Non-rigid registration using free-form deformations for recognition of facial actions and their temporal dynamics

Sander Koelstra
Queen Mary, University of London
Mile End Rd, London, E1 4NS, UK
sander.koelstra@elec.qmul.ac.uk

Maja Pantic
Imperial College London / University of Twente
London, SW7 2AZ, UK / Enschede, 7500 AE, NL
m.pantic@imperial.ac.uk

Abstract

In this paper we propose an appearance-based approach to recognition of facial Action Units (AUs) and their temporal segments in frontal-view face videos. Non-rigid registration using free-form deformations is used to determine motion in the face region of an input video. The extracted motion fields are then used to derive motion histogram descriptors. Per AU, a combination of ensemble learners and Hidden Markov Models detects the presence of the AU in question and its temporal segment in each frame of an input sequence. When tested for recognition of all 27 lower and upper face AUs, occurring alone or in combination in 264 sequences from the MMI facial expression database, an average sequence classification rate of 94.3% was achieved.

1. Introduction

Successful automatic analysis of human facial behaviour cues is an important step towards a more natural interaction between humans and computers. The current methods of interaction rely on the use of input devices such as keyboards and mice for issuing explicit commands. However, a significant part of human-to-human communication relies on the use of other channels, such as facial expressions, body gestures, etc. Enabling computers to understand these messages and adapt the interaction accordingly (e.g. in terms of the user's mood) would likely make the communication more natural, efficient, persuasive and trustworthy [16]. The most important signals through which nonverbal communication occurs are facial expressions [4][5].

Most facial expression recognition systems (FERS) aim to recognise prototypical expressions of six universal basic emotions (surprise, anger, happiness, sadness, fear and disgust), as proposed by Ekman and Friesen [4]. For a survey of such FERS, the reader is referred to [16], [19] and [24]. This categorical representation can be quite useful and intuitive, but it has some important downsides. For one, the categories form only a subset of the total range of possible affect displays and classification is therefore often unnatural. Boredom or interest, for instance, do not seem to fit well in any of the emotion categories. Also, there is no straightforward way of representing the intensity of the emotions.

A different method of categorising facial signals relies on the detection of atomic facial signals (such as frowning, smiling, blinking, etc.) and does not attempt an interpretation of these muscular activities. This interpretation is instead relegated to higher-order systems. The most widely used facial signal taxonomy developed for this goal is called the Facial Action Coding System (FACS). FACS was proposed by Ekman and Friesen in 1978 and revised (and simplified) in 2002 [5]. FACS classifies atomic facial signals into Action Units (AUs) according to the facial muscles that cause them. It defines 9 upper face AUs and 18 lower face AUs. It also defines 20 Action Descriptors for eye and head position. FACS also defines the temporal segments of neutral, onset, apex and offset of each AU display. AUs are considered to be the smallest visible facial movements and are independent of age, sex, culture, etc. The aim of this work is to detect all upper and lower face AUs and their temporal segments in each frame of an input video.

Previous FERS can also be categorised in terms of used features.Approaches that use geometric features usually detect sets of fiducial facial points or fit a face mesh. These points or shapes are then tracked throughout the video and their relative and absolute position, mutual spatial position, speed, acceleration, etc., are used for recognition. Appearance-based features concern motion and tex-

The authors would like to thank Jeffrey Cohn of the University of Pittsburgh for providing the Cohn-Kanade database. The research of Maja Pantic leading to these results has been funded in part by the EU IST Programme Project FP6-0027787 (AMIDA) and the EC's 7th Framework Programme [FP7/2007-2013] under grant agreement no 211486 (SEMAINE). This work has been done while Sander Koelstra was an MSc. student at Delft University of Technology working as an exchange student at Imperial College London.

978-1-4244-2154-1/08/$25.00 2008 IE

ture changes (deformations of the skin) such as wrinkles, bulges and furrows. Surveys of the various approaches can be found in [13] and [19].

A geometric approach that also attempts to automatically detect temporal segments of AUs is the work of Valstar & Pantic [20] [21]. They locate and track 20 facial fiducial points and extract a set of spatio-temporal features from the trajectories. Then, SVMs are combined with an HMM to classify each frame into one of the temporal segments. Using only the movement of 20 feature points makes it difficult to detect certain AUs that do not lead to a clear movement of these points, such as AU 11, 14, 17, 28. On the other hand, these AUs are typical for facial expressions of emotions such as sadness (see EMFACS [5]), and for expressions of more complex mental states including puzzlement and disagreement [6], which are of immense importance if the goal is to realise human-centred adaptive interfaces. Our appearance-based approach is capable of detecting the furrows and wrinkles associated with these AUs and is therefore better equipped to recognise them.

Earlier systems using appearance-based features have used optical flow [1], active appearance models [12], Gabor wavelets [2][3][9] and temporal templates [22]. Bartlett *et al.* [2][3] have tried different methods such as optical flow, Gabor wavelets and others. They report that using Gabor wavelets renders the best results [13]. In [19] and [25] a combination of geometric features (parametric descriptions of facial components) and appearance-based features (Gabor wavelets) was proposed and they claim that the geometric features outperform the appearance-based ones, yet using both yields the best results.

