Dynamic facial expression recognition using a behavioural model

Thomas Robin
Michel Bierlaire
Javier Cruz

STRC 2009 **September 2009**

TRANSP-OR

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Dynamic facial expression recognition using a behavioural model

Thomas Robin
Transp-or
EPFL
1015 Lausanne
phone: +41 21 693 24 35
fax: +41 21 693 80 60
thomas.robin@epfl.ch

Michel Bierlaire
Transp-or
EPFL
1015 Lausanne
phone: +41 21 693 25 37
fax: +41 21 693 80 60
michel.bierlaire@address

Javier Cruz
Transp-or
EPFL
1015 Lausanne
phone: +41 21 693 24 35
fax: +41 21 693 80 60
javier.cruz@address

September 2009

Abstract

A recent interest appears in transportation for users emotion recognition. This permits to adapt car behaviors to drivers mood for safety reasons, or improve public transportation offers. Human emotions are complex and defined by several elements, such as voices intonations or facial expressions. We propose a new dynamic facial expression recognition framework based on Discrete Choice Models (DCM). The aim of the work is to model the choice of a person who is exposed to a video sequence representing a facial expression, and has to label it. The approach originality lies on the absence of ground truth and the explicit modelling of causal effects between facial features and face expression. The model is composed of two parts: the first one captures the dynamic facial expression evaluation across the frames in the sequence, and the second one concerns the frames weighting in order to determine at which moment the person decides the facial expression when looking at the video sequence. A computer vision tool, called Active Appearance Model (AAM) is used to extract facial information in videos. Concerning the dynamic expression evaluation, we assume that the person's perception evolves at regular time intervals (1 second is chosen). For each time interval a utility function is associated with each possible label (happiness, surprise, neutral, fear, anger, disgust, sadness, other, not known) in order to capture the decision maker's instantaneous perception. It contains some measures about the face in the associated frames according to the Facial Action Coding System (FACS), as well as facial texture attributes (different levels of grey on the face). For the frames weighting, a utility function is associated to each frame and contains information about the frame dynamic, such as derivatives of feature characterising the face. Finally both parts are linked with the observed choice in the construction of the likelihood function. The model is then estimated using videos from the Facial Expressions and Emotions Database (FEED). Expressions labels on the videos have been obtained using an internet survey available at *http://transp-or2.epfl.ch/videosurvey/*.

Keywords

facial expression recognition, behavioural model, dynamic model, discrete choice model, random utility model, estimation, prediction

1 INTRODUCTION

The measuring of users emotions in transportation systems has become a very important research topic. This gives information about users satisfaction in public transportation systems. In the car context, it permits to adapt vehicles behaviors to drivers mood for both, well-being and safety reasons. Emotions are defined as psychological and physiological states of users. Some non-intrusive measuring techniques have to be proposed to quantify emotions. In that context, facial expression recognition appears to be fundamental. Several applications of emotions recognition can be cited. Reimer *et al.* (2009) develop the concept of an "Aware" vehicle in order to improve the mobility, performance and safety of older drivers. Information about drivers general states, such as respiration, facial expression or concentration, are crucial to correctly apprehend the immediate drivers capabilities and adapt the vehicle behavior to it. Moreover some cars constructors are currently working on the drivers mood recognition in order to warn drivers from possible dangers generated by other users. The aim is to prevent road rages. For instance, mood recognition is only based on drivers voices. Facial expression recognition can be obviously used as a complementary information source to determine drivers moods. For routine travels, Abou-Zeid (2009) conduct experiments to measure the travel well-being for both, public transportation and car modes. Collected data were employed to estimate behavioral mode choice models. Well-being measures are used as utility indicators, in addition of standard choice indicators. Facial expression recognition could be coupled to such models, in order to better capture the commuters emotional states. Another obvious application is security, for example in airports or train stations. More generally, dynamic facial expression models could be used in any human-machine interfaces.

Some systems have been proposed to describe facial expressions. Ekman and Friesen (1978) have proposed the Facial Action Coding System (FACS), they associate sets of muscles tenseness or relaxations, called Action Units (AU) to each basic expression. A FACS expert can easily recognize AUs activated on a face, and then deduct precisely the facial expressions mixture. This has became the leading system to characterize facial expressions.

Dynamic facial expression recognition is a well known topic in computer vision. Many researches have been conducted in the field. For example, Cohen *et al.* (2003) have developed an expression classifier based on a Bayesian network. They also propose a new architecture of Hidden Markov Model (HMM) for automatically segmenting and recognizing human facial expression from video sequences. Pantic and Patras (2006) present a dynamic system capable to recognize facial AUs and so expressions, based on a particle filtering method. In this context Bartlett *et al.* (2003) use a Support Vector Machine (SVM) classifier. Finally Fasel and Luettin (2003) study and compare methods and systems presented in the literature

to deal with dynamic facial expression recognition. They focus particularly on the robustness comparison in case of environmental changes.

