of the difference vector and thus not being able to detect faces with no movement. The influence of the difference vector in threshold calculation causes at least one peak to always be classed as a detection. The proposed method uses the participant's baseline provided by the SAMM dataset to improve on this difference method and contribute to the growing field of micro-expression spotting.

### 6.1 Individualised Baseline Threshold

To address the limitations of the previous method, we proposed a difference analysis method using individualised baselines of the participants of SAMM [42]. Using the spatial HOG feature extracted from each frame of split blocks of videos, the $\chi^2$ distance was applied and a threshold was obtained using the neutral baseline sequences.

By using the participant's baseline feature, a more Psychological-based approach is employed to differentiate between movement peaks and a neutral sequence with naturally occurring difference peaks from high-speed video noise or head movement.

Further, an adaptive baseline threshold (ABT) is proposed to improve on the individualised baseline approach that uses a combination of the baseline feature and movement currently being analysed.

### 6.2 Results

Using all of the movements from the SAMM dataset, the two difference analysis methods are tested. Moilanen et al. [26] reported 'spotting accuracy' in their original results. This translated to the true positive rate or recall of the method and is calculated by using the Equation $\frac{TP}{(TP+FN)}$, where $TP$ and $FN$ are the true positives and false negatives respectively. The results report the recall and precision of the discussed methods. The F-measure, the harmonic mean of precision and recall, is also calculated using $\frac{2TP}{(2TP+FP+FN)}$, where $FP$ are the false positives.

The results in Table 11 show that the method in [26] does not perform well on our dataset and mirrors a similar problem exhibited in their original results on the higher resolution CASME-A dataset, where the recall was 0.52. As the SAMM dataset has the highest available resolution on micro-movements, it shows that this method struggles to process such data effectively. In contrast, the individualised baseline method presented in [42] returns a recall of 0.8429 and F-measure of 0.7672, owing to the ability of our method to spot actual movement and disregard what is the participant's neutral expression.

However, the precision of 0.7041 reflects high false positives. To improve the results, we propose an Adaptive Baseline Threshold that takes into account the mean of both the movement and baseline feature vector that adapts the threshold level based on a balance between what is happening on the participant's face and what their baseline expression level is. The ABT can be calculated by

$$ABT = \begin{cases} \max(\beta), & \text{if } \max(\beta) > \bar{\epsilon} \\ \frac{\bar{\epsilon}+\bar{\beta}}{2}, & otherwise, \end{cases} \tag{9}$$

where $ABT$ is the calculated adaptive threshold, $\beta$ is the baseline feature vector and $\bar{\beta}$ is its mean. The movement feature vector and its mean is denoted by $\epsilon$ and $\bar{\epsilon}$ respectively. By contrasting and comparing the baseline feature and movement feature using ABT, the proposed method substantially increases the detection rate and produced the best result of 0.9125 and 0.8179 for recall and F-measure, respectively. We observe that the precision has increased by 3 percent when compared to Davison et al. [42], this implies that our proposed method manages to reduce some of the false positives, but the false positives remain a challenge for micro-movement detection. Further, work into reducing the head movement and other inevitable movement of the head will be investigated.

## 7 Data Availability

The SAMM micro-movements will be available online for download (goo.gl/SJmoti) for research purposes and will contain 4 GB of images formatted to jpeg. Each image represents a frame of the video sequence. The raw captured data, that does not crop out the FACS coded movements, is also available. However, written request must be made to obtain this along with a way to receive the data as the size on disk is around 700 GB and unable to be hosted online for direct download.

TABLE 11
The Best Results of the Current Feature Difference Methods and the Proposed Method Using the SAMM Dataset

| Method | Recall | Precision | F-Measure |
|---|---|---|---|
| Moilanen et al. [26] | 0.5171 | 0.6084 | 0.5595 |
| Davison et al. [42] | 0.8429 | 0.7041 | 0.7672 |
| Proposed Method | 0.9125 | 0.7304 | 0.8179 |

Each of the movement sequences are organised into folders based on the movement identification number. The ground truth FACS codes and onset, apex and offset frames are provided in an Excel form. The ID numbers will allow for easy cross-referencing.

The data, as shown in this paper, can be used for micro-expression classification tasks using the FACS coded ground truth for validation. Micro-movement spotting techniques can also use this data as the temporal phase information allows for accurate frame-positions of the micro-movements. All details required to request the raw data can be found with the information provided when the main dataset has been downloaded.