Those works that aim at recognition of AUs of the FACS system recognise only subsets of up to 20 AUs [2][3][19][22][25]. Also, none of the appearance-based approaches classifies the AUs' temporal segments (neutral, onset, apex, offset). In contrast to these past efforts in the field, in this paper we present a novel approach to appearance-based analysis of facial expressions that recognises all 27 AUs and their temporal segments defined by the FACS system. Except of geometric-feature-based methods proposed in [14] [15] [20], none of the existing systems attains automatic recognition of AUs temporal segments.To the best of our knowledge, the presented system is the first appearance-based approach that can achieve such a complete analysis of the AUs displayed in an input face video. It is also the very first effort to model changes in facial expression by using a non-rigid registration method.

2. Methodology

Fig. 1 gives an overview of the complete system. In the preprocessing phase, the face is located in the first frame of the sequence and head motion and inter-subject differences are suppressed by rigid registration. Next, non-rigid registration is used to estimate the remaining motion caused by facial expressions in each frame relative to the previous frame. For each AU, a quadtree decomposition [7] is defined based on a separate training set such as to identify interesting face regions related to that AU. In the regions determined by this decomposition, orientation histogram feature descriptors are extracted. These descriptors are used in the classification part of the system, where a combined GentleBoost classifier and a Hidden Markov Model (HMM) are used to classify each frame in terms of AUs and their temporal segments. In the remainder of this section the details of each processing phase are described.

2.1. Rigid registration

In order to locate the face in the first frame of the sequence, we assume the face is in a near-frontal position in that frame and use the fully automatic face and facial point detection algorithm proposed in [23]. This algorithm uses an adapted version of the Viola-Jones face detector to locate the face. 20 facial characteristic points are detected by using Gabor-feature-based boosted classifiers.

To suppress inter-sequence and intra-sequence variations (such as respectively facial shape differences and rigid head motion throughout the sequence), registration techniques are applied to find a displacement field T that registers each frame to a neutral, expressionless reference frame. This registration consists of two parts:

$$T = T_{inter} \circ T_{intra} \tag{1}$$

The intra-subject displacement field T_{intra} is modelled as a simple affine registration. More specifically, the facial part of each frame in the sequence is registered to the facial part of the first frame, using the squared sum of differences (SSD) of the grey level values as a distance metric, to suppress minor head motions.

The inter-subject displacement field T_{inter} is again modelled as an affine registration. A subset of 10 of the 20 facial points detected earlier in the first frame that are stable (i.e., their location is mostly unaffected by facial expressions) is registered to a predefined reference set of facial points. The first frame of the sequence is assumed to be expressionless. The displacement field T_{inter} is applied to the entire image sequence to eliminate inter-subject differences.

2.2. Non-rigid registration

After preprocessing of each video sequence, we estimate the motion field M_t due to facial expressions between consecutive frames $t-1$ and t. We use an adapted version of the technique developed by Rueckert *et al.* [18], which uses a free-form deformation (FFD) model based on b-splines as described in [11]. This method was originally used to register breast MR images, where the breast undergoes local shape changes as a result of breathing and patient motion.

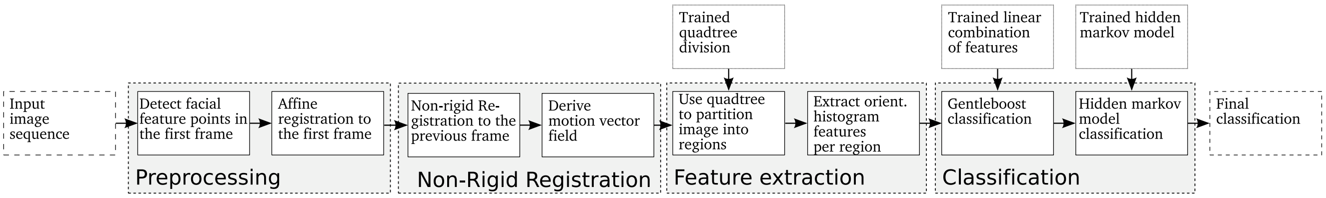

Trained quadtree division
Trained linear combination of features
Trained hidden markov model
Input image sequence
Detect facial feature points in the first frame
Affine registration to the first frame
Non-rigid Registration to the previous frame
Derive motion vector field
Use quadtree to partition image into regions
Extract orient. histogram features per region
Gentleboost classification
Hidden markov model classification
Final classification
Preprocessing
Non-Rigid Registration
Feature extraction
Classification

Figure 1: Outline of the proposed method

```
find the 20 facial feature points in the first frame of the sequence
find $T_{inter}$ (affine transformation to reference facial points) and apply it to the entire sequence
foreach frame $t$ do
    find $T_{intra}$ (affine transformation to frame 1) and apply it
    initialise the control point lattice $\Phi_0$
    foreach control point density $d$ do
        calculate the gradient vector of the cost function $C$ in terms of $\Phi_d$: $\nabla C = \frac{\delta C(\Phi_d)}{\delta \Phi_d}$
        while $||\nabla C|| > \epsilon$ do
            recalculate the control point positions $\Phi_d = \Phi_d + \mu \frac{\nabla C}{||\nabla C||}$
            recalculate the gradient vector $\nabla C$
        end
        increase the density of the control point lattice, adding new points to $\Phi_{d+1}$ from $\Phi_d$ by b-spline interpolation
    end
    use b-spline interpolation to derive $M_t$ from $\Phi$
end
```

Table 1: The registration algorithm. ϵ is a stopping criterion and μ is the step size in the recalculation of control point positions. Both are experimentally determined.

