Differentiating Spontaneous From Posed Facial Expressions Within a Generic Facial Expression Recognition Framework

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

Abstract

In this paper we propose the first method known to the authors that successfully differentiates spontaneous from posed facial expressions using a realistic training corpus. We propose a new spatiotemporal local texture descriptor (CLBP-TOP) that outperforms other descriptors. We demonstrate that our temporal interpolation and visual/near-infrared fusion methods improve the differentiation performance. Finally, we propose a new generic facial expression recognition framework that subdivides the facial expression recognition problem into a cascade of smaller tasks that are simpler to tackle. The system is the first to differentiate spontaneous from posed facial expressions with a realistic corpus and achieves promising results.

1. Introduction

Facial expressions can be broadly classified as either spontaneous (genuine) or posed (faked). For example, posed smiles often only involve movement of the mouth (zygomaticus) whereas spontaneous smiles are more symmetrical and also include the muscles surrounding the eyes (orbicularis oculi) [4]. We propose the first method known to the authors to successfully differentiate spontaneous vs. posed (SVP) facial expressions with a realistic corpus.

There are numerous potential applications for SVP differentiation. Police can use SVP in surveillance systems to detect deceptive facial expressions. Doctors can recognise when patients are experiencing genuine pain so that their pain is taken seriously. Researchers can subdivide the facial expression recognition problem into two tasks and separately optimise their solutions.

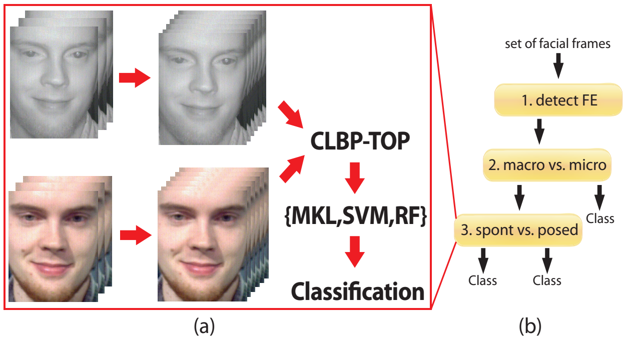

Figure 1. (a) shows an example of a visual and near-infrared facial expression (left) being interpolated through graph embedding (middle); the result from which a spatiotemporal local texture descriptor is extracted (right), enabling differentiation of spontaneous from posed expressions using multiple kernel learning, support vector machines and random forests. (b) shows a generic facial expression recognition framework where a facial expression is detected (top); then classified as a micro or macro-expression (middle); after which spontaneous and posed macro-expressions are differentiated (bottom).

Work on facial expressions in computer vision originally focused on posed expressions, with the focus now shifting towards recognising spontaneous expressions in more realistic situations. Very recently pioneering work was done in facial micro-expression recognition [10] aiming to recognise very short, involuntary spontaneous facial expressions that reveal suppressed affect. As an extension to our work on SVP differentiation we propose a generic facial expression recognition framework that subdivides the facial expression recognition problem into a cascade of application-independent spontaneous vs. posed and micro vs. macro-expression classifiers and application-specific micro, posed and spontaneous expression classifiers.

The main contributions of this paper are: 1. to the best knowledge of the authors the method is the first to differentiate spontaneous from posed facial expressions with a realistic corpus; 2. it proposes a new spatiotemporal local texture descriptor CLBP-TOP which we demonstrate outperforms other descriptors; and 3. it proposes a generic classification cascade that divides the facial expression recognition problem into small simple subtasks. We show that SVP differentiation benefits from both temporal interpolation and near-infrared images. Our system is the first to differentiate spontaneous from posed facial expressions with a realistic corpus and achieves very promising results.

2. Related Work

Very few studies in computer vision have investigated the general problem of differentiating between spontaneous and posed facial expressions. Valstar *et al.* [13] studied SVP differentiation based on eyebrow movements. The authors extended their work in Valstar *et al.* [12] to distinguish spontaneous from posed smiles using head, face and shoulder modalities. However, in both studies data for spontaneous and posed expressions were selected from different corpora. Of these, the spontaneous corpus has so far not been made publicly available. It would be valuable to evaluate the system on a corpus in which data for spontaneous and posed expressions were selected from same corpora so that the classification would be focused more on SVP differences than corpus differences.

Other studies have investigated SVP differentiation for specific emotions but have not developed a generic SVP classifier. Bartlett *et al.* [1] used Gabor wavelet decomposition to differentiate spontaneous from posed pain with 72% accuracy. Cohn and Schmidt [3] used geometric features to SVP differentiate smiles and found temporal patterns particularly useful, highlighting that the task could benefit from spatiotemporal local texture descriptors.

