Dynamic vs. Static Recognition of Facial Expressions

No Author Given

No Institute Given

Abstract. In this paper, we address the dynamic recognition of basic facial expressions. We introduce a view- and texture independent schemes that exploits facial action parameters estimated by an appearance-based 3D face tracker. We represent the learned facial actions associated with different facial expressions by time series. Furthermore, we compare this dynamic scheme with a static one and show that the former performs better than the latter. We provide evaluations of performance using several classification schemes. With the proposed scheme, we developed an application for social robotics, in which an AIBO is mirroring the facial expression recognized.

1 Introduction

Facial expression plays an important role in recognition of human emotions. Psychologists postulate that facial expressions have a consistent and meaningful structure that can be backprojected in order to infer people inner affective state [7, 8]. Basic facial expressions typically recognized by psychologists are: happiness, sadness, fear, anger, disgust and surprise [9]. In the beginning, facial expression analysis was essentially a research topic for psychologists. However, recent progresses in image processing and pattern recognition have motivated significantly research works on automatic facial expression recognition [10–12]. 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 [16], Gabor wavelets [17] and active appearance models [19]. 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. 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 poses

in order to convey their message. In [3], 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 [20]. In [21], 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. In [12], a two-stage approach is used. Initially, a linear classification bank was applied and its output was fused to produce a characteristic signature for each universal facial expression. The signatures thus computed from the training data set were used to train discrete Hidden Markov Models to learn the underlying model for each facial expression. A Bayesian approach to modelling temporal transitions of facial expressions represented in a manifold has been proposed in [18]. However, the fact that the method relies heavily on the gray level of the image can be a serious limitation.

As can be seen, most proposed expression recognition schemes require a frontal view of the face. Moreover, most of them rely on the use of image raw brightness changes. The recognition of facial expressions in image sequences with significant head motion is a challenging problem. It is required by many applications such as human computer interaction and computer graphics animation [13–15] as well as training of social robots [5, 6].

In this paper we propose a novel scheme for dynamic facial expression recognition that is based on an appearance-based 3D face tracker [4]. 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. 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 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 scheme. In section 4 we report some experimental results and method comparisons. Section 5 describes a human robot interaction application that is based on the developed facial expression recognition scheme. Finally, in section 6 we present our conclusions and some guidelines for future work.

2 3D facial dynamics extraction

2.1 A deformable 3D face model

In our work, we use the 3D face model *Candide* [2]. 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:

- θ_x, θ_y, and θ_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.

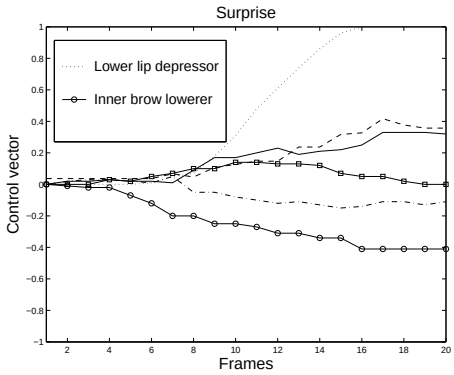
For this purpose, we use the tracker based on Online Appearance Models—described in [4]. This appearance-based tracker aims at computing the 3D head pose and the facial actions, i.e. the vector $\mathbf{b}$, by minimizing a distance between the incoming warped frame and the current *shape-free* appearance 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.