## 8 Discussion

The contributions of this paper focus on the dataset protocol and design, including an intention to make the dataset publicly available for researchers to expand the field of micro-facial expression analysis.

The protocol is designed to be personalised to each participant, rather than generalise the experiment. This process of personalisation involves tailoring each stimulus, mainly videos, to be best suited to what a participant will find emotionally stimulating. This is done by asking each person to fill in a pre-questionnaire before attending the experiment. This process takes more time than using the same stimuli repeatedly, however as each person responds differently to different stimuli, it makes sense to know what to show to a person to invoke a response. Choosing videos for the participant could potentially lead to some participants not feeling emotional towards the generalised stimuli. Previous micro-expression datasets ask participants after the experiment on what emotion they felt towards the stimulus to gauge what triggers their emotions.

Inducing spontaneous micro-facial expressions in a controlled environment can be difficult as people are consciously aware of the need to suppress emotions. To reduce this risk, the environment was set up in a room designed to be comfortable, and minimal distraction occurred between the observer and participant by remotely controlling all stimuli shown in an observation room. The ideal scenario would be to record micro-expressions when people are unaware of being recorded, however it would require the person to suppress without being asked and this raises ethical concerns on recording people without consent.

A further contribution, relating to the limitations of previous datasets, is providing a wider demographic of participants. Using a diverse ethnicity, age and even split of gender provides a better overall view of a population. Doing this contributes a deeper understanding of micro-facial expressions and how they occur across this demographic.

The resolution of the dataset increases significantly in relation to the lower resolutions captured in previous datasets and then used to attempt to detect and recognise micro-expressions. Previous experiments have found that resolution is more important when detecting micro-expressions, and a sharp reduction in true positive rate can be seen as resolution is scaled from $300 \times 310$ to $77 \times 80$ [20].

After a full count of AUs that were coded, a large proportion of the movements were classed as micro-movements, including 29.7 percent of movements up to 100 frames. From these results, shown in Table 5, the experiment was able to successfully induce a large amount of micro-movements relative to the total AU count. A $\chi^2$ test was also performed on the pooled frequency occurrences (see Table 6) of AUs in the dataset to observe the relationship between participant's reliable facial movements and the emotional stimuli they were exposed to. The results showed a highly significant difference in facial movements across different stimuli categories.

The initial results shown for this dataset attempts to take all movement blocks and non-movement blocks and create a descriptor that can describe the two features clearly. For larger facial expressions this can be a much easier task [43], [44] as the movements across people are more clearly defined. For micro-movements, the amount of permutations that subtle motion can take, along with a person's individual facial features, can make generalisation almost impossible.

Current state of the art in micro-expression detection carefully selects parameters that do not allow for real-world generalisation. The results in this paper also do not give an ideal generalisation method, as the performance is not high enough for a real-world scenario. However, the results show promise that further investigation could lead to a high performing micro-movement detection system in the near future.

## 9 Conclusion

### 9.1 Movement and Emotions

The main aim of the dataset is to create a collection of spontaneous micro-facial movements that are FACS coded. The focus is not on the emotional labels but the objective AUs. These movements can then be used for a variety of methods rather than just emotional recognition including, but not limited to, subtle motion detection as in this paper and lie detection [2].

The results using spatio-temporal methods and machine learning using SAMM are promising, however it opens up the potential for further improvements to the field, especially when it comes to generalising a human face so micro-movements can be located automatically without the need for a lab-controlled environment. Results from the proposed ABT outperform the previous state-of-the-art in both machine learning and difference analysis based approaches. Further experiments on other datasets would be advantageous to test the robustness of ABT, however the lack of baseline sequences within these datasets currently limit the experiments to SAMM.

### 9.2 Dataset Summary

To the best of our knowledge, the SAMM dataset has the highest resolution and includes a very diverse demographic of the micro-movement datasets currently available, giving a better representation of a population, which in turn allows for a variety of emotional responses like you would experience in a real-world scenario.

As the camera was recording at 200 fps, there are more frames to potentially reveal micro-movements, and are therefore not missed like they would in a conventional video camera. The high frame rate also means FACS coders have an easier time to coding by stepping through each frame in detail to obtain an accurate onset, apex and offset frame.