Discrete Choice Models (DCM) have been developed in econometrics since the late 50's. They are designed to describe the behavior of people in choice situations. The set of available alternatives, called choice set, has to be finite and discrete. The alternatives are supposed to be mutually exclusive and collectively exhaustive. They have been widely used in transportation or marketing. Their estimation is based on the likelihood maximization. Contrary to classifiers, they need behavioral data to be estimated. Ben-Akiva and Lerman (1985) propose an overview of the theory. It could be adapted to the Dynamic facial expression recognition by considering that we want to model a person who has to decide the expression of a face on a video sequence. The choice set is composed of expressions, in particular we consider the seven basic expressions, as described by Keltner (2000): happiness, surprise, fear, disgust, sadness, anger, neutral. In addition, "don't know" and "other" have been introduced in the choice set, when collecting behavioral data, in order to avoid acquiring noise (see section 3).

All presented Computer Vision systems are classifiers, meaning that they are based on a ground truth. Indeed the modeller has to decide which are the facial characteristics corresponding to an expression, in order to learn the system how to recognize it. Consequently the system is highly modeller dependant. In our approach, this assumption is relaxed. The model is estimated using behavioral data. A specification is proposed, the model estimated by likelihood maximization (the modeller is not interfering in the process), and model fit checked afterwards. Moreover, in the DCM framework, causal links between facial characteristics and expressions are explicitly modelled, parameters signs and values have sense, DCMs are not "black box". In addition, the expression set is finite, and as borders between expressions are sometimes not obvious, a probabilistic distribution among expressions instead of selecting one expression for a face, is preferable. For reasons described above, the discrete choice theory appears to be well adapted. Note that M.Sorci *et al.* (2008) have used successfully DCMs for static facial expression recognition, (static meaning considering images instead of videos). Their proposed model is a simple logit model, with nine alternatives corresponding to the nine expressions cited above. Each expression utility contains measures related to its associated AUs, defined by the FACS, they use also Expression Descriptive Units (EDU), that capture interactions between AUs. Finally some outputs of the computer vision algorithm used to extract measures on faces images, called C parameters, are directly injected in utilities, in order to account for the global facial perception.

Of course dynamic facial expression recognition does not fit into the usual discrete choice

applications, so adjustments have to be done. The model basis is a DCM with latent segmentation. This kind of models has been proposed by J.L.Walker (2001). Moreover we inspire from the work of Choudhury (2007) who uses a dynamic behavioral framework to handle with car line changing models.

We first present the model framework, then the data collected and used to estimate the model, the model specification and estimation, its predictions and finally the conclusions.

2 Modeling framework

As discussed in the introduction, the model lies on a DCM with latent segmentation, which was proposed by J.L.Walker (2001). This is motivated by the dynamic aspect of the model. We hypothesise that the respondent expression perception evolves when watching the video. In addition we consider that the influence of the video frames on the respondent perception is varying depending on their dynamic characteristics. Each second of a video contains 25 frames. As a single frame is considered to be too short to influence directly the perception, a perception evolution time step is defined equal to one second. So the sequence is discretized in groups of 25 frames, each corresponding to one second of the video. The features for each group are the features of the first frame of the group. By extension in the following we call a group of frames, a frame.

The dynamic facial expression recognition model consists of a combination of two DCMs. At each time step is associated a perception state corresponding to the respondent facial expression perception at that moment. A first DCM is used to quantify this perception, the choice set is composed of the nine expressions used in the static case. The second DCM quantifies the frames influences on the respondent observed facial expression choice. The choice set is composed of the frames of the labelled video. So the choice set varies from one video to another, as the frame number is varying. Note that both models are based on latent concepts, the respondent instantaneous expression perception and the frames influences are not observed. Only the video expression choice is observed. In the following, the model details are explained.