3 Facial expression recognition

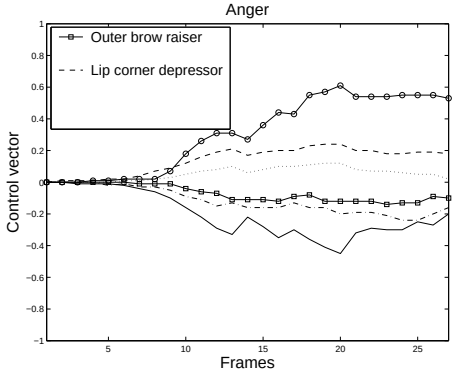
Learning. In order to learn the spatio-temporal structures of the actions associated with facial expressions, we have used a simple supervised learning scheme that consists in two stages. In the first stage, continuous videos depicting different facial expressions are tracked and the retrieved facial 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 [1]. These sequences depict five frontal view universal expressions (surprise, sadness, joy, disgust and anger). Each expression is performed by 7 different subjects, starting from the neutral one. Altogether we select 35 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. Figure 1 shows the retrieved facial action parameters associated with three sequences: surprise, anger, and joy. 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 2 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.

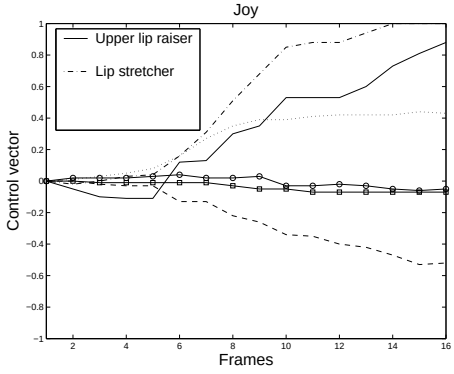
Recognition. In the recognition phase, the 3D head pose and facial actions are recovered from the video sequence using the appearance-based face and facial action tracker. We infer the facial expression associated with the current frame t by considering the estimated trajectory, i.e. the sequence of vectors $\boldsymbol{\tau}_{\mathbf{a}(t)}$ within



(a)



(b)



(c)

Fig. 1. Three examples (sequences) of learned facial action parameters as a function of time. **(a)** Surprise expression. **(b)** Anger expression. **(c)** Joy expression.

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.

It is worth noting that the static recognition scheme will use the facial actions associated with only one single frame. In this case, the training examples correspond to the apex of the expression or to its half apex.

4 Experimental results and method comparisons

In our experiments, we used a subset from the CMU facial expression database, containing 7 persons who are displaying 5 expressions: surprise, sadness, joy, disgust and anger. For dynamical facial expression recognition evaluation, 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 2) and the last one corresponding to the apex state of the facial expression (see the right column of Figure 2). 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. The results reported in this section are based on the "leave-one-out" cross-validation strategy. Several classification methods 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, by comparing frame-wise the instances corresponding to half-apex and apex states respectively. The results for LDA and NDA classifiers are reported in tables 1 and 3, respectively. The SVM results for the dynamic classifier are reported in table 5. The kernel was a radial basis function. Thus, the SVM used has two parameters to tune 'c' and 'g' (gamma). In this case we wanted to see how the variation of parameters 'c' (cost) and 'g' (gamma) affect the recognition performance. We considered 7 values for 'c' and 3 for 'gamma'. Table 8 shows the confusion matrix associated with a given "leave-one-out" for the dynamic classifier using SVM. Since we noticed that 'gamma' doesn't have a significant impact on the classification results, for the study of frame-wise case we set this parameter to its default value ($1/dim(vector) = 1/54$) and considered only different values for the 'c' parameter. The results for the static classifier based on SVM are presented in table 6. 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.

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 and test it using the half-apex frames ('major' static classifier). The results for LDA, NDA and SVM are presented in the tables 2, 4 and 7, respectively. By analyzing the obtained results, we could observe that the 'minor' static classifier has very good generalization capabilities. This was confirmed by the three classification methods: LDA, NDA, and SVM. Even when forced to cope with incomplete data it proved to give reliable hints about facial expression being shown. The explanation for this behavior, could be found by analyzing the spatio-temporal sequence. In other words, we could say that the facial actions extracted by our tracker could give good predictions of the facial expression 'trajectories'. This result may have very practical implications, 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 person's 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.