Wang *et al.* [15] collected a corpus of natural and infrared spontaneous and posed facial expressions. However, they only recorded the apexes for posed facial expressions, making it impossible to exploit temporal differences between spontaneous and posed expression to differentiate between spontaneous and posed facial expressions. Zhao *et al.* [16] found that near-infrared facilitates facial expression recognition, providing a degree of illumination invariance. We show that near-infrared is also very useful for differentiating between spontaneous and posed facial expressions.

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 [9, 17]. Recent work has also investigated temporal models. Shan *et al.* [11] proposed a Bayesian temporal manifold model for deriving a probability distribution measure of posed facial expressions.

3. Proposed Method

Our proposed SVP differentiation method combines temporal interpolation with a new spatiotemporal feature descriptor and state-of-the-art machine learning methods. Unlike previous work, the system presented in this paper 1. eliminates classification of corpus differences by using a new specially collected corpus with both spontaneous and posed expressions; 2. classifies general SVP for any facial expression instead of classifying specific expressions; 3. is the first to also consider near-infrared for SVP differentiation; 4. makes the SVP corpus publicly available; and 5. uses cues from the whole face with a new powerful spatiotemporal local texture descriptor.

We first discuss our methods in detail. As our work is the first to create a realistic corpus for SVP differentiation, we briefly discuss the experimental setting we used to collect the corpus. We finally describe pilot work on a generic facial expression recognition cascade that combines spontaneous and posed facial expression differentiation with recent work on facial micro-expressions to analyse arbitrary facial expressions.

3.1. Differentiating Spontaneous from Posed Facial Expressions

We illustrate how we apply a temporal interpolation model (TIM) together with a new feature descriptor and state-of-the-art machine learning methods to successfully differentiate spontaneous from posed facial expressions with high accuracy. Algorithm 1 shows our framework for SVP differentiation.

Spontaneous facial expressions last a varying length of time depending on subjects and the emotion-eliciting context. However, the length of posed facial expressions is normally set by the posing instructions. To improve generalisation we temporally normalise all facial expressions to a given set of frames $\theta \in T$. For each facial expression image sequence s we compute a temporally interpolated image sequence $\boldsymbol{\xi_{s,\theta}} = \boldsymbol{UM}\mathcal{F}^n(t) + \bar{\boldsymbol{\xi}}_{\boldsymbol{s,\theta}}$ 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 provide details of the temporal graph embedding method in Section 3.4. Temporal normalisation is done after the face has been cropped using the eye positions from a Haar eye detector.

We then apply spatiotemporal local texture descriptors (SLTD) to the video for feature extraction. In particular, LBP-TOP has recently achieved state-of-the-art results in facial expression analysis [9]. We show that by using magnitude differences to neighbours and the centre grey level in addition to signs we can reach significantly better results. The details of this new SLTD are discussed in Section 3.3.

We use Multiple Kernel Learning (MKL) [14] 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 $Z(K, H)$

Algorithm 1 Algorithm for SVP differentiation. N is the base corpus of near-infrared (NIR) and visual (VIS) corpora. C is any subcorpus 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. The temporal interpolation variables are defined in Section 3.4. $\textsc{Poly}(q_{k,r}, q_{o,r}, d)$ and $\textsc{HisInt}(q_{k,r}, q_{o,r})$ compute the polynomial kernel of degree d and the histogram intersection kernel. SVP($\boldsymbol{K}$) returns the result from a multiple kernel learning classifier trained to distinguish spontaneous from posed facial expressions.

```
LAYER3-SVP($N$)

1. Initialise $\Gamma = \{8 \times 8 \times 1, 8 \times 8 \times 2, 8 \times 8 \times 3\}$ and $T = \{10, 20, 25, 30\}$
2. For all $i.C_i \in N$
   (a) For all $j.c_j \in C_i \wedge s = (i, j)$ with frames $\rho_{s,1} \cdots \rho_{s,t}$
       i. Detect face $F_s$ in the first frame $\rho_{s,1}$
       ii. Find eyes $E(F_s)$
       iii. Crop face using $E(F_s)$ with the formula from [10]
       iv. For all $\theta \in T$ compute TIM image sequence $\boldsymbol{\xi_{s,\theta}} = \boldsymbol{UM}\mathcal{F}^n(t) + \boldsymbol{\bar{\xi}_{s,\theta}}$
       v. For all $p \in \Gamma, \theta \in T$ extract $\mu_{s,p,\theta}(\boldsymbol{\xi_{s,\theta}}) = \{q_{s,p,\theta,1} \cdots q_{s,p,\theta,m} \cdots q_{s,p,\theta,M}\}$ set of SLTDs where $M$ is the length of the SLTD feature vector
3. Compute $\boldsymbol{K} = \{\forall k, m, o, C, \theta, p. C \in N \wedge c_k \in C \wedge m = 1...M \wedge c_o \in C \wedge \theta \in T \wedge p \in \Gamma \wedge r = (m, \theta, p) | \textsc{HisInt}(q_{k,r}, q_{o,r}), \textsc{Poly}(q_{k,r}, q_{o,r}, 2), \textsc{Poly}(q_{k,r}, q_{o,r}, 6)\}$
4. Output SVP($\boldsymbol{K}$)
```

