# Natural Facial Expression Recognition Using Dynamic and Static Schemes

Bogdan Raducanu[1] and Fadi Dornaika[2,3]

[1] Computer Vision Center, 08193 Bellaterra, Barcelona, Spain
*bogdan@cvc.uab.es*
[2] IKERBASQUE, Basque Foundation for Science
[3] University of the Basque Country, San Sebastian, Spain
*fadi_dornaika@ehu.es*

**Abstract.** Affective computing is at the core of a new paradigm in HCI and AI represented by human-centered computing. Within this paradigm, it is expected that machines will be enabled with perceiving capabilities, making them aware about users' affective state. The current paper addresses the problem of facial expression recognition from monocular videos sequences. We propose a dynamic facial expression recognition scheme, which is proven to be very efficient. Furthermore, it is conveniently compared with several static-based systems adopting different magnitude of facial expression. We provide evaluations of performance using Linear Discriminant Analysis (LDA), Non parametric Discriminant Analysis (NDA), and Support Vector Machines (SVM). We also provide performance evaluations using arbitrary test video sequences.

## 1 Introduction

There is a new paradigm in Human-Computer Interaction (HCI) and Artificial Intelligence focused on human-centered computing [1]. From the HCI perspective, computers will be enabled with perceptual capabilities in order to facilitate the communication protocols between people and machines. In other words, computers must use natural ways of communication people use in their everyday life: speech, hand and body gestures, facial expression. In the past, a lot of effort was dedicated to recognize facial expression in still images. For this purpose, many techniques have been applied: neural networks [2], Gabor wavelets [3] and active appearance models [4]. A very important limitation to this strategy is the fact that still images usually capture the apex of the expression, i.e., the instant at which the indicators of emotion are most marked. In their daily life, people seldom show apex of their facial expression during normal communication with their counterparts, unless for very specific cases and for very brief periods of time. More recently, attention has been shifted particularly towards modelling dynamical facial expressions [5, 6]. This is because that the differences between expressions are more powerfully modelled by dynamic transitions between different stages of an expression rather than their corresponding static key frames. This is a very relevant observation, since for most of the communication act, people rather use 'subtle' facial expressions than showing deliberately exaggerated expressions in order to convey their

message. In [7], the authors found that subtle expressions that were not identifiable in individual images suddenly became apparent when viewed in a video sequence.

Dynamical classifiers try to capture the temporal pattern in the sequence of feature vectors related to each frame such as the Hidden Markov Models (HMMs) and Dynamic Bayesian Networks [8]. In [9], parametric 2D flow models associated with the whole face as well as with the mouth, eyebrows, and eyes are first estimated. Then, mid-level predicates are inferred from these parameters. Finally, universal facial expressions are detected and recognized using the estimated predicates. Most proposed expression recognition schemes rely on the use of image raw brightness changes, which may require fixing the same imaging conditions for training and testing. The recognition of facial expressions in image sequences featuring significant head motions is a challenging problem. However, it is required by many applications such as human computer interaction and computer graphics animation [10] as well as training of social robots [11].

In this paper we propose a novel scheme for dynamic facial expression recognition that is based on the appearance-based 3D face tracker [12]. Compared to existing dynamical facial expression methods our proposed approach has several advantages. First, unlike most expression recognition systems that require a frontal view of the face, our system is view independent since the used tracker simultaneously provides the 3D head pose and the facial actions. Second, it is texture independent since the recognition scheme relies only on the estimated facial actions—invariant geometrical parameters. Third, its learning phase is simple compared to other techniques (e.g., the HMM). As a result, even when the imaging conditions change, the learned expression dynamics need not to be recomputed. It is worth noting that the proposed expression recognition schemes are only depending on the facial shape deformations (facial actions) and not on the image rawbrightness. Certainly, the shape deformations are retrieved using the rawbrightness of the sequence using the 3D face tracker based on the flexible Online Appearance Models [12].

The proposed approach for dynamic facial expression recognition has been compared afterwards against static frame-based recognition methods, showing a clear superiority in terms of recognition rates and robustness. The paper presents comparisons with several static classifiers that take into account the magnitude of facial expressions. We provide evaluations of performance using Linear Discriminant Analysis (LDA), Non parametric Discriminant Analysis (NDA), and Support Vector Machines (SVM).