The probability for respondent n to choose the expression i when watching the frame t of the video sequence o is written $P_n(i/t, o)$ (first DCM). Then the probability for the respondent n to make her expression choice when watching frame t of the video sequence o is $P_n(t/o)$ (second DCM). The two DCMs are linked by the probability for the respondent n to label the video o

with the expression i, called $P_n(i/o)$. The relation is shown in equation 1, T_o being the video duration in seconds. It is a classical combination of conditional probabilities.

$$P_n(i/o) = \sum_{t=1}^{T_o} P_n(i/t,o)P_n(t/o) \tag{1}$$

As shown for the static model (M.Sorci *et al.* (2008)), $P_n(i/t,o)$ is quite universal, in the sense that no clear socio-economic characteristic seems to interact with the expression perception. We expect that it is not the case for $P_n(t/o)$ which should strongly depends on the respondent n. Indeed the frame dynamic perception depends on the current respondent attention. This leads to take into account the panel data effect. ξ_n is defined as a random term specific to the respondent n. So equation 1 can be transformed as shown in equation 2.

$$P_n(i/o,\xi_n) = \sum_{t=1}^{T_o} P_n(i/t,o)P_n(t/o,\xi_n) \tag{2}$$

In order to obtain a closed form of $P_n(i/o,\xi_n)$, we need to integrate over ξ_n. By default ξ_n is supposed to be normally distributed $N(0,\sigma)$. $f(\xi)$ is the probability density distribution of ξ_n, and O_n is the number of observations associated to the respondent n. By integration, we obtain $P_n(i/o)$, the formula is expressed in equation 3).

$$\prod_{o=1}^{O_n} P_n(i/o) = \int \prod_{o=1}^{O_n} \sum_{t=1}^{T_o} P_n(i/t,o)P_n(t/o,\xi_n)f(\xi)d\xi \tag{3}$$

Theoretically $P_n(i/t,o)$ can be of any DCM type, such as MEV, or mixture of logit models. But as mentioned before, the model is similar to the static model proposed by M.Sorci *et al.* (2008). In a first time a simple logit model will be used, and the utility specification will be near from the one proposed in the static model version. In a second step, utilities will take into account the perception memory effect. Indeed we will consider that the previous frame expression perception influences the current one. Practically $V_n(i/t,o)$, the utility associated with the expression i in the frame t of the video o for individual n will be defined as follows (see equation 4). $Vgeneric_n(i/t,o)$ denotes the generic specification of $V_n(i/t,o)$. In case of no memory effect, of course $V_n(i/t,o) = Vgeneric_n(i/t,o)$. $a_{i,n}$ is the memory parameter, different assumptions can be made on it, such as being considered independent from the expression i and individual n, or specific to the expression i and independent from respondent

n.

$$V_n(i/t, o) = Vgeneric_n(i/t, o) + a_{i,n} Vgeneric_n(i/t-1, o) \tag{4}$$

Concerning $P_n(t/o, \xi_n)$, it is a mixture of logit models, due to the panel data effect term. We prefer to use a quite simple model form, such as mixture of logit models. Mixtures of MEV models are not considered since a correlation between frames is difficult to define. Moreover, the frames number vary from one video to another. The utility specification has to contain attributes which reflect the frame dynamic, such as derivatives of the attributes used in the first DCM. The idea of using a simple correlation structure is also motivated by the fact that both models are estimated jointly by likelihood maximization, (as a classical DCM). The combination of such models can imply high non linearities in the likelihood function, and the optimization algorithm has to deal with such difficulties. β is the parameters vector which has to be estimated. $c_{i,o,n}$ represents the choice indicator, i.e. $c_{i,o,n} = 1$ if respondent n chose expression i for video o, 0 otherwise. The likelihood $l(\beta)$ has the form described in equation 5.

$$l(\beta) = \prod_{n=1}^{N} \prod_{o=1}^{O_n} \prod_{i=1}^{9} P_n(i/t, o, \beta)^{c_{i,o,n}} \tag{5}$$

By mixing equation 3 and equation 5 we obtain equation 6. Moreover for numerical reasons the logarithm of the likelihood function, $L(\beta)$, is used instead of $l(\beta)$ during the estimation process, it is described in equation 7.

$$l(\beta) = \prod_{n=1}^{N} \left(\int \prod_{o=1}^{O_n} \sum_{t=1}^{T_o} \left(\prod_{i=1}^{9} P_n(i/t, o, \beta)^{c_{i,o,n}} \right) P_n(t/o, \xi_n, \beta) f(\xi) d\xi \right) \tag{6}$$

$$L(\beta) = \sum_{n=1}^{N} log\left(\int \prod_{o=1}^{O_n} \sum_{t=1}^{T_o} \left(\prod_{i=1}^{9} P_n(i/t, o, \beta)^{c_{i,o,n}} \right) P_n(t/o, \xi_n, \beta) f(\xi) d\xi \right) \tag{7}$$

We conclude this section by underlying the fact that the model specification will depend on the number of observations provided by the internet video survey (see section 4). For instant the number of collected observations is quite narrow. This little number constraints the number of alternative specific parameters in the perception model to be reduced, compared to the static

model version.