where K is a combination of basic kernels. As illustrated in Algorithm 1, we combine a histogram-intersection kernel HISINT and polynomial kernels POLY of degrees 2 and 6 with different SLTD parameters $p \in \Gamma$ over different temporal interpolations $\theta \in T$ where

$$\textsc{HisInt}(q_{k,r}, q_{o,r}) = \sum_{a=1}^{b} \min \{q_{k,r}^a, q_{o,r}^a\} \tag{1}$$

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

and b is the number of bins in $q_{k,r}, q_{o,r}$ and $r = (m, \theta, p)$. As alternative classifiers we use Random Forest, SVM, LINEAR and their fusion through majority voting. We ran pilot experiments to determine the optimal values of Γ and T for our corpora that are given in Algorithm 1. Finally, SVP($\boldsymbol{K}$) either runs MKL on the computed kernels or another classifier $w \in \phi$ on SLTD features μ to differentiate spontaneous from posed facial expressions.

3.2. New SVP Differentiation Corpus (SPOS)

Previous approaches to SVP differentiation suffer from using different corpora and different subjects for each training class, thereby to some extent learning to differentiate the corpora in addition to differentiating SVP.

Our new spontaneous vs. posed (SPOS) corpus provides spontaneous and posed expressions for the same subjects in the same session. This allows us to train for SVP differentiation rather than corpus differentiation. The initial corpus consists of 7 subjects (4 male and 3 female; 4 Asian and 3 Caucasian) with 84 posed and 147 spontaneous expressions. Five subjects wore glasses.

The corpus was recorded in an indoor bunker environment. Two cameras running at 640x480 with 25fps were used, one recording data from the visual and the other from the near-infrared spectrum. The two streams were automatically synchronised with manual checking to ensure consistency. Each subject was recorded watching 14 carefully selected film clips chosen to induce 6 basic emotions. The emotions with the number of spontaneous expressions in parenthesis are: anger (13), disgust (20), fear (32), happiness (66), sadness (5) and surprise (11). After the experiment subjects were asked to pose each expression twice, yielding 12 posed expressions per subject.

In total 720 minutes of data with 1 080 252 frames were obtained. The data were segmented and labelled for onset, apex, offset and end by two annotators according to subjects' self-reported emotions. For our experiments we focus on the onset phase since the length from onset to end of spontaneous expressions can vary significantly and be very long. In total 22462 frames from onset to apex were used for classification. The average expression lengths were about 6 seconds (147 frames) and 13 seconds (323 frames) for posed and spontaneous expressions respectively. The average lengths from onset to apex were about 1 seconds (28 frames) and 3 seconds (69 frames). We are in the process of adding more subjects to the corpus.

3.3. CLBP-TOP

Completed local binary patterns (CLBP) proposed by Guo *et al.* [6] represent the original image as its centre grey level (C) and the sign (S) and magnitude (M) of the local difference $d_p = g_p - g_c$ where g_c is the central pixel with P circularly and evenly spaced neighbours $g_p, p = 0, 1, \ldots, P-1$. The local difference d_p is decomposed into

40	35	14
96	**15**	6
2	27	66

(a)

25	20	-1
81		-9
-13	12	51

(b)

1	1	-1
1		-1
-1	1	1

(c)

25	20	1
81		9
13	12	51

(d)

Figure 2. (a) A sample 3×3 block; (b) the local difference d_p; (c) the sign component S and (d) magnitude component M.

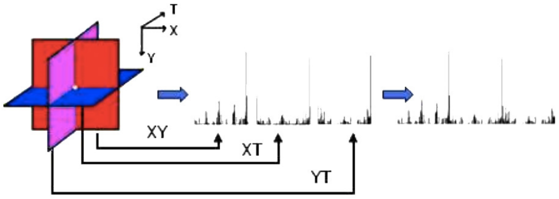

Figure 3. Concatenated histogram from the three planes XY, XT and YT [17].