### 9.3 Future Work

Moving forward, the dataset will be tested on different methods of data representation to further investigate the ideal method to represent micro-movements. This includes testing optical flow based methods such as the MDMO feature [27], which would be best applied using a normalised block-based feature rather than applying it for individual block recognition as in this paper. Unsupervised clustering methods could be used to compare movements against a participant's baseline neutral face, however to detect micro-movements in real-time, the use of machine learning may have to be minimised or left out and be replaced with the proposed difference analysis methods.

sUsing a block-based approach for splitting the face into local temporal video cubes is relatively simple and has been used in other techniques for the analysis of movements [20], [26], [33], [42], [45]. However, by splitting into $m \times n$ blocks, the chance of introducing irrelevant facial features is higher. Recent approaches to this problem have used Delaunay triangulation [46] and specifically chosen regions of interest using facial feature points to define the region boundaries for analysis [27]. For detecting micro-movements, the 'noise' and unwanted data captured on the face, like head and eye movements, need to be minimised through a better definition of face regions based on FACS.

Other applications of the dataset include the study of using FACS-coded data for deception detection [47] and facial paresis [48], where people have a weakened ability to voluntarily move the muscles of the face. The dataset and method could also be used in the study of other issues that create facial twitches, and how to differentiate them from a suppressed or repressed emotional response that leads to micro-facial movements.

## Acknowledgments

The authors would like to acknowledge the Emotional Intelligence Academy for coding the dataset using FACS. Thanks also goes to all who participated in the emotional inducement experiment and expert consultations on the statistical analysis of data.

## References

[1] P. Ekman, *Emotions Revealed: Understanding Faces and Feelings*. Campbell, CA, USA: Phoenix, 2004.

[2] P. Ekman, "Lie catching and microexpressions," in *The Philosophy of Deception*, C. W. Martin, Ed. New York, NY, USA: Oxford Univ. Press, 2009, pp. 118–133.

[3] D. Matsumoto, S. H. Yoo, and S. Nakagawa, "Culture, emotion regulation, and adjustment." *J. Pers. Soc. Psychol.*, vol. 94, no. 6, pp. 925–937, 2008.

[4] X.-B. Shen, Q. Wu, and X.-L. Fu, "Effects of the duration of expressions on the recognition of microexpressions," *J. Zhejiang Univ. Sci. B*, vol. 13, no. 3, pp. 221–230, 2012.

[5] W.-J. Yan, Q. Wu, J. Liang, Y.-H. Chen, and X. Fu, "How fast are the leaked facial expressions: The duration of micro-expressions," *J. Nonverbal Behav.*, vol. 37, no. 4, pp. 217–230, 2013.

[6] P. Ekman, *Telling Lies: Clues to Deceit in the Marketplace, Politics, and Marriage*. New York, NY, USA: Norton, 2001.

[7] P. Ekman and E. L. Rosenberg, *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System*. New York, NY, USA: Oxford Univ. Press, 2005.

[8] M. G. Frank, C. J. Maccario, and V. l. Govindaraju, "Behavior and security," in *Protecting Airline Passengers in the Age of Terrorism*. Goleta, CA, USA: Greenwood Pub. Group, 2009.

[9] G. O. Paradiso, D. I. Cunic, C. A. Gunraj, and R. Chen, "Representation of facial muscles in human motor cortex," *J. Physiol.*, vol. 567, no. 1, pp. 323–336, 2005.

[10] D. Matsumoto and H. S. Hwang, "Evidence for training the ability to read microexpressions of emotion," *Motivation Emotion*, vol. 35, pp. 181–191, 2011.

[11] P. Ekman and R. J. Davidson, "Voluntary smiling changes regional brain activity," *Psychological Sci.*, vol. 4, no. 5, pp. 342–345, 1993.

[12] T. L. Kraft and S. D. Pressman, "Grin and bear it the influence of manipulated facial expression on the stress response," *Psychological Sci.*, vol. 23, no. 11, pp. 1372–1378, 2012.

[13] S. Polikovsky, Y. Kameda, and Y. Ohta, "Facial micro-expressions recognition using high speed camera and 3D-gradient descriptor," in *Proc. 3rd Int. Conf. Imag. Crime Detect. Prevention*, 2009, pp. 16–21. [Online]. Available: http://digital-library.theiet.org/content/conferences/10.1049/ic.2009.0244

[14] X. Li, T. Pfister, X. Huang, G. Zhao, and M. Pietikäinen, "A spontaneous micro-expression database: Inducement, collection and baseline," in *Proc. 10th IEEE Int. Conf. Automat. Face Gesture Recog.*, 2013, pp. 1–6.

[15] M. Shreve, S. Godavarthy, D. Goldgof, and S. Sarkar, "Macro- and micro-expression spotting in long videos using spatio-temporal strain," in *Proc. IEEE Int. Conf. Automat. Face Gesture Recog. Workshops*, 2011, pp. 51–56.