The rest of the paper is organized as follows. Section 2 briefly presents the proposed 3D face and facial action tracking. Section 3 describes the proposed recognition schemes. In section 4 we report some experimental results and method comparisons. Finally, in section 5 we present our conclusions.

## 2 3D Facial Dynamics Extraction

### 2.1 A deformable 3D face model

In our work, we use the 3D face model *Candide* [13]. This 3D deformable wireframe model was first developed for the purpose of model-based image coding and computer

animation. The 3D shape of this wireframe model is directly recorded in coordinate form. It is given by the coordinates of the 3D vertices $\mathbf{P}_i, i = 1, \ldots, n$ where $n$ is the number of vertices. Thus, the shape up to a global scale can be fully described by the $3n$-vector $\mathbf{g}$; the concatenation of the 3D coordinates of all vertices $\mathbf{P}_i$. The vector $\mathbf{g}$ is written as:

$$\mathbf{g} = \mathbf{g}_s + \mathbf{A}\,\boldsymbol{\tau}_\mathbf{a} \tag{1}$$

where $\mathbf{g}_s$ is the static shape of the model, $\boldsymbol{\tau}_\mathbf{a}$ the animation control vector, and the columns of $\mathbf{A}$ are the Animation Units. The static shape is constant for a given person. In this study, we use six modes for the facial Animation Units (AUs) matrix $\mathbf{A}$. We have chosen the following AUs: lower lip depressor, lip stretcher, lip corner depressor, upper lip raiser, eyebrow lowerer, and outer eyebrow raiser. These AUs are enough to cover most common facial animations. Moreover, they are essential for conveying emotions. Thus, for every frame in the video, the state of the 3D wireframe model is given by the 3D head pose parameters (three rotations and three translations) and the internal face animation control vector $\boldsymbol{\tau}_\mathbf{a}$. This is given by the 12-dimensional vector $\mathbf{b}$:

$$\mathbf{b} = [\theta_x, \theta_y, \theta_z, t_x, t_y, t_z, \boldsymbol{\tau}_\mathbf{a}^T]^T \tag{2}$$

where:

- $\theta_x$, $\theta_y$, and $\theta_z$ represent the three angles associated with the 3D rotation between the 3D face model coordinate system and the camera coordinate system.
- $t_x$, $t_y$, and $t_z$ represent the three components of the 3D translation vector between the 3D face model coordinate system and the camera coordinate system.
- Each component of the vector $\boldsymbol{\tau}_\mathbf{a}$ represents the intensity of one facial action. This belongs to the interval $[0, 1]$ where the zero value corresponds to the neutral configuration (no deformation) and the one value corresponds to the maximum deformation. In the sequel, the word "facial action" will refer to the facial action intensity.

### 2.2 Simultaneous face and facial action tracking

In order to recover the facial expression one has to compute the facial actions encoded by the vector $\boldsymbol{\tau}_\mathbf{a}$ which encapsulates the facial deformation. Since our recognition scheme is view-independent these facial actions together with the 3D head pose should be simultaneously estimated. In other words, the objective is to compute the state vector $\mathbf{b}$ for every video frame.

For this purpose, we use the tracker based on Online Appearance Models (OAMs)—described in [12]. This appearance-based tracker aims at computing the 3D head pose and the facial actions, i.e. the vector $\mathbf{b}$, by minimizing a distance between the incoming warped frame and the current *shape-free* appearance of the face. This minimization is carried out using a gradient descent method. The statistics of the *shape-free* appearance as well as the gradient matrix are updated every frame. This scheme leads to a fast and robust tracking algorithm. We stress the fact that OAMs are more flexible than Active Appearance Models which heavily depend on the imaging conditions under which these models are built.


**Fig. 1.** Three video examples associated with the CMU database depicting surprise, anger, and joy expressions. The left frames illustrate the half apex of the expression. The right frames illustrate the apex of the expression.