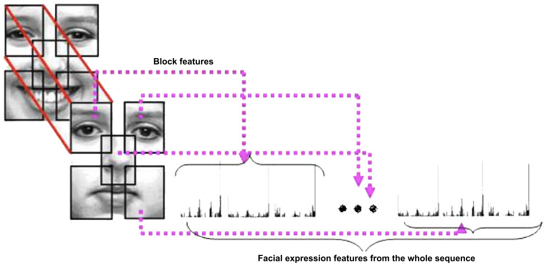

Figure 4. The concatenation of features from different blocks [17].

the sign and magnitude components:

$$d_p = s_p * m_p, \qquad \begin{array}{l} s_p = \mathrm{sgn}(d_p) \\ m_p = |d_p| \end{array} \tag{3}$$

The three operators proposed to code the features S, M and C are

$$CLBPS_{P,R} = \sum_{p=0}^{P-1} s(g_p - g_c)2^p, s(x) = \begin{cases} 1, x \geq 0 \\ 0, x < 0 \end{cases} \tag{4}$$

$$CLBPM_{P,R} = \sum_{p=0}^{P-1} t(m_p, c)2^p, t(x, c) = \begin{cases} 1, x \geq c \\ 0, x < c \end{cases} \tag{5}$$

$$CLBPC_{P,R} = t(g_c, c_l) \tag{6}$$

where R is the radius of the neighbourhood, c is a threshold set to the mean value of m_p for the whole image and c_l is a threshold set to the mean grey level of the whole image. CLBPS is equivalent to the original LBP. Figure 2 shows an example computation of the sign (S) and magnitude (M) of the local difference.

We extend the purely spatial CLBP to a dynamic texture descriptor which we call CLBP from Three Orthogonal Planes (CLBP-TOP). To the best knowledge of the authors, this is the first time that CLBP has been used for facial expression recognition and the first time CLBP has been extended into an SLTD.

We concatenate the CLBP histograms to

$$CLBPH = [CLBPS, CLBPM, CLBPC]. \tag{7}$$

$CLBPH$ is computed on three orthogonal planes XY, XT and YT and the results are concatenated as shown in Figure 3. This results in $3 \cdot 2 \cdot (2^P + 1)$ bins, where P is the number of local neighbouring points around the central pixel in a frame. We denote the resulting SLTD as $CLBPTOP_{P_{XY}, P_{XT}, P_{YT}, R_{XY}, R_{XT}, R_{YT}}$ where $P_{XY}, P_{XT}, P_{YT}, R_{XY}, R_{XT}, R_{YT}$ are the number of neighbouring points in XY, XT and YT planes and the radii in X, Y and T planes, respectively. To encode spatial information in the descriptors we divide the video into several video volumes, compute CLBP-TOP for these separately and concatenate the results as shown in Figure 4.

Given an $X \times Y \times T$ dynamic texture (DT) with $x_c \in \{0, \ldots, X-1\}, y_c \in \{0, \ldots, Y-1\}, t_c \in \{0, \ldots, T-1\}$ we can define a histogram of the DT by

$$H_{i,j} = \sum_{x,y,t} I(f_j(x, y, t) = i), i = 0, \ldots, Z, j = 0, 1, 2 \tag{8}$$

where $f_j(x, y, t)$ presents the CLBP code of the central pixel (x, y, t) in the jth plane computed by $CLBPH$ in Equation 7, and $Z = 1$ for $CLBPC$ and $Z = 2^P - 1$ for $CLBPS$ and $CLBPM$. Only the central part is considered because a sufficiently large neighbourhood cannot be used at the borders of the volume.

The histogram is normalised to a consistent description for different spatial and temporal sizes:

$$N_{i,j} = \frac{H_{i,j}}{\sum_{k=0}^{Z} H_{k,j}} \tag{9}$$

This yields CLBP-XY, CLBP-XT and CLBP-YT histograms that are concatenated to build the final CLBP-TOP feature vector shown in Figure 3.

3.4. Temporal Interpolation Model

In this subsection we briefly discuss how we use graph embedding to temporally interpolate facial expression im-

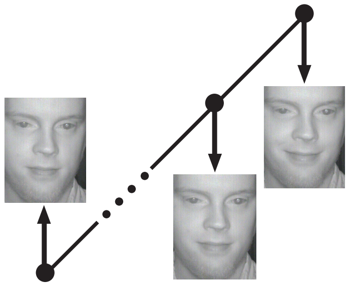

Figure 5. The graph representation of a near-infrared facial expression image sequence.

age sequences. Interpolation enables us to achieve more statistically stable feature extraction results by increasing the number of frames we use for extraction. Zhou *et al.* [18] previously proposed a similar method for synthesising a talking mouth, and Pfister *et al.* [10] applied it to recognising spontaneous facial micro-expressions. In this paper we show that this method can also yield improved results for differentiating spontaneous from posed facial macro-expressions.

