Recognising Spontaneous Facial Micro-expressions

Tomas Pfister, Xiaobai Li, Guoying Zhao and Matti Pietikäinen
Machine Vision Group, Department of Computer Science and Engineering, University of Oulu
PO Box 4500, 90014 Oulu, Finland
{tpfister,lxiaobai,gyzhao,mkp}@ee.oulu.fi

Abstract

Facial micro-expressions are rapid involuntary facial expressions which reveal suppressed affect. To the best knowledge of the authors, there is no previous work that successfully recognises spontaneous facial micro-expressions. In this paper we show how a temporal interpolation model together with the first comprehensive spontaneous micro-expression corpus enable us to accurately recognise these very short expressions. We designed an induced emotion suppression experiment to collect the new corpus using a high-speed camera. The system is the first to recognise spontaneous facial micro-expressions and achieves very promising results that compare favourably with the human micro-expression detection accuracy.

1. Introduction

Humans are good at recognising full facial expressions which present a rich source of affective information [6]. However, psychological studies [4, 8] have shown that affect also manifests itself as micro-expressions. These are very rapid (1/3 to 1/25 second; the precise length definition varies [4, 13, 14]) involuntary facial expressions which give a brief glimpse to feelings that people undergo but try not to express. Currently only highly trained individuals are able to distinguish them, but even with proper training the recognition accuracy is only 47% [5]. We demonstrate that using temporal interpolation together with Multiple Kernel Learning (MKL) and Random Forest (RF) classifiers on a new spontaneous micro-expression corpus we can achieve very promising results that compare favourably with the human micro-expression detection accuracy.

There are numerous potential applications for recognising micro-expressions. Police can use micro-expressions to detect abnormal behaviour. Doctors can detect suppressed emotions in patients to recognise when additional reassurance is needed. Teachers can recognise unease in students and give a more careful explanation. Business negotiators can use glimpses of happiness to determine when they have proposed a suitable price. Since the human recognition accuracy is so low, an alternative method for recognising micro-expressions would be very valuable.

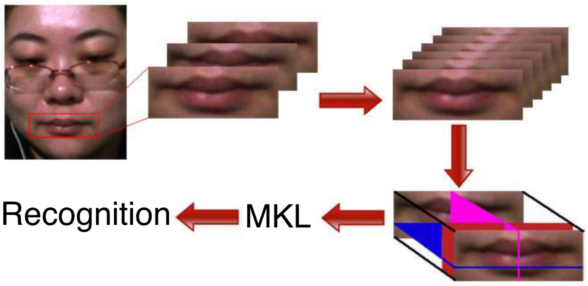


Figure 1. An example of a facial micro-expression (top-left) being interpolated through graph embedding (top-right); the result from which spatiotemporal local texture descriptors are extracted (bottom-right), enabling recognition with multiple kernel learning.

The major challenges in recognising micro-expressions involve their very short duration and involuntariness. The short duration means only a very limited number of frames are available for analysis with a standard 25fps camera. To allow accurate recognition, a camera with high frame rate can be used or an alternative method needs to be developed. Furthermore, with large variations in facial expression appearance, a supervised machine learning approach based on a training corpus is expected to best suit the problem. Acted facial expression corpora are least challenging to gather. However, since micro-expressions are involuntary [4], acted micro-expressions will likely differ greatly from spontaneous ones. Gathering a comprehensive training corpus therefore requires considerable psychological insights and time-consuming experiments to successfully induce spontaneous micro-expressions.

In this paper we propose a framework for recognising spontaneous facial micro-expressions that achieves very promising results. To the best of our knowledge, no previous work has successfully recognised these very short spontaneous facial expressions. Inside the framework, we use temporal interpolation to counter short video lengths, spa-

978-1-4577-1102-2/11/$26.00 ©2011 IEEE

tiotemporal local texture descriptors to handle dynamic features and $\{\mathrm{SVM}, \mathrm{MKL}, \mathrm{RF}\}$ to perform classification. To address the challenging task of collecting a training corpus of expressions that are involuntary we worked with psychologists to design an induced emotion suppression experiment. The resulting spontaneous micro-expression (SMIC) corpus was recorded using a 100fps high-speed camera. We demonstrate that temporal interpolation using graph embedding enables us to achieve equivalent micro-expression detection performance with a standard 25fps camera. The system is the first to recognise real spontaneous facial micro-expressions and achieves very promising results that compare favourably with the human accuracy.

2. Related Work

We present a brief history of micro-expressions and a summary of related work in psychology. We also provide a review of previous work on acted micro-expressions. For a more comprehensive summary of related work on facial expressions we refer the reader to a survey by Zeng *et al*. [19].

In analysing facial expressions, the main focus has long been on detecting the six basic emotions and to provide facial action unit labelling using FACS. Little work within computer vision has been done on analysing more complex facial expressions. In social psychology, however, micro-expressions as a complex form of facial expressions have been thoroughly studied by Gottman [8] and Ekman [4]. Ekman first discovered the existence of facial micro-expressions when examining a video of a psychiatric patient who tried to conceal a plan to commit suicide. By analysing the video in slow-motion, Ekman discovered a very short expression of intense anguish that was subsequently covered up by a smile. As these expressions are very rapid they are easily missed during casual observation. Micro-expressions have later been empirically observed in psychological studies [17] and training programs for learning to observe them have been created [4].