## 3 Facial Expression Recognition

*Learning.* In order to learn the spatio-temporal structures of the actions associated with facial expressions, we have used a simple supervised learning scheme that consists in two stages. In the first stage, continuous videos depicting different facial expressions are tracked and the retrieved facial actions $\boldsymbol{\tau}_{\mathbf{a}}$ are represented by time series. In the second stage, the time series representation of all training videos are registered in the time domain using the Dynamic Time Warping technique. Thus, a given example (expression) is represented by a feature vector obtained by concatenating the registered $\boldsymbol{\tau}_{\mathbf{a}}$.

Video sequences have been picked up from the CMU database [14]. These sequences depict five frontal view universal expressions (surprise, sadness, joy, disgust and anger). Each expression is performed by 70 different subjects, starting from the neutral one. Altogether we select 350 video sequences composed of around 15 to 20 frames each, that is, the average duration of each sequence is about half a second. The learning phase consists of estimating the facial action parameters $\boldsymbol{\tau}_{\mathbf{a}}$ (a 6-element vector) associated with each training sequence, that is, the temporal trajectories of the action

parameters. The training video sequences have an interesting property: all performed expressions go from the neutral expression to a high magnitude expression by going through a moderate magnitude around the middle of the sequence. Therefore, using the same training set we get two kinds of trajectories: (i) an entire trajectory which models transitions from the neutral expression to a high magnitude expression, and (ii) a truncated trajectory (the second half part of a given trajectory) which models the transition from small/moderate magnitudes (half apex of the expression) to high magnitudes (apex of the expression). Figure 1 show the half apex and apex facial configurations for three expressions: surprise, anger, and joy. In the final stage of the learning all training trajectories are aligned using the Dynamic Time Warping technique by fixing a nominal duration for a facial expression. In our experiments, this nominal duration is set to 18 frames.

*Recognition.* In the recognition phase, the 3D head pose and facial actions are recovered from the video sequence using the appearance-based face and facial action tracker. We infer the facial expression associated with the current frame $t$ by considering the estimated trajectory, i.e. the sequence of vectors $\boldsymbol{\tau}_{\mathbf{a}(t)}$ within a temporal window of size 18 centered at the current frame $t$. This trajectory (feature vector) is then classified using classical classification techniques that rely on the learned examples. We have used three different classification schemes: (i) Linear Discriminant Analysis, (ii) Non-parametric Discriminant Analysis, and (iii) Support Vector Machines with a Radial Basis Function.

## 4 Experimental Results

In our experiments, we used a subset from the CMU facial expression database, containing 70 persons who are displaying 5 expressions: surprise, sadness, joy, disgust and anger. For training and testing we used the truncated trajectories, that is, the temporal sequence containing 9 frames, with the first frame representing a "subtle" facial expression (corresponding more or less with a "half apex" state, see the left column of Figure 1) and the last one corresponding to the apex state of the facial expression (see the right column of Figure 1). We decided to remove in our analysis the first few frames (from initial, "neutral" state to "half-apex") since we found them irrelevant for the purposes of the current study.

It is worth noting that the static recognition scheme will use the facial actions associated with only one single frame, that is, the dimension of the feature vector is 6. However, the dynamic classifier use the concatenation of facial actions within a temporal window, that is, the feature vector size is $6 \times n$ where $n$ is the number of frames within the temporal window. In the sequel, $n$ is set to 9.

### 4.1 Classification results using the CMU data

The results reported in this section are based on the "leave-one-out" cross-validation strategy. Several machine learning techniques have been tested: Linear Discriminant Analysis (LDA), Non-parametric Discriminant Analysis (NDA) and Support Vector

Machines (SVM). For LDA and NDA, the classification was based on the K Nearest Neighbor rule (KNN). We considered the following cases: K=1, 3 and 5.

In order to assess the benefit of using temporal information, we performed also the "static" facial expression recognition. Three static classifier schemes have been adopted. In the first scheme, training and test data are associated to the apex frames. In the second scheme, training and test data are associated to the half-apex frames. In the third schemes, we considered all the training frames in the 9-frame sequence belonging to the same facial expression, but with different magnitudes. However, during testing every frame is recognized individually and the recognition rate concerns the recognition of individual frames.