Our temporal interpolation model (TIM) views a video of a facial expression as a set of images sampled along a curve and creates a continuous function in a low-dimensional manifold by representing the facial expression video as a path graph P_n with n vertices. Vertices correspond to video frames and edges to adjacency matrix $\boldsymbol{W}\in\{0,1\}^{n\times n}$. A sample graph is shown in Figure 5. To embed the manifold in the graph we map P_n to a line that minimises the distance between connected vertices. We minimise

$$\sum_{i,j}(y_i - y_j)^2\, W_{ij}, \quad i,j = 1,2,\ldots,n \qquad (10)$$

to obtain mapping $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)^{\mathrm{T}}$, which is equivalent to calculating the eigenvectors of the Laplacian graph of P_n. We compute 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 functions

$$f_k^n(t) = \sin\left(\pi k t + \pi(n-k)/(2n)\right),\ t \in [1/n, 1] \qquad (11)$$

sampled at $t = 1/n, 2/n, \ldots, n/n$. We use the resulting curve $\mathcal{F}^n(t) = [f_1^n(t) \ldots f_{n-1}^n(t)]$ to temporally interpolate images at arbitrary positions within a facial 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 a linear extension of graph embedding to learn a transformation vector $\boldsymbol{w}$ that minimises

$$\sum_{i,j}\left(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i - \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_j\right)^2 W_{ij}, \quad i,j = 1,2,\ldots,n \qquad (12)$$

where $\boldsymbol{x}_i = \boldsymbol{\xi}_i - \bar{\boldsymbol{\xi}}$ is a mean-removed vector for vectorised image $\boldsymbol{\xi}_i$. Zhou *et al.* showed that we can interpolate a new image $\boldsymbol{\xi}$ by

$$\boldsymbol{\xi} = \boldsymbol{UM}\mathcal{F}^n(t) + \bar{\boldsymbol{\xi}} \qquad (13)$$

where $\boldsymbol{M}$ is a square matrix and $\boldsymbol{X} = \boldsymbol{U\Sigma V}^{\mathrm{T}}$ is the solution to the resulting eigenvalue problem. The validity of this interpolation depends on assuming that all frames of the input video $\boldsymbol{\xi}_i$ are linearly independent. The assumption held for both the NIR and VIS subcorpora.

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

3.5. Generic Algorithm for Facial Expression Recognition

Algorithm 2 presents a generic facial expression recognition framework that combines solutions to subproblems and a solution to facial micro-expression recognition with our algorithm for SVP differentiation. Our SVP differentiation algorithm provides one of two missing links in the framework. A solution to the other link is also proposed. This generic framework addresses the facial expression recognition task by combining solutions to spontaneous vs. posed, micro vs. macro-expression recognition with classifiers for micro, posed and spontaneous expressions. Each solution in the cascade can be independently optimised or substituted. In this way the complex facial expression recognition task is subdivided into several simpler problems.

Layer1-FED requires a classifier that distinguishes image sequences with and without facial expressions. We combine the publicly available SMIC corpus [10] and the spontaneous part of the SPOS corpus proposed in this work to perform the classification. The SMIC corpus consists of 6 subjects with 77 spontaneous micro-expressions recorded with a 100fps camera running at 640x480. As negative data we use randomly selected image sequences without facial expressions from both SMIC and SPOS. This Layer1-FED system can be used for automatic segmentation of facial expressions by classifying sliding windows of frames.

If Layer1-FED detects a facial expression, Layer2-MicMac uses a publicly available facial micro-expression classifier [10] to detect micro-expressions from the image sequence. The classifier uses temporal interpolation to counter short video lengths, LBP-TOP to handle dynamic features and $\{\mathrm{SVM}, \mathrm{MKL}, \mathrm{RF}\}$ to perform classification. Faces are normalised using a Local Weighted Mean transformation on 68 facial feature points from an Active Shape Model. After extracting LBP-TOP features classifier MKL-Phase1 recognises the occurrence of a micro-expression and MKL-Phase2 classifies it into an arbitrary number of

Algorithm 2 Generic algorithm for facial expression recognition. γ is an input image sequence. Layer1-FED detects facial expressions (FED) by classifying image sequences with and without facial expressions in a sliding window. Layer2-MicMac distinguishes micro from macro expressions. Layer3-SVP distinguishes spontaneous from posed expressions. Micro classifies a micro-expression into a set of classes. Spont and Posed classify spontaneous and posed macro-expressions respectively into an application-specific set of classes.

```
Analyse-FE($\gamma$)

1. If Layer1-FED($\gamma$)= FE [10]
    (a) If Layer2-MicMac($\gamma$)= micro [10]
        i. Output Micro($\gamma$) [10]
    (b) Else
        i. If Layer3-SVP($\gamma$)= spont [this paper]
            A. Output Spont($\gamma$) [various work]
        ii. Else
            A. Output Posed($\gamma$) [7]
```

classes. The system was evaluated on two new corpora and achieved promising results.