Studies of micro-expressions in psychology strongly suggest that humans are naturally weak at recognising micro-expressions. Frank *et al*. [5] conducted a micro-expression recognition test with real-life videos and found that US undergraduates and coast guards achieved accuracies of 32% and 25% without training and 40% and 47% with training respectively (chance is 20%), with very low absolute levels of detection. Further, Ekman [4] reports that even when showing micro-expressions out of context with no sound very few people are able to recognise them.

Most facial expression studies to date use training corpora consisting of people acting out facial expressions. These have been found to significantly differ from natural facial expressions occurring in everyday life [1]. As expected, focus is now shifting towards using induced and natural data for training [10]. These data are more challenging to deal with as they demand more freedom for expression and hence less control over recording conditions.

The few computer vision studies on recognising facial micro-expressions have used acted data, and none of the studies have made their data publicly available. Polikovsky *et al*. [13] gathered data from 10 university students acting out micro-expressions and used gradient orientation histogram descriptors. Shreve *et al*. [14, 15] similarly collected 100 acted facial micro-expressions from an unreported number of subjects and used strain patterns for feature description. Subjects were shown example videos of micro-expressions and were asked to mimic them.

However, micro-expressions are involuntary according to psychological research [4] and should not be elicitable through acting. Not surprisingly, Shreve *et al*. [14] report chance accuracy (50%) when attempting to evaluate their classifier on 24 spontaneous micro-expressions in the Canal-9 political debate corpus. Although the poor result may be partly attributable to head movement and talking, it is clear that to achieve reasonable practical accuracy a different method and a larger spontaneous facial micro-expression corpus are needed. In our paper we present both a significantly larger corpus of spontaneous facial micro-expressions and a method that succeeds at classification.

More tangentially related work includes Michael *et al*. [11] who proposed a method for automated deception detection using body movement. Although the authors briefly mention micro-expressions, no analysis of their occurrence in the training data is given.

Many successful facial expression recognition approaches to date have involved using spatiotemporal local texture descriptors. One such texture descriptor is LBP-TOP which has recently achieved state-of-the-art results in facial expression analysis [10, 20].

3. Proposed Method

Our proposed method to recognise facial micro-expressions combines a temporal interpolation model with state-of-the-art facial expression recognition methods. We first explain these in detail. As our study is the first to successfully recognise spontaneous facial micro-expressions, we briefly discuss the experimental method we employed to create the first comprehensive spontaneous micro-expression corpus.

3.1. Algorithm for Micro-expression Recognition

We illustrate how we apply a temporal interpolation model (TIM) together with state-of-the-art machine learning methods to successfully recognise facial micro-expressions with high accuracy. Algorithm 1 shows our framework for recognising spontaneous micro-expressions.

To address the large variations in the spatial appearances of micro-expressions, we crop and normalise the face ge-

Algorithm 1 Algorithm for recognising spontaneous micro-expressions. C is the corpus of image sequences c_i. Γ is the set of SLTD parameters where $x \times y \times t$ are the number of rows, columns and temporal blocks into which the SLTD feature extraction is divided. T is the set of frame counts into which image sequence c_i is temporally interpolated. LWM(Ψ, ω, ρ) computes the Local Weighted Mean transformation for frame ρ using feature points Ψ and model facial feature points ω according to Equation 1. The temporal interpolation variables are defined in Section 3.2. Poly$(q_{j,r}, q_{k,r}, d)$ and HisInt$(q_{j,r}, q_{k,r})$ compute the polynomial kernel of degree d and the histogram intersection kernel according to Equations 2 and 3. MKL-Phase1$(\boldsymbol{K})$ and MKL-Phase2$(\boldsymbol{K})$ output results from multiple kernel learning classifiers trained for Phase 1 (detection) and Phase 2 (classification) respectively.

```
Detect-Micro($C$)

1. Initialise $\Gamma = \{8\times8\times1, 5\times5\times1, 8\times8\times2, 5\times5\times2\}$
   and $T = \{10, 15, 20, 30\}$

2. For all $i.c_i \in C$ with frames $\rho_{i,1}...\rho_{i,s}$

   (a) Detect face $F_i$ in the first frame $\rho_{i,1}$
   (b) Extract $h$ facial feature points $\Psi = \{(a_1, b_1)...(a_h, b_h)\}$ from ASM
   (c) Normalise face to model face by computing LWM transformation
       $\zeta = \text{LWM}(\Psi, \omega, \rho_{i,1})$ where $\omega$ is the matrix of
       feature points for the model face and $\rho_{i,1}$ is the first neutral frame
   (d) Apply transformation $\zeta$ to frames $\rho_{i,2}...\rho_{i,s}$
   (e) Find eyes $E(F_i) = \{(x_{i,l}, y_{i,l}), (x_{i,r}, y_{i,r})\}$; set
       distance $\delta_i = \sqrt{(x_{i,l} - x_{i,r})^2 + (y_{i,l} - y_{i,r})^2}$
   (f) Crop face by setting topleft $= (x_{i,l}, y_{i,l}) + 0.4(y_{i,l} - y_{i,r}) - 0.6(x_{i,r} - x_{i,l})$;
       height $= 2.2\delta_i$; width $= 1.8\delta_i$
   (g) For all $\theta \in T$ compute TIM image sequence
       $\boldsymbol{\xi_{i,\theta}} = \boldsymbol{UM}\mathcal{F}^n(t) + \boldsymbol{\bar{\xi}_{i,\theta}}$
   (h) For all $p \in \Gamma, \theta \in T$ extract set $\mu_{i,p,\theta}(\boldsymbol{\xi_{i,\theta}}) = \{q_{i,p,\theta,1}...q_{i,p,\theta,M}\}$
       of SLTDs with SLTD feature vector length $M$

3. Compute kernels $\boldsymbol{K} = \{\forall j, k, m, \theta, p.c_j \in C \wedge$
   $c_k \in C \wedge m = 1...M \wedge \theta \in T \wedge p \in \Gamma \wedge r = (m, \theta, p)|$
   HisInt$(q_{j,r}, q_{k,r})$, Poly$(q_{j,r}, q_{k,r}, 2)$,
   Poly$(q_{j,r}, q_{k,r}, 6)\}$

4. If MKL-Phase1$(\boldsymbol{K})$ = micro, output classification
   result of MKL-Phase2$(\boldsymbol{K})$
```