The whole results (dynamic and static) for LDA and NDA are reported in tables 1 and 2, respectively. The SVM results for the dynamic classifier are reported in table 3. The kernel was a radial basis function. Thus, the SVM used has two parameters to tune 'C' and 'g' (gamma). The first parameter controls the number of training errors, and the second one controls the RBF aperture. In general, gamma is taken as the inverse of the feature dimension, that is, it is set to $1/dim(vector) = 1/54$ for the dynamic classifier and to $1/dim(vector) = 1/6$ for the static classifier. In this case we wanted to see how the variation of the parameters 'C' (cost) affects the recognition performance. We considered six values for 'C'.

To conclude this part of the experimental results, we could say that, in general, the dynamic recognition scheme has outperformed all static recognition schemes. Moreover, we found out that the SVM clearly outperforms LDA and NDA in classification accuracy. Moreover, by inspecting the recognition results obtained with SVM we can observe that the dynamic classifiers and the static classifiers based on the apex frames are slightly more accurate than the static classifiers (half-apex) and (all frame) (third and fourth columns of Table 3). This can be explained by the fact that these static classifiers are testing separately individual frames that may not contain high magnitude facial actions.

| Classifier type | K=1 | K=3 | K=5 |
|---|---|---|---|
| Dynamic | 94.2857% | 88.5714% | 82.8571% |
| Static (apex) | 91.4286% | 91.4286% | 88.5714% |
| Static (half-apex) | 85.7143% | 82.8571% | 80.0000% |
| Static (all frames) | 84.1270% | 91.4286% | 89.5238% |

**Table 1.** LDA - Overall classification results for the dynamic and static classifiers.

### 4.2 Cross-check validation using the CMU data

Besides the experiments described above, we performed also a cross-check validation. In the first experiment, we trained the static classifier with the frames corresponding to half-apex expression and use the apex frames for test. We refer to this case as 'minor' static classifier. In a second experiment, we trained the classifier with the apex frames

| Classifier type | K=1 | K=3 | K=5 |
|---|---|---|---|
| Dynamic | 88.5714% | 88.5714% | 85.7143% |
| Static (apex) | 85.7143% | 88.5714% | 91.4286% |
| Static (half-apex) | 82.8571% | 80.0000% | 80.0000% |
| Static (all frames) | 90.7937% | 90.1587% | 91.1111% |

**Table 2.** NDA - Overall classification results for the dynamic and static classifiers.

| C | Dynamic | Apex | Half-apex | All frames |
|---|---|---|---|---|
| 5 | 94.2857% | 97.1428% | 82.8571% | 87.9364% |
| 10 | 97.1428% | 100.0000% | 85.7142% | 88.8888% |
| 50 | 100.0000% | 94.2857% | 94.2857% | 86.6666% |
| 100 | 97.1428% | 94.2857% | 94.2857% | 86.3491% |
| 500 | 97.1428% | 94.2857% | 94.2857% | 87.3015% |
| 1000 | 97.1428% | 94.2857% | 91.4285% | 88.5714% |

**Table 3.** SVM - Overall classification results for the dynamic and static classifiers.

| Static classifier | K=1 | K=3 | K=5 |
|---|---|---|---|
| Minor | 82.8571% | 85.7143% | 85.7143% |
| Major | 57.1429% | 65.7143% | 62.8571% |

**Table 4.** LDA - Cross-check validation results for the static classifier. Minor: train with half-apex frames and test with apex. Major: train with apex frames and test with half-apex.

| Static classifier | K=1 | K=3 | K=5 |
|---|---|---|---|
| Minor | 94.2857% | 88.5714% | 85.7143% |
| Major | 65.7143% | 62.6571% | 60.0000% |

**Table 5.** NDA - Cross-check validation results for the static classifier. Minor: train with half-apex frames and test with apex. Major: train with apex frames and test with half-apex.

| C | Minor | Major |
|---|---|---|
| 5 | 80.0000% | 60.0000% |
| 10 | 85.7142% | 51.4285% |
| 50 | 85.7142% | 45.7142% |
| 100 | 80.0000% | 48.5714% |
| 500 | 82.8571% | 48.5714% |
| 1000 | 82.8571% | 48.5714% |