5 Application for AIBO: mirroring human facial expression

With the dynamic facial expression recognition scheme proposed in this paper, we developed an application for human-robot interaction. For this purpose, we have used an AIBO robot which has the advantage of being specifically designed for interaction with people. Among its communication mechanisms, it is endowed with the ability of displaying 'facial expressions' (according to its inner state) through a set of colored LEDs located in the frontal part of the head (see figure 3).

From its concept design, AIBO's affective states are triggered by an emotion generator engine. This occurs as a response to its internal state representation, captured through a multi-sensorial mechanism (vision, audio, and touch). For instance, it can display the happiness feeling when it detects a face or it hears a voice. But it does not possess a built-in system for vision-based automatic facial-expression recognition. For this reason, the current developed application can be considered as a natural extension of its pre-built behaviors. This application is a very simple one, in which the robot is just imitating the expression of a human subject. In other words, we wanted to see its reaction according to the emotional state displayed by a person. Some extracted frames are depicted in figure 4.

6 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 corresponding learned facial actions associated with different facial expressions by time series. In order to show even better the benefits of employing a dynamic classifier, we compared it with static classifiers, built on the half-apex and apex frames of the corresponding facial expressions. We also showed that only by using half-apex frames to train the static classifiers, we still get very reliable predictions about the real facial expression (test were done with apex frames). With the proposed scheme, we developed an application for social robotics, in which an AIBO is mirroring the facial expression recognized.

In the future, we want to further explore the results obtained in this paper by focusing on two directions: using facial actions as a hint to assess persons' level of interest during an event and trying to discriminate between a fake and a genuine facial expression.

Table 1. LDA - Overall classification results:

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%

Table 2. 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	82.8571%	85.7143%	85.7143%
Major	57.1429%	65.7143%	62.8571%

Table 3. NDA - Overall classification results

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%

Table 4. 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.

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

Table 5. SVM - Overall recognition rate for the dynamic classifier.

c (g=1/54)	g/2	g	2 g
1	91.4285%	91.4285%	91.4285%
5	91.4285%	94.2857%	97.1428%
10	97.1428%	97.1428%	97.1428%
50	97.1428%	100.0000%	97.1428%
100	100.0000%	97.1428%	97.1428%
500	97.1428%	97.1428%	97.1428%
1000	97.1428%	97.1428%	97.1428%

Table 6. SVM - Overall recognition rate for the static classifier.

Static type	c=1	c=5	c=10	c=50	c=100	c=500	c=1000
Apex	82.8571%	97.1428%	100.0000%	94.2857%	94.2857%	94.2857%	94.2857%
Half-apex	82.8571%	82.8571%	85.7142%	94.2857%	94.2857%	94.2857%	91.4285%

Table 7. 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.

Static type	c=1	c=5	c=10	c=50	c=100	c=500	c=1000
Minor	80.0000%	80.0000%	85.7142%	85.7142%	82.8571%	80.0000%	82.8571%
Major	48.5714%	60.0000%	51.4285%	45.7142%	48.5714%	48.5714%	48.5714%

Table 8. Confusion matrix for the dynamic classifier. The results correspond to the case when c=1 and g=0.0185

	Surprise	Sadness	Joy	Disgust	Anger
Surprise (7)	7	0	0	0	0
Sadness (7)	0	6	0	1	0
Joy (7)	0	0	7	0	0
Disgust (7)	0	0	0	7	0
Anger (7)	0	1	1	1	4