To estimate M_t, a lattice of control points is overlaid on the face box in frame t. These control points are then moved to find the optimal alignment between frame t and frame $t-1$. Next, cubic b-splines are used to interpolate the motion field in between the control points, resulting in a smooth and C^2-continuous deformation. The advantage of using b-splines is that they have local support (the interpolation is only affected by the location of control points in the direct neighbourhood), so that incremental transformations can be computed efficiently.

Let $\Omega_t = \{(x, y)|0 <= x <= X, 0 <= y <= Y\}$ represent the part of frame t containing the face (after applying T_{inter} and T_{intra} to the sequence). Let Φ be an n_x x n_y lattice of control points $\phi_{i,j}$ overlaid on Ω_t, with uniform spacing δ. In addition, for a point in Ω_t at location (x, y), let $\phi_{u,v}$ be the control point at location (x', y') that satisfies the following conditions:

$$x' < x < x' + \delta, \qquad y' < y < y' + \delta \tag{2}$$

The control points are displaced such that a cost function C describing the alignment of the images is minimised according to the algorithm displayed in Table 1. Rueckert *et al.* [18] use normalised mutual information as the image alignment criterion. However, in the simple 2D low-resolution case considered in this paper, not enough sample data is available to make a good estimate of the image probability density function from the joint histograms. Therefore, we use the sum of squared differences (SSD) as the image alignment criterion. Then, to find the new position of the point at location (x, y), we use a b-spline interpolation between it's 16 closest neighbouring control points, which gives us the displacement field M_t (depicting the motion between frame t and $t-1$) as

$$M_t(x, y) = \sum_{k=0}^{3} \sum_{l=0}^{3} B_k(a) B_l(b) \phi_{(u+k-1),(v+l-1)}, \tag{3}$$

where $a = x - x', b = y - y'$ and B_n is the nth basis function of the uniform cubic b-spline, i.e.:

$$\begin{aligned}
B_0(a) &= (-a^3 + 3a^2 - 3a + 1)/6, \\
B_1(a) &= (3a^3 + 6a^2 + 4)/6, \\
B_2(a) &= (-3a^3 + 3a^2 + 3a + 1)/6, \\
B_3(a) &= a^3/6.
\end{aligned}$$

To speed up the process, a coarse-to-fine search is used, where the density of the control point lattice is increased at each iteration (the location of new control points is determined by the b-spline interpolation). To prevent folding

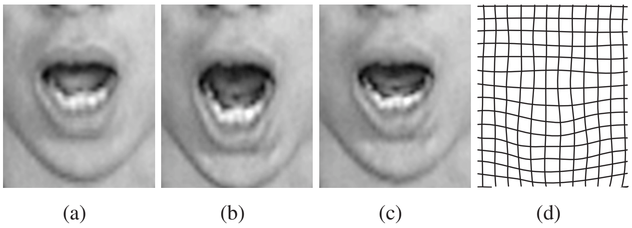


Figure 2: An illustration of the non-rigid registration process. (a) and (b): parts of frames $t-1$ and t, (c): t deformed by M_t, (d): visualization of M_t

of the control points (where one control point is moved beyond an adjacent one, leading to corruption in the image), the maximum displacement of control points cannot exceed the spacing of the lattice at that iteration. The algorithm for finding the optimal transformation is outlined in Table 1. M_t gives us a motion field depicting the facial motion between frame $t-1$ and t, from which orientation histogram features will be extracted.

Since the amount of motion between consecutive frames is usually small and may not provide enough information for AU detection, we use a temporal sliding window containing n frames wherein the motion is simply summed. A sliding window of size n for the current frame t_c gives the following transformation :

$$M_t^n = \sum_{t=t_c-n/2}^{t_c+n/2-1} M_t \tag{4}$$

In any given frame, each AU can be in one of four different temporal segments: neutral (inactive), onset, apex, or offset. Since the system only looks at motion between successive frames, there is no point in trying to detect the neutral and apex activation levels (where there is no motion). Therefore, two GentleBoost classifiers are trained per AU: one to detect the onset and another to detect the offset.

Different AUs have different onset and offset durations, therefore we consider several sizes of the sliding window n. The onset of AU 45 (blink), for instance, has an average duration of 2.4 frames (in our data set). Conversely, the offset of AU 12 (smile) lasts 15.4 frames on average. A window of 2 frames is well-suited to find the onset of AU 45, but larger windows can make it harder to detect. Thus, several sizes of the window, ranging from 2 to 20 frames, are tested. A window size of 20 frames is large enough to encompass 96.4% of all segments in our data set.