If a micro-expression is detected, an application-dependent classifier Micro is used to output the final class. In this work we use the MKL-Phase2 classifier to classify the micro-expression as either negative or positive. If no micro-expression is detected the facial expression is a macro-expression.

Macro-expressions can be either spontaneous or posed. Layer3-SVP distinguishes spontaneous from posed facial expressions and enables them to be separately classified. If the expression is spontaneous, Spont is used to classify it. This classifier can be trained on an arbitrary dataset for an arbitrary set of classes, for example on the visible and infrared corpus collected by Wang *et al.* [15]. If the expression is posed, the Posed classifier determines the class. Posed facial expression recognition has been thoroughly studied, so we use a standard optimised approach by Huang *et al.* [7] on the Cohn-Kanade (CK) corpus [8]. Alternatively the spontaneous and posed parts of the SPOS corpus can be used.

4. Experiments and Results

We evaluate our proposed SVP differentiation system by leave-one-subject-out evaluation on the new SPOS corpus.

As the SLTD our experiments use CLBP-TOP and LBP-TOP. For MKL we use the block sizes given in Algorithm 1. Non-MKL classification results are reported with SLTD$_{8\times8\times3}$, where the image is split in 8×8 blocks in the spatial domain and 3 blocks in the temporal domain. Unless stated otherwise, CLBP-TOP results are reported using all components (S, M and C). 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 $\phi = \{\mathrm{SVM}, \mathrm{MKL}, \mathrm{RF}, \mathrm{LIN}, \mathrm{FUS}\}$ that gave the best leave-one-subject-out results. RF is the Random Forest [2] decision tree ensemble classifier, LIN is the LibLINEAR classifier [5] and FUS denotes fusion of SVM, LIN and RF through majority voting.

Channel	Method	Accuracy LBP (%)	Accuracy CLBP (%)
NIR	SVM	49.3	66.6
NIR	FUS+TIM10	55.7	73.0
NIR	LIN+TIM25	58.0	**78.2**
NIR	LIN+TIM30	62.8	76.9
VIS	SVM	65.3	70.3
VIS	FUS+TIM20	66.0	**72.0**
VIS	SVM+TIM25	66.6	70.0
VIS	SVM+TIM30	66.6	70.0
NIR+VIS	MKL+TIM25	66.8	**80.0**

Table 1. Leave-one-subject-out results on the SPOS corpus with CLBP-TOP and LBP-TOP. NIR denotes the near-infrared channel; VIS denotes the visual channel; SVM denotes support vector machines; MKL denotes Multiple Kernel Learning; TIMn denotes temporal interpolation to n frames; LIN denotes the LINEAR classifier; FUS denotes fusion of SVM, LINEAR and Random Forest through majority voting.

4.1. Experiment 1: SVP Classification

Table 1 shows the SVP differentiation results for CLBP-TOP and LBP-TOP for near-infrared, visual and combined channels using a variety of methods described in Section 3.

Interestingly the results for CLBP computed on near-infrared data are better than for visual data. This confirms the finding by Zhao *et al.* [16] that facial expression recognition most certainly can benefit from using NIR. Even with strong, stable illumination near-infrared provides a very valuable source of information for SVP classification. By computing CLBP on near-infrared data we outperform any classifier trained on visual data.

The high performance of CLBP on NIR image sequences is likely due to the illumination-invariance of NIR. Particularly the C component of CLBP is very sensitive to illumination changes, as is the M component to a lesser extent. NIR eliminates illumination variances, leaving only monotonic grey level changes.

Fusing the NIR and VIS modalities with MKL improves the results by 1.8% beyond the best performance on the NIR data. We compute separate Poly and HisInt kernels for near-infrared and visual data and use multiple kernel learning to combine them.

We used classifier fusion for the tasks that MKL did not perform well in. FUS denotes fusion of the SVM, LINEAR and Random Forest classifiers using majority voting. This yields the best performance for CLBP on the visual corpus.

As expected, SVM and fusion of other classifiers (FUS) performed best for CLBP on visual data. NIR data was best classified by LINEAR.