References

1. T. Kanade and J. Cohn and Y.L. Tian. Comprehensive database for facial expression analysis. Proc. of IEEE International Conference on Automatic Face and Gesture Recognition, 2000.
2. J. Ahlberg. Model-based coding: extraction, coding and evaluation of face model parameters. *Ph.D. Thesis*, Dept. of Elec. Eng., Linköping Univ., Sweden, September 2002
3. Z. Ambadar, J. Schooler and J. Cohn. Deciphering the enigmatic face: the importance of facial dynamics to interpreting subtle facial expressions. *Psychological Science*, 16(5):403-410, 2005.
4. F. Dornaika and F. Davoine. On appearance based face and facial action tracking. *IEEE Trans. on Circuits and Systems for Video Technology*, 16(9):1107-1124, 2006
5. C. Breazeal. Robot in society: friend or appliance? *Proc. of Wksp on Emotion-Based Agent Architectures*, pp. N/A, 1999
6. C. Breazeal. Sociable machines: Expressive social exchange between humans and robots. *Ph.D. dissertation, Dept. Elect. Eng. & Comput. Sci.*, MIT, Cambridge, 2000
7. P. Ekman. Facial expression and emotion. *American Psychologist*, 48(4):384-392, 1993
8. P. Ekman and R. Davidson. The nature of emotion: fundamental questions. *Oxford Univ. Press*, New York, 1994
9. P. Ekman. Facial expressions of emotions: an old controversy and new findings. *Philos. Trans. of the Royal Society of London*, **B** 335, pp.63-69, 1992
10. B. Fasel and J. Luettin. Automatic facial expression analysis: a survey. *Pattern Recognition*, 36(1):259-275, 2003
11. Y. Kim, S.Lee, S. Kim and G. Park. A fully automatic system recognizing human facial expressions In M. Negoita, R.J. Howlett and L.C. Jain (Eds.), *LNCS3215: Knowledge-Based Intelligent Information and Engineering Systems*, Springer-Verlag, pp.203-209, 2004
12. M. Yeasin, B. Bullot and R. Sharma. Recognition of facial expressions and measurement of levels of interest from video *IEEE Trans. on Multimedia*, 8(3):500-508, 2006
13. L. Cañamero and P. Gaussier. Emotion understanding: robots as tools and models. In J. Nadel and D. Muir (Eds.), *Emotional Development: Recent Research Advances*, Oxford Univesity Press, pp. 235-258, 2005
14. M. Pantic. Affective Computing. In M. Pagani (Ed.), *Encyclopedia of Multimedia Technology and Networking*, Idea Group Publishing, vol. I, pp. 8-14, 2005
15. R.W. Picard, E. Vyzas and J. Healy. Toward machine emotional intelligence: analysis of affective physiological state. *IEEE Trans. on Patt. Anal. and Machine Intell.*, 23(10)-1175-1191, 2001
16. Y. Tian, T. kanade and J.F. Cohn. Recognizing action units for facial expression analysis. *IEEE Trans. on Patt. Anal. and Machine Intell.*, 23:97-115, 2001
17. M. Bartlett, G. Littlewort, C. Lainscsek, I. Fasel and J. Movellan. Machine learning methods for fully automatic recognition of facial expressions and facial actions. *Proc. of IEEE Intl. Conf. on SMC*, vol. I, pp. 592-597, The Hague, The Netherlands, 2004
18. Dynamic facial expression recognition using a bayesian temporal manifold model. *Proc. of British Machine Vision Conference*, Vol. I, pp.297-306, Edinburgh, UK, 2006

19. J. Sung, S. Lee and D. Kim. A real-time facial expression recognition using the STAAM. *Proc. of Intl. Conf. on Pattern Recognition*, Vol. I, pp. 275-278, Hong Kong, PR China, 2006
20. Y. Zhang and Q. Ji. Active and dynamic information fusion for facial expression understanding from image sequences. *IEEE Trans. on Patt. Anal. and Machine Intell.*, 27(5):699-714, 2005
21. M.J. Black and Y. Yacoob. Recognizing facial expressions in image sequences using local parameterized models of image motion. *Intl. Journal of Comp. Vision*, 25(1):23-48, 1997

Fig. 2. 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.

Fig. 3. AIBO showing facial expressions.

Fig. 4. Left column: Some detected keyframes associated with the 1600-frame original video. **Middle column:** The recognized expression. **Right column:** The corresponding robot's response.