2.3. Feature Extraction

The face region in each frame of an input image sequence is divided into sub-regions and for each sub-region an orientation histogram of 8 directions, the divergence, the curl, and the motion magnitude are calculated, resulting in 11 features per sub-region. Some AUs are very much alike in appearance but differ greatly in the temporal domain. For instance, AU 43 (close and open eyes) looks exactly like AU 45 (blink) but lasts significantly longer. Therefore, we also use a number of temporal regions to extract features. To decide where to extract features, we first select the set of all sliding windows Θ in a labelled training set that show a specific AU and a specific temporal segment. Then, three projections of this set are made showing the motion magnitude, the motion over time in the x-direction, and the motion over time in the y-direction:

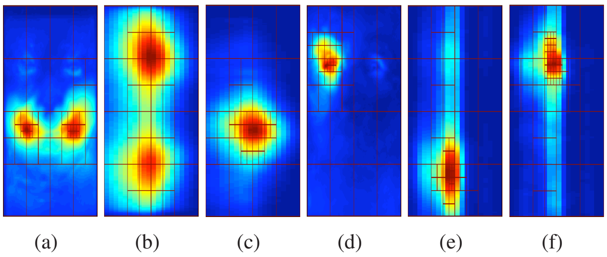


Figure 3: Quadtree decompositions: (a,b,c) Onset of AU 12(smile); (d,e,f) Onset of AU 46L(left eye wink). Shown for each AU are the three projections P_{mag} (a,d), P_{tx} (b,e), P_{ty} (c,f), as well as the resulting quadtree decompositions.

$$P_{mag}(x,y) = \sum_{\theta\in\Theta}\sum_{t\in\theta} u_t(x,y)^2 + v_t(x,y)^2, \tag{5}$$

$$P_{tx}(t,x) = \sum_{\theta\in\Theta}\sum_{y} u_t(x,y)^2, \tag{6}$$

$$P_{ty}(t,y) = \sum_{\theta\in\Theta}\sum_{x} v_t(x,y)^2, \tag{7}$$

where $u_t(x,y)$ and $v_t(x,y)$ are the components of the motion vector at location (x,y) of frame t in window θ. Since any classification algorithm can only handle a limited number of features, we aim to allocate the amount of features we can use such that unimportant areas are less covered than important ones. The projections mentioned show us exactly where there occurs much motion for a particular AU and temporal segment and where there is less. Quadtree decompositions were introduced in 1974 by Finkel & Bentley [7] and are an efficient and simple method to partition a 2D image. We use these to partition the regions so that areas showing much motion are divided in a large number of smaller sub-regions, while those showing little motion are divided into a small number of large sub-regions. Examples of motion magnitude images and resulting quadtree decompositions are shown in Fig. 3.

After generating the quadtree decompositions, the features mentioned above are extracted from each defined region in each of the projections for each frame.

2.4. Classification

To reduce the amount of features used for classification we use the GentleBoost algorithm [8], which proved successful for classification and feature selection in the domain of face and object detection. For each AU and each temporal segment (onset, offset), we train a dedicated one-vs-all GentleBoost classifier. Since our data set is rather unbalanced (over 95% of the frames depict neutral faces), to prevent all frames being classified as neutral we initialise the weights such that both the positive and negative classes carry equal weight. The GentleBoost algorithm is used to select a linear combination of features one at a time until the classification no longer improves by adding more features or a maximum of 100 features is reached.

Each onset/offset GentleBoost classifier returns a single number per frame indicating the confidence that that frame depicts the target AU in the target temporal segment. In order to combine the onset/offset GentleBoost classifiers into one AU recogniser, a continuous HMM is used. Using an HMM enables us to use the information contained in the training set about the prior probabilities of each temporal segment of an AU and its duration (represented in the HMM's transition matrix). Hence, an HMM is trained for the classification of each AU, where the outputs of the GentleBoost classifiers are used as the emissions for the HMM.

The HMM facilitates a degree of temporal filtering. For instance, given the training data, it's very unlikely to have an apex followed by a neutral phase. Also, the HMM tends to smooth out the results of the GentleBoost classifiers (for instance, short incorrect detections are usually filtered out). However, it only captures the temporal dynamics to a limited degree, for example, the HMM does not explicitly prevent onsets that last only one frame (even though minimum onset durations are much longer). Using HMMs with state duration models may help remediate this issue.

3. Experiments

The used data set consists of 264 image sequences, distributed over 15 subjects, taken from the MMI facial expression database [17] (www.mmifacedb.com). Each image sequence depicts a near-frontal view of a face showing one or more AUs, with some sequences exhibiting significant out-of-image-plane head motion. The image sequences are chosen such that each AU is present in at least 10 sequences. The image sequences on average last 3.4 seconds and were all manually coded for the presence of AUs. Ten-fold cross-validation was used, where the folds were divided such that each fold contains at least one example of each AU. Unfortunately, due to this constraint, we could not perform leave-one-out cross-validation, since some AUs are not performed by all subjects.