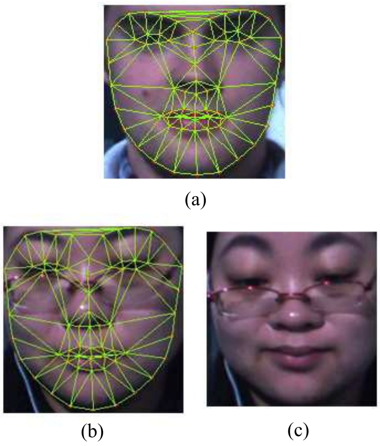


Figure 2. Normalisation in space domain to a model face. (a) is the model face onto which the feature points in the example face are mapped. (b) is the example face with its feature points detected. (c) shows the image after the feature points of the example face have been mapped to the model face.

ometry according to the eye positions from a Haar eye detector and the feature points from an Active Shape Model (ASM) [3] deformation. ASMs are statistical models of the shape of an object that are iteratively deformed to fit an example of the object. It starts the search from a mean shape aligned to the position and size of the face determined by a face detector and repeats until convergence: 1. suggest tentative shape by template matching of image texture around points to change feature point locations; and 2. fit tentative shape to the global shape model.

Using 68 ASM feature points shown in Figure 2 we compute a Local Weighted Mean (LWM) [7] transformation of frame $p_{i,1}$ for sequence i. LWM computes the weighted mean of all polynomials passing over each point by setting the value of an arbitrary point (x, y) to

$$f(x,y) = \frac{\sum_{i=1}^{N} V(\sqrt{(x-x_i)^2 + (y-y_i)^2}/R_n) S_i(x,y)}{\sum_{i=1}^{N} V(\sqrt{(x-x_i)^2 + (y-y_i)^2}/R_n)} \tag{1}$$

where $S_i(x, y)$ is the polynomial with n parameters passing through a measurement for control point (x_i, y_i) and $n - 1$ other measurements nearest to it, V is the weight and R_n is the distance of (x_i, y_i) from its $(n - 1)$th nearest control point in the reference image. We then apply the transformation to $p_{i,2}...p_{i,s}$ for an expression with s frames. Figure 2 illustrates the LWM transformation of the facial feature points in an example face to a model face. Haar eye detection results were checked against ASM feature points and were used to crop the image as shown in Algorithm 1.

We further temporally normalise all micro-expressions to a given set of frames $\theta \in T$. For each micro-expression image sequence i we compute a temporally interpolated image sequence $\boldsymbol{\xi_{i,\theta}} = \boldsymbol{UM}\mathcal{F}^n(t) + \bar{\boldsymbol{\xi}}_{\boldsymbol{i,\theta}}$ for all $\theta \in T$, where $\boldsymbol{U}$ is the singular value decomposition matrix, $\boldsymbol{M}$ is a square matrix, $\mathcal{F}^n(t)$ is a curve and $\bar{\boldsymbol{\xi}}_{\boldsymbol{i}}$ is a mean vector.

We then apply spatiotemporal local texture descriptors (SLTD) to the video for feature extraction. It is worth noting that SLTD require the input video to have a minimum length. In our case we use LBP-TOP [20] with radius $R = 3$ and the block sizes given in Algorithm 1. These parameters require the first and last 3 frames to be removed since the descriptor cannot be placed here [12]. To enable extraction of at least 1 frame for a segment we therefore need a minimum of 7 frames of data. With a 25fps camera 1/3 to 1/25 second micro-expressions would range between 1 to 8 frames. A method to derive more frames is therefore essential to use SLTD for any but the longest micro-expressions. Moreover, we would expect to achieve more statistically stable histograms with a higher number of frames. We demonstrate a temporal graph embedding method that achieves this in Section 3.2.