4.2. Experiment 2: LBP-TOP Versus CLBP-TOP

Table 1 shows a comparison of the performance of LBP-TOP and our proposed SLTD. CLBP-TOP outperforms the popular LBP-TOP feature descriptor in all our experiments. In particular, for the best result with near-infrared the difference is over 20%. For other experiments the differences are more modest but significant. The increase in performance for SVP differentiation is particularly significant. This shows that the magnitude ($CLBPM$) and centre grey level component ($CLBPC$) added in CLBP-TOP preserve important information.

The difference in accuracy of LBP-TOP and CLBP-TOP is considerable for NIR data (13.2–20.2%). For the VIS channel it is much smaller (3.4–6.0%). This shows that the magnitude and centre grey level added are particularly exploitable for near-infrared data.

The improvement from using TIM is particularly high for NIR (CLBP 11.6%, LBP 13.5%) while the improvement for VIS is more modest (CLBP 1.7%, LBP 1.3%). This suggests that particularly the NIR image sequences contain a lot of redundant data (average 235 frames) that worsens the performance when temporal interpolation and normalisation are not used.

4.3. Experiment 3: CLBP-TOP Components

Table 2 shows the results for different components of CLBP-TOP using data from the well-performing NIR channel. As explained in Section 3.3, results with only the sign component (S) are equivalent to LBP-TOP. The results from using only M or C were constantly lower than S+M.

The best results for CLBP-TOP are achieved by using all three components. This finding agrees with Guo *et al.* [6] who found all components useful for static texture classification of two texture corpora. It is interesting to note that even though the tasks are quite different (static texture recognition on visual data and dynamic SVP differentiation on near-infrared data) the results of the component division experiments follow the same pattern. In general, S+M+C yields the best results, followed closely by S+M and M+C, and more distantly by S.

However, the improvement from adding the centre grey level (C) is small. Using only S+M gets similar results. On the other hand, the magnitude M of the local difference is clearly very valuable in particular for NIR data, yielding a 17.0–18.2% improvement for the classifiers in Table 2. This is consistent with the findings of Guo *et al.* [6].

Components	Method	Accuracy (%)
S+M+C	FUS+TIM10	**73.0**
S+M	FUS+TIM10	72.7
M+C	FUS+TIM10	71.7
S+C	FUS+TIM10	56.4
S	FUS+TIM10	55.7
S+M+C	LIN+TIM25	**78.2**
S+M	LIN+TIM25	76.2
M+C	LIN+TIM25	73.0
S+C	LIN+TIM25	62.1
S	LIN+TIM25	58.0

Table 2. Leave-one-subject-out results on the SPOS corpus comparing different CLBP-TOP components. NIR data were used for this experiment. C is the centre grey level; S is the sign and M is the magnitude of the local difference d_p. TIMn denotes temporal interpolation to n frames; LIN denotes the LINEAR classifier; FUS denotes fusion of SVM, LINEAR and Random Forest through majority voting.

4.4. Experiment 4: Generic Facial Expression Recognition

As explained in Section 3.5, previous work already provides solutions to Layer2-MicMac, Micro, Spont and Posed. We refer the reader to the publications cited in Section 3.5 for the details of their performance.

Our paper is the first to provide a solution to Layer3-SVP with a realistic corpus. This solution was evaluated in Section 4.1. However, no solution previously exists for Layer1-FED. In this section we therefore evaluate our solution for this final unsolved part of the cascade.

We combine the visual facial expression part of the SPOS corpus with the SMIC micro-expression corpus [10] to create a corpus that can distinguish any facial expression from a set of frames. We use randomly selected $\frac{1}{2}$ to 5 second image sequences without facial expression as negative data. The system can be used for on-line facial expression detection by classifying sliding windows of a varying number of frames. Our experiments are equivalent to off-line runs of the sliding window classifier.

Table 3 shows the results of our solution for Layer1-FED which involves detection of both facial macro and micro-expressions. The results are reported without temporal interpolation to conform with more practical run-time performance requirements. Using an SVM on CLBP with all components we achieve 58.8% accuracy. MKL with HisInt and Poly kernels improves the result to 64.7%. Using the Random Forest classifier improves on the result

Method	Accuracy (%)
CLBP+SVM	58.8
CLBP+MKL	64.7
CLBP+RF	68.6

Table 3. Leave-one-subject-out results for Layer1-FED with visual data. SVM denotes support vector machines; MKL denotes multiple kernel learning; RF denotes the Random Forest decision tree classifier.

by 9.8% from SVM to 68.6%. Bearing in mind that recognising micro-expressions is very difficult without first temporally interpolating the image sequence, this is a promising result that sets a good baseline for future work on this topic.