AU	NUM	WIN	CR	RC	PR	F_1
1	13	20	97.73	61.54	88.89	**72.73**
2	11	20	97.73	66.67	80.00	**72.73**
4	35	20	91.29	74.29	65.00	**69.33**
5	12	20	93.56	66.67	38.10	**48.48**
6	17	20	96.21	82.35	66.67	**73.68**
7	11	8	92.05	54.55	27.27	**36.36**
9	11	20	96.97	81.82	60.00	**69.23**
10	14	20	97.35	78.57	73.33	**75.86**
11	18	12	94.70	77.78	58.33	**66.67**
12	17	20	93.56	82.35	50.00	**62.22**
13	10	12	95.45	90.00	45.00	**60.00**
14	16	16	91.29	75.00	38.71	**51.06**
15	12	8	94.70	75.00	45.00	**56.25**
16	14	16	96.97	85.71	66.67	**75.00**
17	93	16	83.71	75.27	77.78	**76.50**
18	22	16	91.67	63.64	50.00	**56.00**
20	11	20	95.08	45.45	41.67	**43.48**
22	11	12	93.18	72.73	34.78	**47.06**
23	12	16	92.42	58.33	31.82	**41.18**
24	18	16	89.39	61.11	34.38	**44.00**
25	75	8	90.53	92.00	78.41	**84.66**
26	33	20	95.45	81.82	81.82	**81.82**
27	13	20	99.62	100.00	92.86	**96.30**
28	14	16	93.56	92.86	44.83	**60.47**
28B	11	16	95.45	72.73	47.06	**57.14**
28T	10	12	92.42	80.00	30.77	**44.44**
43	15	20	95.08	60.00	56.25	**58.06**
45	109	8	93.56	90.83	93.40	**92.09**
46L	11	8	99.24	90.91	90.91	**90.91**
46R	11	8	99.24	81.82	100.00	**90.00**
avg	-	-	94.31	75.73	59.66	65.12

AU=Action Unit, NUM=Number of instances
WIN=Optimal window size, CR=Classification Rate
RC=Recall Rate, PR=Precision Rate, F_1=F_1-measure

Table 2: Results for testing the system for 27 AUs (+4 partial AUs) on 264 sequences.

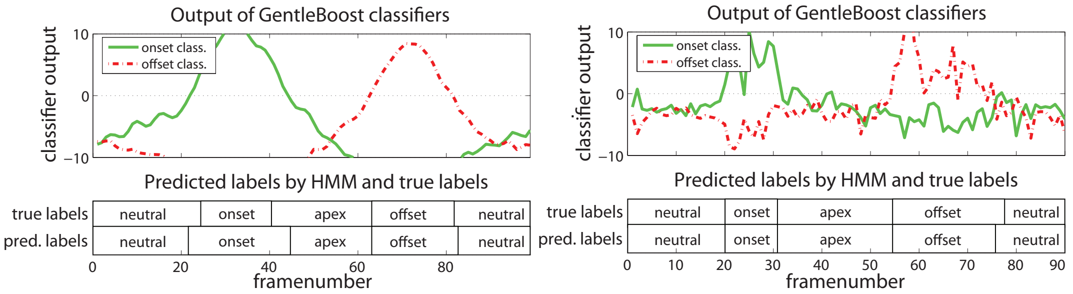
Fig. 4 shows two typical results. Fig. 4a shows the optimal situation; the GentleBoost classifiers yield very good results and the resulting labelling is almost perfect. In Fig. 4b a temporal window width of 2 frames was used, and we can see that the GentleBoost classifiers yield less smooth results. Even so, the HMM filters out the jitter effectively.

Fig. 5 shows the results for all AU classifiers for all tested window widths. AU 46 (wink) has been split up into 46L and 46R, since the appearance differs greatly depending on which eye is used to wink. Similarly, AU 28 (lip suck) is scored when both lips are sucked into the mouth, and AU 28B and AU 28T are scored when only the lower or upper lip is sucked in. The F_1-measure, which is defined as

$$F_1 = \frac{2 \cdot recall \cdot precision}{recall + precision}, \tag{8}$$

is a good indicator of the quality of the results. Overall, we clearly see that the F_1-measure improves as the temporal window increases. Exceptions include AUs with short durations, such as 7 (eye squint), 45 (blink), 46L, and 46R.

Table 2 gives a more in-depth look into the results of the best classifiers (per AU, the window width that gave the highest F_1-score is mentioned). The relatively high values of the classification rate, defined as the ratio of correct sequences to the total number of sequences, can be explained



(a) AU 27, $n = 20$

(b) AU 27, $n = 2$

Figure 4: Example classification results. The output of the GentleBoost-classifiers are shown in the top plots. The true and the estimated frame labels are shown in the bottom plots. n is the size of the used temporal window.