We use Multiple Kernel Learning (MKL) [16] to improve our classification results. Given a training set $H = \{(\boldsymbol{x_1}, l_1)...(\boldsymbol{x_n}, l_n)\}$ and set of kernels $\{K_1...K_M\}$ where $K_k \in \mathbb{R}^{n \times n}$ and K_k is positive semi-definite, MKL learns weights for linear/non-linear combinations of kernels over different domains by optimising a cost function $Z(K, H)$ where K is a combination of basic kernels. As shown in Algorithm 1, we combine polynomial kernels POLY of degrees 2 and 6 and a histogram-intersection kernel HISINT with different SLTD parameters $p \in \Gamma$ over different temporal interpolations $\theta \in T$ where

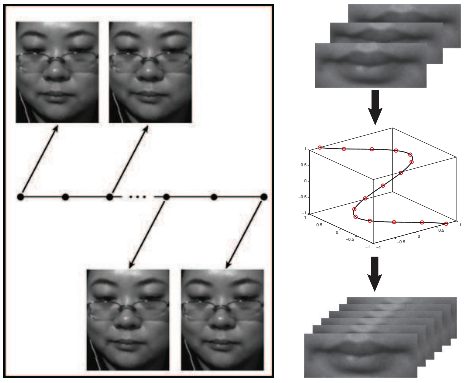
$$\textsc{Poly}(q_{j,r}, q_{k,r}, d) = (1 + q_{j,r} q_{k,r}^{\mathrm{T}})^d \qquad (2)$$

$$\textsc{HisInt}(q_{j,r}, q_{k,r}) = \sum_{a=1}^{b} \min\{q_{j,r}^a, q_{k,r}^a\} \qquad (3)$$

and $r = (m, \theta, p)$ and b is the number of bins in $q_{j,r}, q_{k,r}$. As alternative classifiers we use Random Forest and SVM. We ran pilot experiments to determine the optimal values of Γ and T for our corpora that are given in Algorithm 1.

Our classification system is two-phased. MKL-PHASE1($\boldsymbol{K}$) recognises the occurrence of a micro-expression. We train it with spatiotemporally normalised and size-balanced corpora labelled with $L = \{micro, \neg\, micro\}$.

If MKL-PHASE1($\boldsymbol{K}$) = micro, MKL-PHASE2($\boldsymbol{K}$) classifies the micro-expression into an arbitrary set of classes $L = \{l_1...l_n\}$. Dividing the task into two pipelined phases enables us to 1. optimise the phases separately; and 2. tailor L for Phase 2 to a given application whilst retaining the original optimised Phase 1. Further, because data labelling for Phase 2 requires a one step deeper analysis, it is subject to a greater labelling error. By separating the two phases we avoid a subjective labelling of expressions (Phase 2) affecting the performance of the more objective recognition step (Phase 1).



(a) (b)

Figure 3. (a) shows the graph representation of a micro-expression; (b) illustrates the temporal interpolation method in which a video is mapped onto a curve along which a new video is sampled.

3.2. Temporal Interpolation Model

In this subsection we show how we use graph embedding to interpolate images at arbitrary positions within a micro-expression. This allows us to input a sufficient number of frames to our feature descriptor even for very short expressions with very small number of frames. It also enables us to achieve more statistically stable feature extraction results by increasing the number of frames we use for extraction. Zhou *et al.* [21] previously proposed a similar method for synthesising a talking mouth. To the knowledge of the authors, this is the first time when such a method has been used to recognise facial expression.

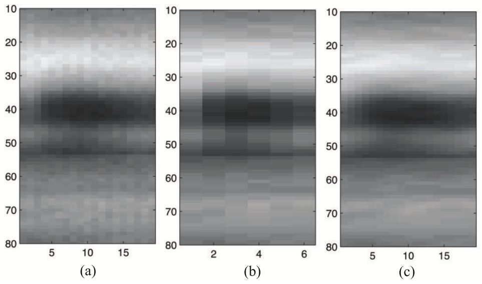
We view a video of a micro-expression as a set of images sampled along a curve and create a continuous function in a low-dimensional manifold by representing the micro-expression video as a path graph P_n with n vertices as shown in Figure 3. Vertices correspond to video frames and edges to adjacency matrix $\boldsymbol{W} \in \{0,1\}^{n \times n}$ with $W_{i,j} = 1$ if $|i - j| = 1$ and 0 otherwise. To embed the manifold in the graph we map P_n to a line that minimises the distance between connected vertices. Let $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)^{\mathrm{T}}$ be the map. We minimise

$$\sum_{i,j} (y_i - y_j)^2 \, W_{ij}, \quad i, j = 1, 2, \ldots, n \qquad (4)$$

to obtain $\boldsymbol{y}$, which is equivalent to calculating the eigenvectors of the Laplacian graph of P_n. We computed the Laplacian graph such that it has eigenvectors $\{\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_{n-1}\}$ and enables us to view $\boldsymbol{y}_k$ as a set of points described by

$$f_k^n(t) = \sin\left(\pi k t + \pi(n - k)/(2n)\right), t \in [1/n, 1] \qquad (5)$$

sampled at $t = 1/n, 2/n, \ldots, 1$. We can use the resulting



(a) (b) (c)