[16] W.-J. Yan, Q. Wu, Y.-J. Liu, S.-J. Wang, and X. Fu, "Casme database: A dataset of spontaneous micro-expressions collected from neutralized faces," in *Proc. 10th IEEE Conf. Workshops Automat. Face Gesture Recog.*, 2013, pp. 1–7.

[17] W.-J. Yan, et al., "Casme II: An improved spontaneous micro-expression database and the baseline evaluation," *PLoS ONE*, vol. 9, no. 1, Jan. 2014, Art. no. e86041.

[18] T. Pfister, X. Li, G. Zhao, and M. Pietikäinen, "Recognising spontaneous facial micro-expressions," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 1449–1456.

[19] S.-J. Wang, W.-J. Yan, X. Li, G. Zhao, and X. Fu, "Micro-expression recognition using dynamic textures on tensor independent color space," in *Proc. 22nd Int. Conf. Pattern Recog.*, 2014, pp 4678–4683.

[20] M. Shreve, J. Brizzi, S. Fefilatyev, T. Luguev, D. Goldgof, and S. Sarkar, "Automatic expression spotting in videos," *Image Vis Comput*, vol. 32, no. 8, pp. 476–486, 2014.

[21] S. Afzal and P. Robinson, "Natural affect data—collection & annotation in a learning context," in *Proc 3rd Int. Conf. Affect. Comput. Intell. Interact. Workshops*, 2009, pp. 1–7.

[22] P. Ekman and W. V. Friesen, *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Palo Alto, CA, USA: Consulting Psychologists Press, 1978.

[23] P. Ekman and W. V. Friesen, *Facial Action Coding System: Investigator's Guide*. Washington, DC, USA: Consulting Psychologists Press, 1978.

[24] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recog.*, 2005, pp. 886–893.

[25] S. Polikovsky, Y. Kameda, and O. Yuichi, "Facial micro-expression detection in hi-speed video based on facial action coding system (FACS)," *IEICE Trans. Inf. Syst.*, vol. 96, no. 1, pp. 81–92, Jan. 2013.

[26] A. Moilanen, G. Zhao, and M. Pietikainen, "Spotting rapid facial movements from videos using appearance-based feature difference analysis," in *Proc. 22nd Int. Conf. Pattern Recog.*, 2014, pp. 1722–1727.

[27] Y.-J. Liu, J.-K. Zhang, W.-J. Yan, S.-J. Wang, G. Zhao, and X. Fu, "A main directional mean optical flow feature for spontaneous micro-expression recognition," *IEEE Trans. Affect. Comput.*, 2015, Doi: 10.1109/TAFFC.2015.2485205.

[28] D. Patel, G. Zhao, and M. Pietikäinen, "Spatiotemporal integration of optical flow vectors for micro-expression detection," in *Advanced Concepts for Intelligent Vision Systems*. New York, NY, USA: Springer, 2015, pp. 369–380.

[29] P. J. Lang, "Behavioral treatment and bio-behavioral assessment: Computer applications," in *Technology in Mental Health Care Delivery Systems*, J. B. Sidowski, J. H. Johnson, and T. A. Williams, Eds. Norwood, NJ, USA: Ablex, 1980, pp. 119–137.

[30] M. M. Bradley and P. J. Lang, "Measuring emotion: The self-assessment manikin and the semantic differential," *J. Behav. Ther. Exp. Psychiatry*, vol. 25, no. 1, pp. 49–59, 1994. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0005791694900639

[31] A. Mehrabian and J. A. Russell, *An Approach to Environmental Psychology*. Cambridge, MA, USA: MIT Press, 1974.

[32] A. Mammucari, "Spontaneous facial expression of emotions in brain-damaged patients," *Cortex*, vol. 24, no. 4, pp. 521–533, 1988.

[33] A. K. Davison, M. H. Yap, N. Costen, K. Tan, C. Lansley, and D. Leightley, "Micro-facial movements: An investigation on spatio-temporal descriptors," in *Proc. ECCVW*, 2014, pp. 111–123.

[34] J. A. Ruiz-Hernandez, A. Lux, and J. L. Crowley, "Face detection by cascade of Gaussian derivates classifiers calculated with a half-octave pyramid," in *Proc. 8th Int. Conf. Automat. Face Gesture Recog.*, 2008, pp. 1–6.

[35] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, Jul. 2002.

[36] P. Dollár, Z. Tu, P. Perona, and S. Belongie, "Integral channel features." in *Proc. Brit. Mach. Vis. Conf.*, vol. 2, no. 3, 2009, pp. 19.1–19.11.