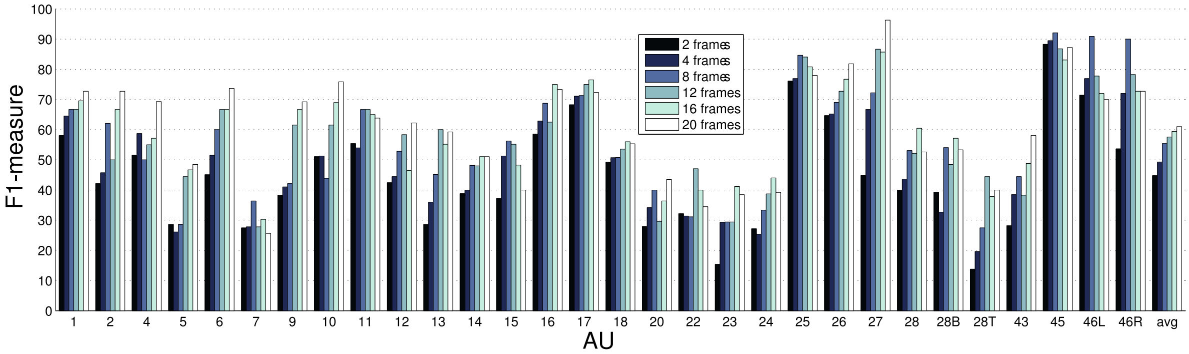


Figure 5: F_1-measure per AU for different window sizes

by the high number of true negative sequences for each AU. AUs that give the best performance are the ones that are not easily confused with other AUs and are usually less subtle, such as AU 27 (mouth stretch) and 45 (blink). The F_1-measure is reasonably high for most AUs, but there is still room for improvement. In particular, there are many false positives. Most of these occur in AUs that have a similar appearance. The worst results are achieved for AUs 5 (eye opener), 7 (eye squint) and 23 (mouth tightener). For all three AUs, the reasons for the inaccurate performance lie in the confusion of the target AU with other AUs. For instance, the onset of AU 7 is often confused with the onset of AU 45, the offset of AU 5 is very similar to the onset of AU 45 (and vice versa), and AU 23 is often confused with AU 24 (lips presser). Another cause of false positives is the sometimes poor performance of the rigid registration meant to stabilise the face throughout the sequence. Out-of-image-plane head motions for instance, are not handled very well. As a result, many classifiers will classify remaining rigid face motions as AU activations. One can see clearly from the results that AUs with shorter durations such as AU 45 benefit from a smaller window size, whereas most others perform best with the largest window size tested.

We were also interested in the timings of the temporal segment detections with respect to the timings delimited by the ground truth. This test was run using the optimal window widths as summarised in Table 2. Only sequences that were correctly classified were considered in this test. Four different temporal segment transitions can be detected, *neutral* → *onset*, *onset* → *apex*, *apex* → *offset*, and *offset* → *neutral*. Fig. 6 shows the average absolute frame deviations per AU and temporal segment transition. The overall average deviation is 2.46 frames. 44.12% of the detections are early and 38.18% are late. The most likely cause of late detection is that most AUs start and end in a very subtle manner, visible to the human eye but not sufficiently pronounced to be detected by the system. Early detections usually occur when a larger temporal window width is used, where the AU's segment in question is already visible in the later frames of the window, but it is not actually occurring at the frame under consideration (this can also be seen in Fig. 4a). In general, AUs of shorter duration also show smaller deviations. Also, the transitions that score badly are usually subtle ones. The high deviations for *apex* → *offset* in AUs 6 (cheek raiser and lid compressor) and 7 (eye squint) can be explained by considering that these transitions are first only slightly visible in the higher cheek region before becoming readily apparent in the motion of the eyelids. Since

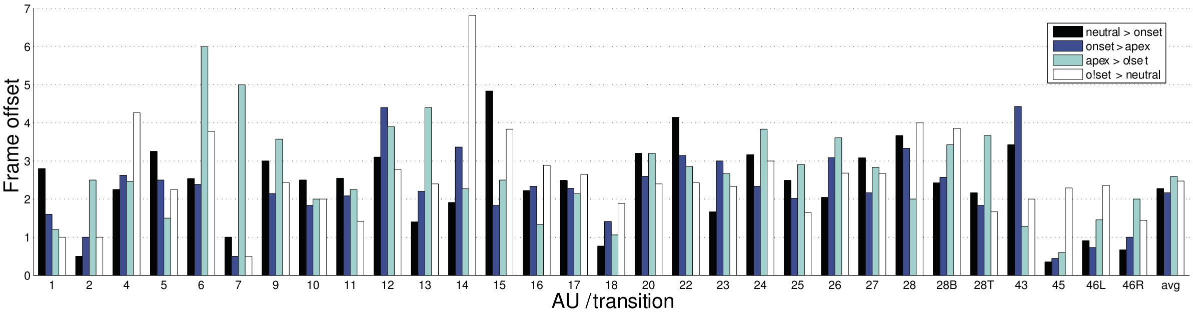


Figure 6: Average detection offsets per AU and temporal segment transition.

the eyelid motion is much clearer, our method targets that motion and misses the cheek raising in the start of the transition. Similarly, the *offset → neutral* transition in AU 14 (lip corner dimpler) has almost all of the motion in the first few frames and then continues very slowly and subtly. Our method picks up only the first few frames of this transition.