Figure 4. Temporal interpolation. The figures show vertical temporal patterns at the 121st column of an original, downsampled and interpolated micro-expression video. (a) shows the original 19-frame expression. (b) shows the downsampled 6-frame micro-expression. (c) shows (b) interpolated to 19 frames using TIM.

curve

$$\mathcal{F}^n(t) = \begin{bmatrix} f_1^n(t) \\ f_2^n(t) \\ \vdots \\ f_{n-1}^n(t) \end{bmatrix} \tag{6}$$

to temporally interpolate images at arbitrary positions within a micro-expression. To find the correspondences for curve $\mathcal{F}^n$ within the image space, we map the image frames to points defined by $\mathcal{F}^n(1/n), \mathcal{F}^n(2/n), \ldots, \mathcal{F}^n(1)$ and use the linear extension of graph embedding [18] to learn a transformation vector $\boldsymbol{w}$ that minimises

$$\sum_{i,j} \left(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i - \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_j\right)^2 W_{ij}, \quad i, j = 1, 2, \ldots, n \tag{7}$$

where $\boldsymbol{x}_i = \boldsymbol{\xi}_i - \bar{\boldsymbol{\xi}}$ is a mean-removed vector and $\boldsymbol{\xi}_i$ is the vectorised image. He *et al.* [9] solved the resulting eigenvalue problem

$$\boldsymbol{XLX}^{\mathrm{T}}\boldsymbol{w} = \lambda^{'}\boldsymbol{XX}^{\mathrm{T}}\boldsymbol{w} \tag{8}$$

by using the singular value decomposition with $\boldsymbol{X} = \boldsymbol{U\Sigma V}^{\mathrm{T}}$. Zhou *et al.* showed that we can interpolate a new image $\boldsymbol{\xi}$ by

$$\boldsymbol{\xi} = \boldsymbol{UM}\mathcal{F}^n(t) + \bar{\boldsymbol{\xi}} \tag{9}$$

where $\boldsymbol{M}$ is a square matrix. The validity of this interpolation depends on assuming that $\boldsymbol{\xi}_i$ are linearly independent. The assumption held for our corpora.

Figure 4 shows how the interpolation affects the vertical temporal pattern. It can be seen that the interpolated frames preserve well the characteristics of the original frames whilst smoothing the temporal profile.

We compute a temporally interpolated image sequence $\boldsymbol{\xi_{i,\theta}} = \boldsymbol{UM}\mathcal{F}^n(t) + \bar{\boldsymbol{\xi}}_{\boldsymbol{i,\theta}}$ for all $\theta \in T, c_i \in C$, compute all combinations of them with different SLTD block parameters Γ and choose the number of frames $\theta \in T$ and parameters $p \in \Gamma$ that maximise the accuracy for a given C.

3.3. Corpus 1: York Deception Detection Test (YorkDDT)

Warren *et al.* [17] recorded 20 videos for a deception detection test (DDT) as part of a psychological study. Subjects either truthfully or deceptively described an emotional or non-emotional film clip. The emotional clip was of a surgery and the non-emotional clip was a mildly positive depiction of a sunny beach. In the truthful scenario, subjects viewing the clips were asked to describe their actual content. In the deceptive scenario, subjects viewing an emotional clip were asked to describe the non-emotional clip, and vice versa. The authors reported a number of micro-expressions occurring during both deceptive and truthful scenarios. The videos were recorded at 320x240 resolution.

We obtained the original DDT video, segmented the micro-expressions and labelled them as (truthful/deceptive, emotional/non-emotional) according to which scenarios the video corresponded to. Micro-expressions were found in 9 subjects (3 male and 6 female). This gave 18 micro-expressions: 7 from the emotional and 11 from the non-emotional scenario; 11 from the deceptive and 7 from the truthful scenario. The shortest expression was 7 frames at the 25fps video frame rate.

3.4. Corpus 2: New 100fps Spontaneous Micro-expression Corpus (SMIC)

The YorkDDT corpus suffers from small-training-sample-size (STSS) and low-resolution problems. To solve these problems we gathered a new corpus.

Our new corpus records spontaneous facial micro-expressions using a 100fps camera. The initial corpus consists of 6 subjects (3 male, 3 female) with 77 spontaneous micro-expressions. Four subjects wore glasses.

The corpus was recorded in an indoor bunker environment designed to resemble an interrogation room. A PixeLINK PL-B774U camera running at 640x480 with 100fps was used. Each subject was recorded watching 16 carefully selected film clips chosen to induce disgust, fear, happiness, sadness and surprise. The experimental instructions were: 1. attempt to suppress your facial expressions whilst carefully watching the clips; 2. experimenters are watching your face and are trying to guess which film clip you are watching; 3. if your facial expression leaks and the experimenter guesses the clip you are watching correctly you will be asked to fill in a dull 500-question survey as a punishment; and 4. after each clip fill in a short self-report questionnaire specifying what emotions you experienced.

An interrogation room setting with a punishment threat and highly emotional clips were chosen to create a high-stake situation where subjects undergoing high emotional arousal are motivated to suppress their facial expressions. Ekman [4] previously argued that these are the ideal conditions for inducing micro-expressions. Although the stake in our experiments is rather low, our results show that the combination is very successful in inducing micro-expressions.