[37] P. Dollár. (2015). *Piotr's Computer Vision Matlab Toolbox (PMT)*. [Online]. Available: http://vision.ucsd.edu/~pdollar/toolbox/doc/index.html

[38] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part based models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–1645, Sep. 2010.

[39] M. Inc., Face++ research toolkit, Dec. 2013. [Online]. Available: www.faceplusplus.com

[40] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: An update," *ACM SIGKDD Explor. Newslett.*, vol. 11, no. 1, pp. 10–18, 2009.

[41] B. W. Matthews, "Comparison of the predicted and observed secondary structure of T4 phage lysozyme," *Biochimica et Biophysica Acta*, vol. 405, no. 2, pp. 442–451, 1975.

[42] A. K. Davison, M. H. Yap, and C. Lansley, "Micro-facial movement detection using individualised baselines and histogram-based descriptors," in *Proc. IEEE Int. Conf. Syst. Man Cybernetics*, 2015, pp. 1864–1869.

[43] B. Fasel and J. Luettin, "Automatic facial expression analysis: A survey," *Pattern Recog.*, vol. 36, pp. 259–275, 2003.

[44] M. S. Bartlett, G. C. Littlewort, M. G. Frank, C. Lainscsek, I. R. Fasel, and J. R. Movellan, "Automatic recognition of facial actions in spontaneous expressions," *J. Multimedia*, vol. 1, no. 6, pp. 22–35, 2006.

[45] G. Zhao and M. Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 6, pp. 915–928, Jun. 2007.

[46] Z. Lu, Z. Luo, H. Zheng, J. Chen, and W. Li, "A Delaunay-based temporal coding model for micro-expression recognition," in *Proc. Asian Conf. Comput. Vis. Workshops*, 2014, pp. 698–711.

[47] M. H. Yap, H. Ugail, and R. Zwiggelaar, "Facial behavioral analysis: A case study in deception detection," *Brit. J. Appl. Sci. Technol.*, vol. 4, no. 10, pp. 1485–1496, 2014.

[48] H. C. Hopf, W. Muller-Forell, and N. J. Hopf, "Localization of emotional and volitional facial paresis," *Neurology*, vol. 42, no. 10, pp. 1918–1918, 1992.

**Adrian K. Davison** received the BSc (Hons.) degree in multimedia computing from the Manchester Metropolitan University in 2012. He is currently working toward the PhD degree from the School of Computing, Mathematics and Digital Technology, Manchester Metropolitan University. He maintains an active role as a student representative at the MMU. Alongside he is a co-chaired the internal MMU Science and Engineering Symposium 2015. He is a student member of the IEEE.

**Cliff Lansley** received the graduate degree in education/psychology at Manchester University, United Kingdom. He has more than 25 years experience working at senior levels in public and private sector organisations facilitating leadership, communications, emotional intelligence and coaching programmes. His mission has been to gather the science and push forward the research that can 'harden' the often termed 'soft-skill' of emotional awareness and management (self and others) so that it is embraced more confidently by public/private employers and schools.

**Nicholas Costen** received the BA degree in experimental psychology from the University of Oxford, and the PhD degree in mathematics and psychology from the University of Aberdeen. He has undertaken research with the Advanced Telecommunications Research Laboratory, Kyoto, and the Division of Imaging Science and Biomedical Engineering, University of Manchester. He is currently a reader in Cognitive Computer Vision with Manchester Metropolitan University, where his interests include face recognition and human motion analysis. He is a member of the IEEE.

**Kevin Tan** received the BSc (Hons.) degree in computer science (software engineering) and the MSc degree in computer vision, visual, and virtual environments from the University of Leeds, and the PhD degree in bi-manual interaction within virtual environments from the Salford University, in 2008. After his PhD, he was a post-doctoral research assistant at the Materials Science Centre, University of Manchester. He is currently a senior lecturer at the Manchester Metropolitan University. His research interests are gamification, augmented and virtual reality for cross disciplinary application.

**Moi Hoon Yap** received the BSc (Hons.) degree in statistics and MSc degree in information technology from the Universiti Putra Malaysia, and the PhD degree in computer science from the Loughborough University, in 2009. After her PhD, she was a post-doctoral research assistant at the Centre for Visual Computing, University of Bradford. She is currently a senior lecturer at the Manchester Metropolitan University. Her research interests are facial analysis, medical image analysis, and image and video processing. She is a member of the IEEE.