We also performed a test on the Cohn-Kanade (CK) data set [10], arguably the most often used data set in the field. This data set consists of 500 image sequences over 100 subjects. We only tested the system on those AUs for which more than ten examples existed in the data set, resulting in 20 AUs in 143 sequences. The data set does not contain offsets; the sequences are cut in the middle of apex segments. The image sequences are therefore a lot shorter than in the MMI data set; on average 0.8 seconds versus 3.4 seconds. The 10-fold cross-validation results are shown in Table 3. As a reference, the F_1-scores for the tests done on the MMI data set are repeated. The results achieved for the CK data set are in general better than those achieved for the MMI data set. One of the reasons is that out-of-image-plane head motions are rare in the CK data set. Exceptions are AUs 16 (lower lip depressor) and 26 (jaw drop). An explanation for the differences in the results lies in the differences in ground truth labelling. More specifically, in the CK database, trace activations (FACS intensity A) were also coded, especially in AU 26, whereas in the MMI data set only AUs of FACS intensity B and higher were considered. Also contributing to the differences is the higher co-occurrence of AU activations in the CK data set, making it harder to distinguish individual AUs. Another difference between the results is that for the CK data set, lower window sizes are selected than for the MMI data set. This is due to the CK sequences ending at the apex of the expression, meaning there are no offset segments. This means that no GentleBoost classifiers could be trained for the detection of offsets and the HMM classification relies solely on the onset detections. Since onsets are generally shorter than offsets, an increased window size does not benefit the classification as much.

A cross-database test was also performed, training on the MMI data set and testing on the CK data set. The results are shown in Table 4. Due to space constraints, only average results are shown. The tests were run on those AUs available in both data sets using a temporal window size of 20 frames. The average result is slightly lower than the result for training and testing on the MMI data set, but this is to be expected given the different coding styles and other differences between the data sets.

AU	NUM	WIN	CR	RC	PR	F_1	F_1^M
1	61	2	88.81	86.89	86.89	**86.89**	72.73
2	39	4	94.41	92.31	87.80	**90.00**	72.73
4	57	20	74.83	85.96	63.64	**73.13**	69.33
5	29	2	92.31	75.86	84.62	**80.00**	48.48
6	19	16	94.41	84.21	76.19	**80.00**	73.68
7	25	16	71.33	72.00	34.62	**46.75**	36.36
9	19	8	93.01	89.47	68.00	**77.27**	69.23
10	15	16	89.51	46.67	50.00	**48.28**	75.86
11	12	4	88.81	50.00	37.50	**42.86**	66.67
12	20	8	95.10	90.00	78.26	**83.72**	62.22
14	10	8	93.01	33.33	42.86	**37.50**	51.06
15	19	8	92.31	68.42	72.22	**70.27**	56.25
16	11	2	89.51	27.27	30.00	**28.57**	75.00
17	51	4	83.92	72.55	80.43	**76.29**	76.50
20	34	20	90.91	73.53	86.21	**79.37**	43.48
24	17	4	90.21	70.59	57.14	**63.16**	44.00
25	82	2	95.10	92.68	98.70	**95.60**	84.66
26	20	16	75.52	30.00	22.22	**25.53**	81.82
27	22	8	95.80	95.45	80.77	**87.50**	96.30
45	27	2	92.31	81.48	78.57	**80.00**	92.09
avg	-	-	89.06	70.93	65.83	67.63	67.42

AU=Action Unit, NUM=Number of instances, WIN=Optimal window size
CR=Classification Rate, RC=Recall Rate, PR=Precision Rate
F_1=F_1-measure, F_1^M=F_1-measure on MMI data set

Table 3: Results for testing the system for 20 AUs on 143 sequences of the Cohn-Kanade data set

	MMI→CK	MMI → MMI	CK → CK
Classification Rate	**85.75%**	91.57%	88.53%
Recall Rate	**51.30%**	72.48%	68.51%
Precision Rate	**68.89%**	47.99%	69.48%
F_1-measure	**53.04%**	55.88%	68.19%

MMI→CK: trained on MMI data set, tested on CK data set
MMI→MMI: trained on MMI data set, tested on MMI data set
CK→CK: trained on CK data set, tested on CK data set
All three are average results of 20 AUs with window size n = 20 frames

Table 4: Results for cross database testing

4. Conclusion and Future Work

In this paper we have used non-rigid registration using free-form deformations to model facial motion in frontal face image sequences. From this motion, motion orientation histograms were extracted as feature descriptors to train a classification system for the automatic frame-by-frame recognition of AUs and their temporal dynamics using a combination of ensemble learning and HMMs. To the best of our knowledge, this is the first appearance-based facial expression recognition system that can detect all 27 AUs and their temporal segments. On average, the system achieved a 76% recall and 60% precision rate when tested on the MMI facial expression database. For each correctly detected temporal segment transition, the mean of the offset between the actual and the predicted time of its occurrence is 2.46 frames. For the Cohn-Kanade database, the system achieved on average a 71% recall and 66% precision rate. The proposed system still suffers from the detection of many false positives, mainly due to confusion between AUs that are very similar in appearance. These AUs, though very similar in appearance, differ in the temporal domain. In future work we will look at employing HMMs with explicit state duration models.