**Table 6.** SVM - Cross-check validation results for the static classifier. Minor: train with half-apex frames and test with apex. Major: train with apex frames and test with half-apex.

and test it using the half-apex frames ('major' static classifier). The results for LDA, NDA and SVM are presented in the tables 4, 5 and 6, respectively. By analyzing the obtained results, we could observe that the 'minor' static classifier has comparable results to the static half apex classifier. This was confirmed by the three classification methods: LDA, NDA, and SVM. This means that a learning based on data featuring half apex expressions will have very good generalization capabilities since the tests with both kinds of data (half-apex and apex expressions) have a high recognition rate. Also, one can notice that the recognition rate of the minor static classifier is higher than that of the major static classifier.

This result may have very practical implications assuming that training data contain non-apex expressions, specially for real-world applications. In human-computer interaction scenarios, for instance, we are interested in quantifying human reaction based on its natural behavior. For this reason, we have to acquire and process data online without any external intervention. In this context, it is highly unlikely to capture automatically a persons apex of the facial expression. Most of the time we are tempted to show more subtle versions of our expressions and when we indeed show apex, this is in very specific situations and for very brief periods of time.

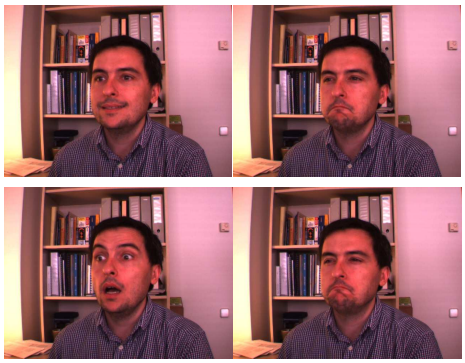

**Fig. 2.** Four snapshots from the second video sequence.

### 4.3 Dynamic vs. static recognition on non-aligned videos

In order to assess the robustness of our method, we also tested the recognition schemes on three arbitrary video sequences. The length of the shortest one is 300 frames and that of the longest is 1600 frames. Figure 2 shows four snapshots associated with the second test video sequence. These sequences depicted unseen subjects displaying a variety of different facial expressions. For training, we employed all the videos from the CMU database used in the previous sections (for which the dynamic expressions are represented by aligned 9-frame sequences). It is worth mentioning that the CMU videos and these three test videos are recorded at different frame rates. Moreover, the displayed expressions are not so similar to those depicted in the CMU data.

We compare the recognized expressions by the static and dynamic classifiers with the ground-truth displayed expressions. Since the test videos are not segmented, we perform the dynamic and static recognition only at some specific frames of the test videos. These keyframes correspond to significant facial deformations and are detected using the heuristic developed in [15]. These keyframes does not correspond to a specific frame in the time domain (onset, apex, offset of the expression). As a result of this, the task of the dynamic classifier will be very hard since the temporal window of 9 frames centered at this detected keyframe will be matched against the learned aligned trajectories. The static recognizer will not be so affected since the recognition is based on comparing the attributes of the individual detected keyframe with those of a set of learned individual frames depicting several amplitudes of the expression.

In the tables 7 and 8, we present the results for the dynamic and static classifiers, respectively. The static scheme has outperformed the dynamic scheme for these three sequences. This confirms that the dynamic classifiers need better temporal alignment. As can be seen, the recognition rates obtained with both recognition schemes are lower than those obtained with a cross validation test based on the same database. This is due to the fact that the test was performed only on two subjects displaying arbitrary facial expressions.

| Sequence name | LDA | NDA | SVM |
|---|---|---|---|
| Data_1 | 43.4783% | 34.7826% | 60.8696% |
| Data_2 | 60.0000% | 40.0000% | 60.0000% |
| Data_3 | 61.1111% | 55.5556% | 66.6667% |

**Table 7.** Recognition results for the dynamic classifier on arbitrary non aligned video sequences.

| Sequence name | LDA | NDA | SVM |
|---|---|---|---|
| Data_1 | 69.5652% | 65.2174% | 65.2174% |
| Data_2 | 80.0000% | 80.0000% | 60.0000% |
| Data_3 | 66.6667% | 66.6667% | 72.2222% |

**Table 8.** Recognition results for the static classifier on the three arbitrary non aligned video sequences. We considered only the keyframes.