The run-time performance of Layer1-FED and Layer3-SVP is chiefly limited by the performance of the temporal interpolation and feature extraction phases. For a 25 frame facial expression sequence, the average classification delay of a MATLAB implementation of TIM10+MKL over 100 runs on a 2.66 GHz PC with 4 GB RAM is 1.1 seconds. 40% of this time is spent computing the TIM; 35% computing CLBP-TOP and 25% on the other steps in Algorithm 1. Further speed improvements are possible by parallelising the implementation and rewriting it in C++. The performance overhead added by layering is minimal for classification since TIM $\boldsymbol{\xi}$, feature set μ and kernel $\boldsymbol{K}$ can be shared between layers when their methods are similar.

5. Conclusions

We have shown the first method to successfully differentiate spontaneous from posed facial expressions and described a generic facial expression recognition cascade. Our method uses graph embedding to temporally interpolate image sequences and inputs the resulting frames through a new SLTD into a set of classifiers. We have illustrated that our new spatiotemporal local texture descriptor CLBP-TOP outperforms other descriptors and that SVP differentiation benefits from both temporal interpolation and near-infrared images. Our system is the first to differentiate spontaneous from posed facial expressions with a realistic corpus and achieves promising results.

Future work includes expanding the SPOS corpus to more participants, continuing the evaluation of the generic facial expression recognition framework, and investigating alternative temporal interpolation methods. We hope to encourage further work in this area by publishing the SPOS corpus for public use.[1]

Acknowledgements. We gratefully acknowledge the support of the Finnish Funding Agency for Technology and Innovation (Tekes), Academy of Finland, Infotech Oulu, TES and the Ella and Georg Ehrnrooth Foundation.

References

[1] M. Bartlett, G. Littlewort, E. Vural, K. Lee, M. Cetin, A. Ercil, and J. Movellan. Data Mining Spontaneous Facial Behavior With Automatic Expression Coding. *Verbal and Nonverbal Features of Human-Human and Human-Machine Interaction*, pages 1–20, 2008. 2

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

[3] J. Cohn and K. Schmidt. The Timing of Facial Motion in Posed and Spontaneous Smiles. *J. Wavelets Multiresolution and Information Processing*, 2:121–132, 2005. 2

[4] Ekman, P. and O'Sullivan, M. Who Can Catch a Liar. *American Psychologist*, 46(9):913–920, 1991. 1

[5] R. Fan, K. Chang, C. Hsieh, X. Wang, and C. Lin. LIBLINEAR: A Library for Large Linear Classification. *J. Machine Learning Research*, 9:1871–1874, 2008. 6

[6] Z. Guo, L. Zhang, and D. Zhang. A Completed Modeling of Local Binary Pattern Operator for Texture Classification. *TIP*, 19(6):1657–1663, 2010. 3, 7

[7] X. Huang, G. Zhao, M. Pietikäinen, and W. Zheng. Dynamic Facial Expression Recognition Using Boosted Component-based Spatiotemporal Features and Multi-classifier Fusion. In *ACIVS*, 2010. 6

[8] T. Kanade, Y. Tian, and J. Cohn. Comprehensive Database for Facial Expression Analysis. In *FG*, 2000. 6

[9] 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

[10] T. Pfister, X. Li, G. Zhao, and M. Pietikäinen. Recognising Spontaneous Facial Micro-expressions. In *ICCV*, 2011. 1, 3, 5, 6, 7

[11] C. Shan, S. Gong, and P. McOwan. Dynamic Facial Expression Recognition Using a Bayesian Temporal Manifold Model. In *BMVC*, 2006. 2

[12] M. Valstar, H. Gunes, and M. Pantic. How to Distinguish Posed from Spontaneous Smiles Using Geometric Features. In *ICMI*, 2007. 2

[13] M. Valstar, M. Pantic, Z. Ambadar, and J. Cohn. Spontaneous vs. Posed Facial Behavior: Automatic Analysis of Brow Actions. In *ICMI*, 2006. 2

[14] M. Varma and D. Ray. Learning the Discriminative Power-Invariance Trade-Off. In *ICCV*, 2007. 2

[15] S. Wang, Z. Liu, S. Lv, Y. Lv, G. Wu, P. Peng, F. Chen, and X. Wang. A Natural Visible and Infrared Facial Expression Database for Expression Recognition and Emotion Inference. *TMM*, 12(7):682–691, 2010. 2, 6

[16] G. Zhao, X. Huang, M. Taini, S. Z. Li, and M. Pietikäinen. Facial Expression Recognition from Near-infrared Videos. *IMAVIS*, 2011. 2, 6

[17] 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

[18] Z. Zhou, G. Zhao, and M. Pietikäinen. Towards a Practical Lipreading System. In *CVPR*, 2011. 5

[1] http://tomas.pfister.fi/spontaneous_vs_posed