References

[1] K. Anderson and P. McOwan. A real-time automated system for recognition of human facial expressions. *IEEE Trans. Systems, Man, and Cybernetics*, 36(1):96–105, 2006.

[2] M. Bartlett, G. Littlewort-Ford, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan. Recognizing facial expression: machine learning and application to spontaneous behavior. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, pages 568–573, 2005.

[3] M. Bartlett, G. Littlewort-Ford, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan. Fully automatic facial action recognition in spontaneous behavior. In *Proc. IEEE Conf. Face and Gesture Recognition*, pages 223–230, 2006.

[4] P. Ekman. Facial expression and emotion. *American Psychologist*, 48:384–392, 1993.

[5] P. Ekman, W. Friesen, and J. Hager. *The Facial Action Coding System: A Technique for the Measurement of Facial Movement*. A Human Face, Salt Lake City, UT, 2002.

[6] P. Ekman and E. Rosenberg. *What the Face Reveals*. Oxford University Press, Oxford, 2005.

[7] R. Finkel and J. Bentley. Quad trees a data structure for retrieval on composite keys. *Acta Informatica*, 4(1):1–9, 1974.

[8] J. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: a statistical view of boosting. *The Annals of Statistics*, 28(2):337–407, 2000.

[9] G. Guo and C. Dyer. Learning from examples in the small sample case - face expression recognition. *IEEE Trans. Systems, Man, and Cybernetics*, 35(3):477–488, 2005.

[10] T. Kanade, J. Cohn, and Y. Tian. Comprehensive database for facial expression analysis. In *Proc. Conf. Face and Gesture Recognition*, pages 46–53, 2000.

[11] S. Lee, G. Wohlberg, and S. Shin. Scattered data interpolation with multilevel b-splines. *IEEE trans. visualization and computer graphics*, 3(3):228–244, July 1997.

[12] S. Lucey, A. Ashraf, and J. Cohn. Investigating spontaneous facial action recognition through aam representations of the face. In K. Delac and M. Grgic, editors, *Face Recognition*, pages 275–286. I-Tech Education and Publishing, Vienna, Austria, 2007.

[13] M. Pantic and M. Bartlett. Machine analysis of facial expressions. In K. Delac and M. Grgic, editors, *Face Recognition*, pages 377–416. I-Tech Education and Publishing, Vienna, Austria, 2007.

[14] M. Pantic and I. Patras. Detecting Facial Actions and their Temporal Segments in Nearly Frontal-View Face Image Sequences. In *Proc. IEEE Intl Conf. Systems, Man and Cybernetics*, volume 4, pages 3358–3363, 2005.

[15] M. Pantic and I. Patras. Dynamics of facial expressions - recognition of facial actions and their temporal segments from face profile image sequences. *IEEE Trans. Systems, Man, and Cybernetics*, 36(2):433–449, 2006.

[16] M. Pantic and L. Rothkrantz. Toward an affect-sensitive multimodal human computer interaction. *Proc. IEEE*, 91(9):1370–1390, 2003.

[17] M. Pantic, M. Valstar, R. Rademaker, and L. Maat. Web-based database for facial expression analysis. In *Proc. IEEE Conf. Multimedia and Expo*, pages 317–321, July 2005.

[18] D. Rueckert, L. Sonoda, C. Hayes, D. Hill, M. Leach, and D. Hawkes. Nonrigid registration using free-form deformations: Application to breast mr images. *IEEE Transactions on medical imaging*, 18(8):712–721, august 1999.

[19] Y. Tian, T. Kanade, and J. Cohn. Facial expression analysis. In S. Li and A. Jain, editors, *Handbook of Face Recognition*, pages 247–276. Springer, New York, USA, 2005.

[20] M. Valstar and M. Pantic. Fully automatic facial action unit detection and temporal analysis. *Proc. IEEE Conf. Comp. Vision and Pattern Recognition*, 3:149, 2006.

[21] M. Valstar and M. Pantic. Combined Support Vector Machines and Hidden Markov Models for Modeling Facial Action Temporal Dynamics. *IEEE Int'l workshop on HCI, LNCS 4796*, pages 118–127, 2007.

[22] M. Valstar, M. Pantic, and I. Patras. Motion history for facial action detection in video. In *Proc. IEEE Int'l Conf. Systems, Man and Cybernetics*, pages 635–640, 2004.

[23] D. Vukandinovic and M. Pantic. Fully automatic facial feature point detection using gabor feature based boosted classifiers. In *Proc. IEEE Int'l Conf. Systems, Man and Cybernetics*, volume 2, pages 1692–1698, 2005.

[24] Z. Zeng, M. Pantic, G. Roisman, and T. Huang. A survey of affect recognition methods: Audio, visual and spontaneous expressions. In *Proc. ACM Int'l Conf. Multimodal Interfaces*, pages 126–133, 2007.

[25] Y. Zhang and Q. Ji. Active and dynamic information fusion for facial expression understanding from image sequence. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 27(5):699–714, 2005.