## 5 Conclusions and Future Work

In this paper, we addressed the dynamic facial expression recognition in videos. We introduced a view and texture independent scheme that exploits facial action parameters estimated by an appearance-based 3D face tracker. We represented the universal expressions by time series associated with learned facial expressions. Facial expressions are recognized using several machine learning techniques. In order to show even better the

benefits of employing a dynamic classifier, we compared it with static classifiers, built on half-apex, apex, and all frames of the corresponding facial expressions.

In the future, we want to further explore the results obtained in this paper by focusing on two directions: trying to discriminate between a fake and a genuine facial expression, and solving simultaneously the alignment and recognition.

**Acknowledgements**

B. Raducanu is supported by MEC Grant TIN2006-15308-C02 and CONSOLIDER-INGENIO 2010 (CSD2007-00018), Ministerio de Educación y Ciencia, Spain.

## References

1. Lisetti, C., Schiano, D.: Automatic facial expression interpretation: Where HCI, AI and cognitive science intersect. Pragmatics and Cognition **8** (2000) 185–235
2. Tian, Y., Kanade, T., Cohn, J.: Recognizing action units for facial expression analysis. IEEE Trans. on Patt. Anal. and Machine Intell. **23** (2001) 97–115
3. Bartlett, M., Littlewort, G., Lainscsek, C., Fasel, I., Movellan, J.: Machine learning methods for fully automatic recognition of facial expressions and facial actions. In: Proc. of IEEE. Int'. Conf. on SMC. Volume I., The Hague, The Netherlands (2004) 592–597
4. Sung, J., Lee, S., Kim, D.: A real-time facial expression recognition using the staam. In: Proc. of Int'l. Conf. on Pattern Recognition. Volume I., Hong Kong, PR China (2006) 275–278
5. Shan, C., Gong, S., McOwan, P.: Dynamic facial expression recognition using a bayesian temporal manifold model. In: Proc. of British Machine Vision Conference. Volume I., Edinburgh, UK (2006) 297–306
6. Yeasin, M., Bullot, B., Sharma, R.: Recognition of facial expressions and measurement of levels of interest from video. IEEE Trans. on Multimedia **8** (2006) 500–508
7. Ambadar, Z., Schooler, J., Cohn, J.: Deciphering the enigmatic face: the importance of facial dynamics to interpreting subtle facial expressions. Psychological Science **16** (2005) 403–410
8. Zhang, Y., Ji, Q.: Active and dynamic information fusion for facial expression understanding from image sequences. **27** (2005) 699–714
9. Black, M., Yacoob, Y.: Recognizing facial expressions in images sequences using local parameterized models of image motion. Int'l. Journal of Comp. Vision **25** (1997) 23–48
10. Pantic, M.: Affective computing. In Pagani, M.e.a., ed.: Encyclopedia of Multimedia Technology and Networking. Volume I. Idea Group Publishing (2005) 8–14
11. Breazeal, C.: Sociable machines: Expressive social exchange between humans and robots. Ph.D. dissertation, Dept. Elect. Eng. & Comput. Sci., MIT, Cambridge, US (2000)
12. Dornaika, F., Davoine, F.: On appearance based face and facial action tracking. IEEE Trans. on Circuits and Systems for Video Technology **16** (2006) 1107–1124
13. Ahlberg, J.: Model-based coding: extraction, coding and evaluation of face model parameters. Ph.D. Thesis, Dept. of Elec. Eng., Linköping Univ., Sweden (2002)
14. Kanade, T., Cohn, J., Tian, Y.: Comprehensive database for facial expression analysis. In: Proc. IEEE Intl. Conf. on Automatic Face and Gesture Recognition, Grenoble, France (2000) 46–53
15. Dornaika, F., Raducanu, B.: Inferring facial expressions from videos: Tool and application. Signal Processing: Image Communication **22** (2007) 769–784