In total 210 minutes of data with 1 260 000 frames were obtained. The data were segmented and labelled by two annotators according to subjects' self-reported emotions. The annotators followed the advice by Ekman to first view the video frame by frame and then with increasing speed. The shortest recorded expression was about 1/9 seconds (11 frames at 100fps) and the average expression length was about 3/10 seconds (29 frames). We are in the process of adding more subjects to the corpus.

4. Experiments and Results

We evaluate our proposed micro-expression recognition system by leave-one-subject-out evaluation on two corpora.

As the SLTD our experiments use LBP-TOP. For MKL we use the block sizes given in Algorithm 1. Non-MKL classification results are reported with $\text{SLTD}_{8\times8\times1}$, where the image is split in 8×8 blocks in the spatial domain. SVM results without MKL use a polynomial kernel of degree 6. We report the results for combinations of parameters $p \in \Gamma$, $\theta \in T$ and classifiers $\phi = \{\text{SVM}, \text{MKL}, \text{RF}\}$ that gave the best leave-one-subject-out results. RF is the Random Forest [2] decision tree ensemble classifier. The system can be used for a general detection task by classifying a sliding window of facial frames. The experiments are equivalent to off-line runs of the sliding window classifier and do not require manual segmentation.

4.1. Experiment 1: YorkDDT Corpus

YorkDDT poses several challenges common in practical micro-expression recognition. First, the data is spontaneous and hence has high variability. Second, the subjects are constantly talking, so facial movement is not limited to facial expressions. Third, the resolution and frame rate of the camera are very limited.

Despite these inherit challenges in the corpus, we show that using the methods described in Section 3 we can successfully build a subject-independent micro-expression recognition system using this limited corpus.

Table 1 shows a summary of the leave-one-subject-out results on the YorkDDT corpus.

Phase 1 distinguishes micro-expressions from other facial activity. For the purpose of this experiment, we randomly selected 18 image sequences from sections of the data that did not contain any micro-expressions, but were allowed to contain speaking or facial expressions that were not micro-expressions. Using $\text{SLTD}_{8\times8\times1}$ with an SVM we achieve 65% accuracy. By combining temporal interpolation of all expressions to 10 frames with the MKL kernels computed on the SLTD block sizes given in Algorithm 1 we achieve 83% leave-one-subject-out accuracy. Interpolating to over 10 frames did not yield any significant improvement. This may be because the original videos are fairly short, so interpolating to more than 10 frames only adds redundant data and leads to deteriorating performance due to the *curse of dimensionality*. However, we see a very significant boost of 15% by normalising all sequences to 10 frames.

Phase	Classes	Method	Accuracy (%)
1	detection	SVM	65.0
1	detection	MKL	67.0
1	detection	MKL+TIM10	**83.0**
2	lie/truth	SVM	47.6
2	lie/truth	MKL	57.1
2	lie/truth	MKL+TIM10	**76.2**
2	emo/¬ emo	SVM	69.5
2	emo/¬ emo	MKL	**71.5**
2	emo/¬ emo	MKL+TIM10	**71.5**

Table 1. Leave-one-subject-out results on the YorkDDT corpus. MKL denotes Multiple Kernel Learning; TIMn denotes temporal interpolation to n frames; emo/¬ emo denotes classifying emotional vs. unemotional micro-expressions.

Phase 2 recognises the type of a micro-expression. For the YorkDDT corpus we have two sets of labels: emotional vs. non-emotional and deceptive vs. truthful.

For distinguishing deceptive from truthful micro-expressions, without MKL or TIM we achieve a below-chance accuracy 47.6% with an SVM trained on $\text{SLTD}_{8\times8\times1}$. By combining TIM10 and MKL with our selection of SLTD block sizes and kernels we boost the result to 76.2%. Again, interpolating to a higher number of frames did not yield any significant improvement. Out of the three classifiers in ϕ MKL constantly yielded the best result.

Combining Phase 1 and Phase 2 corresponds to pipelining the videos of positive detections from Phase 1 to be classified by Phase 2 as shown in Algorithm 1. Since 83% of the micro-expressions are correctly detected using MKL with TIM to 10 frames, we can detect and classify deceptive/truthful and emotional/unemotional micro-expressions with 63.2% and 59.3% accuracy respectively. Such a pipelined system could be used to detect lies by requiring $\text{MKL-P{\small HASE}1}(\boldsymbol{K}) = \text{micro} \wedge \text{MKL-P{\small HASE}2}(\boldsymbol{K}) = \text{lie}$.

4.2. Experiment 2: SMIC Corpus

In our new SMIC corpus we addressed the resolution and frame rate problems of the YorkDDT corpus. A summary of the results are given in Table 2.

The most notable difference in the results compared to the YorkDDT corpus is that whereas TIM still gave high performance boosts, MKL with different combinations of kernels $\boldsymbol{K}$, TIMs $\theta \in T$ and parameters $p \in \Gamma$ did not always offer the best performance, but was occasionally outperformed by decision tree ensemble classifier Random

Phase	Classes	Method	Accuracy (%)
1	detection	RF+TIM15	67.7
1	detection	SVM	70.3
1	detection	RF+TIM20	70.3
1	detection	MKL	71.4
1	detection	RF+TIM10	**74.3**
2	neg/pos	SVM	54.2
2	neg/pos	SVM+TIM15	59.8
2	neg/pos	MKL	60.2
2	neg/pos	MKL+TIM10	**71.4**

Table 2. Leave-one-subject-out results on the SMIC corpus. MKL denotes Multiple Kernel Learning; TIMn denotes temporal interpolation to n frames; RF denotes the Random Forest decision tree classifier; neg/pos denotes classifying negative vs. positive micro-expressions.

Forest. This demonstrates that the optimal classifier depends on the data and highlights that alternative classifiers should always be investigated. Fusion with MKL would potentially yield even better performance.

A notable similarity to the experiments on YorkDDT is that TIM10 continued performing well. This is even though the frame rate quadrupled from 25fps in YorkDDT to 100fps in SMIC. TIM to 10 frames in fact downsamples micro-expressions from the average length of 29 frames. This indicates that a higher frame rate may produce redundant data that deteriorates the performance of the classifier.

In Phase 1 we distinguish micro-expressions from other facial data. As for the YorkDDT corpus, we randomly selected 77 image sequences of the data that did not contain facial micro-expressions but could contain other facial movement. By running an SVM on this data we achieve 70.3% micro-expression detection accuracy. Using MKL we improve our results slightly. The highest improvement to 74.3% is achieved by using the Random Forest decision tree classifier together with TIM to 10 frames.

In Phase 2 we classify the recognised micro-expression as negative vs. positive using 18 and 17 samples respectively. With SVM only we achieve a rather poor accuracy of 54.2% (50% chance). However, by incorporating MKL and temporal interpolation we improve the result to 71.4%.

4.3. Experiment 3: Recognising Micro-expressions with Standard 25fps Frame Rate

In an ideal case, spontaneous micro-expression recognition would work with standard cameras without special hardware. In this experiment we show how our temporal interpolation method enables high recognition accuracy even when using a standard 25fps frame rate.

We downsampled the 100fps SMIC corpus by selecting every 4th frame from the original data. This resulted in sequences of length between 2 and 8 frames.

Phase	Classes	Method	Accuracy (%)
1	detection	RF+TIM10	58.5
1	detection	SVM+TIM10	65.0
1	detection	MKL+TIM10	70.3
1	detection	RF+TIM15	76.3
1	detection	RF+TIM20	**78.9**
2	neg/pos	SVM+TIM10	51.4
2	neg/pos	MKL+TIM10	60.0
2	neg/pos	MKL+TIM10	60.0
2	neg/pos	SVM+TIM15	62.8
2	neg/pos	MKL+TIM15	**64.9**

Table 3. Leave-one-subject-out results on the SMIC corpus downsampled to 25fps. MKL denotes Multiple Kernel Learning; TIMn denotes temporal interpolation to n frames; RF denotes the Random Forest decision tree classifier; neg/pos denotes classifying negative vs. positive micro-expressions.

The results on the downsampled SMIC corpus are shown in Table 3. As explained in Section 3.1 our SLTD method only allows sequences with at least 7 frames to be used, so TIM is essential for the downsampled corpus. We notice a 5.3% increase in accuracy for Phase 1 to 70.3% from using an MKL kernel set instead of pure SVM. The best results were achieved by combining the Random Forest classifier with a temporal interpolation to 20 frames, yielding 78.9% accuracy. This compares favourably with the human micro-expression recognition accuracy reported by Frank *et al*. [5].

For Phase 2 classification we significantly improve the results through MKL and TIM to 15 frames (64.9%), but are not quite able to match the performance achieved with the full 100fps frame rate (71.4%). For Phase 1 detection, however, even with $1/4$ less frames we are able to match and even slightly outperform the accuracy achieved with the full data from the 100fps camera. This is not surprising as detection only requires a small movement in the relevant facial region to be recognised. This could intuitively be done even with a small number of frames. A higher frame rate may only add irrelevant features and hence lower the performance. Classification of the micro-expression type, however, may need more detailed spatiotemporal data about the movement, so a higher frame rate is more useful.

We further experimented on how the recognition rate varies with frame rate. The results of downsampling the SMIC corpus to various frame rates are given in Figure 5. We note that the result is relatively stable using TIM20. Without interpolation we get a more significant drop in accuracy with a lower frame rate. We observed the peak accuracy when downsampling the corpus to 50fps. This suggests that by using TIM a camera with lower frame rate is sufficient for accurately classifying micro-expressions.

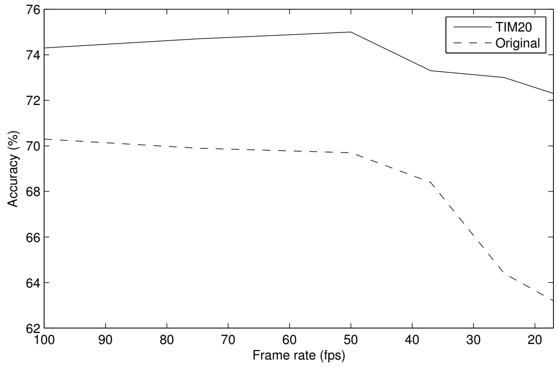


Figure 5. Variation of Phase 1 detection accuracy with level of downsampling for the SMIC corpus. The solid line shows the leave-one-subject-out accuracy with MKL and TIM to 20 frames; the dotted line shows accuracy with MKL only.

5. Conclusions

We have shown the first framework to successfully recognise spontaneous facial micro-expressions. Inside the framework, we use temporal interpolation to counter short video lengths, SLTD to handle dynamic features and $\{\text{SVM}, \text{MKL}, \text{RF}\}$ to perform classification. We designed an induced emotion suppression experiment which successfully induced 77 spontaneous micro-expressions in 6 subjects. We evaluated the system on two new corpora and achieved very promising results. We showed that temporal interpolation enables our system to match the micro-expression detection accuracy of a 100fps camera even with a standard 25fps frame rate. The system is the first to successfully recognise spontaneous micro-expressions and achieves very promising results that compare favourably with the human micro-expression detection accuracy.

Future work includes expanding the SMIC corpus to more participants, comparing our system to the performance achieved by trained humans on our dataset, enabling real-time recognition, and investigating whether classifier fusion methods yield performance improvements. We publish the SMIC corpus and the micro-expression recognition code for public use[1] to encourage further work in this area.

Acknowledgements. The authors are grateful to Ziheng Zhou for his advice and to Peter Bull and Gemma Warren for sharing the YorkDDT. The research was sponsored by Infotech Oulu and the Academy of Finland.

References

[1] S. Afzal and P. Robinson. Natural affect data-collection & annotation in a learning context. In *ACII*, pages 1–7, 2009. 2

[2] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. 6

[3] T. Cootes, C. Taylor, D. Cooper, and J. Graham. Active shape models – their training and application. *Computer Vision and Image Understanding*, 61(1):38–59, 1995. 3

[4] P. Ekman. Lie catching and microexpressions. *The Philosophy of Deception*, Oxford University Press, 2009. 1, 2, 5

[5] M. G. Frank, M. Herbasz, K. Sinuk, A. Keller, and C. Nolan. I see how you feel: Training laypeople and professionals to recognize fleeting emotions. *International Communication Association, Sheraton New York City*, 2009. 1, 2, 7

[6] A. Freitas-Magalhães. The psychology of emotions: The allure of human face. *Uni. Fernando Pessoa Press*, 2007. 1

[7] A. Goshtasby. Image registration by local approximation methods. *IMAVIS*, 6(4):255–261, 1988. 3

[8] J. Gottman and R. Levenson. A two-factor model for predicting when a couple will divorce: Exploratory analyses using 14-year longitudinal data. *Family process*, 41(1):83–96, 2002. 1, 2

[9] X. He, D. Cai, S. Yan, and H. Zhang. Neighborhood preserving embedding. In *ICCV*, pages 1208–1213, 2005. 5

[10] S. Koelstra, M. Pantic, and I. Patras. A dynamic texture based approach to recognition of facial actions and their temporal models. *PAMI*, 32(11):1940–1954, 2010. 2

[11] N. Michael, M. Dilsizian, D. Metaxas, and J. Burgoon. Motion profiles for deception detection using visual cues. In *ECCV*, pages 462–475, 2010. 2

[12] T. Ojala, M. Pietikäinen, and T. Mäenpää. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *PAMI*, 24(7):971–987, 2002. 4

[13] S. Polikovsky, Y. Kameda, and Y. Ohta. Facial micro-expressions recognition using high speed camera and 3D-gradient descriptor. In *ICDP*, 2009. 1, 2

[14] M. Shreve, S. Godavarthy, D. Goldgof, and S. Sarkar. Macro- and micro-expression spotting in long videos using spatio-temporal strain. In *FG*, 2011. 1, 2

[15] M. Shreve, S. Godavarthy, V. Manohar, D. Goldgof, and S. Sarkar. Towards macro- and micro-expression spotting in video using strain patterns. In *Workshop on Applications of Computer Vision*, pages 1–6, 2010. 2

[16] M. Varma and D. Ray. Learning the discriminative power-invariance trade-off. In *ICCV*, pages 1–8, 2007. 4

[17] G. Warren, E. Schertler, and P. Bull. Detecting deception from emotional and unemotional cues. *J. Nonverbal Behavior*, 33(1):59–69, 2009. 2, 5

[18] S. Yan, D. Xu, B. Zhang, H. Zhang, Q. Yang, and S. Lin. Graph embedding and extensions: A general framework for dimensionality reduction. *PAMI*, 29(1):40–51, 2007. 5

[19] Z. Zeng, M. Pantic, G. Roisman, and T. Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *PAMI*, 31(1):39–58, 2008. 2

[20] G. Zhao and M. Pietikäinen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *PAMI*, 29(6):915–928, 2007. 2, 4

[21] Z. Zhou, G. Zhao, and M. Pietikäinen. Towards a Practical Lipreading System. In *CVPR*, 2011. 4

[1] http://tomas.pfister.fi/micro-expressions