3 Face Video Sequence Databases and Features Extraction

In order to estimate such models we need data. First, face video sequences are required, two databases were retained: the Cohn-Kanade and the Facial Expressions and Emotions Database (FEED). T.Kanade (2000) collected face video of actors playing artificial expressions according to the Facial Action Coding System (FACS). The FACS have been proposed by Ekman and Friesen (1978), they characterize expressions with sets of muscles tenseness or relaxations called action units (AU). The figure 1 shows some examples of AUs.

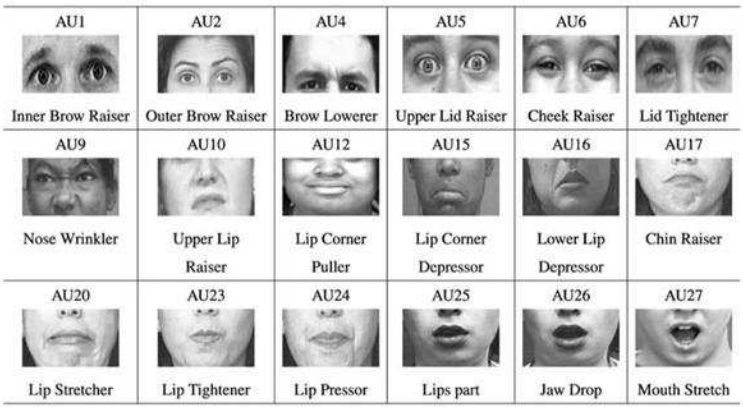

Figure 1: Examples of Action Units (AU)

A snapshot of a Cohn-Kanade video is shown in figure 2. At the video beginning, actors have neutral faces and move toward others expressions. The advantage of such database is to provide clear facial features evolutions, but they are artificial and videos are very short (around 1s). The database contains 69 sequences from 11 subjects.

The FEED database contains natural expressions. Wallhoff (2004) proceeds by filming students watching television. Different kind of videos are presented to students in order to generate a large spectrum of expressions. In the collected facial video sequence, students start with neutral faces and go to others expressions, but contrary to the Cohn-Kanade database, expressions fluctuations can appear, due to the less artificial nature of the collecting procedure. In addition, videos last between 3 and 6 seconds. So they seem particularly adapted to estimate the dynamic model. A snapshot of a FEED database video is shown in figure 3. The database contains 95 sequences from 18 subjects.

Information about the faces are extracted using an Active Appearance Model (AAM). This computer vision tool has been introduced by Cootes *et al.* (2002). It permits to extract facial distances from images as well as facial texture information (different levels of grey). This

Figure 2: Snapshot of a Cohn-Kanade database video

Figure 3: Snapshot of a FEED database video

is based on Principal Component Analysis (PCA), considering a face image as a coloured pixels array. Face video sequences are considered as succession of frames, and for each frame the facial attributes mentioned above are calculated. The algorithm permits to track a facial mask along the videos. Then measures corresponding to distances between mask points, are calculated. An example of mask mapped to a face is shown at figure 4, note that the number of mask points is constant and equal to 55. This is the first type of facial attribute. The second one is a direct output of the computed PCA, a vector describing both facial texture and shape, called vector of C parameters.

For instance evoked attributes are calculated considering each video frame separately, but as a

Figure 4: Mask tracked by AAM along a video sequence

dynamic process is modelled we need to pay attention to dynamic features, such as derivatives of static attributes calculated by finite difference. In addition in both databases cases, face video sequence start from the neutral expression, so distance between first frames attributes and the current frame ones are considered. The total number of attributes is then equal to 564. 88 are distances between mask points, 100 are C parameters (this is the static part); 188 are derivatives from previous 188 static ones; and last 188 are distances between first frames and current ones.

4 Behavioural Data

The approach's originality leads in the expression ambiguity modelling. Indeed the developed model is not a classifier, it is not designed to predict an expression for a face video, but a probability distribution among the expressions set. In order to model the human expression perception and relax the ground truth assumption, we need to collect behavioral data. An internet survey has been conducted in order to obtain expression labels on face video sequences presented in section 3. It is a good way to collect a large number of observations from an heterogeneous group of respondents. This is available at *http://transp-or2.epfl.ch/videosurvey/*. At first connection, respondents are asked to create accounts using their e-mail address and fill a socio-economic form. This is asked because in future modelling steps socio-economic characteristics could be included in the model. Once accounts are created, they have to choose how many facial video they want to label, 5, 10 or 20, videos are taken randomly from the two databases presented in the section 3. Then the expression labelling process can start, a screen snapshot is shown at figure 5. The nine expressions are displayed below the video, and respondents have to choose the one they perceived, and pass to the following video. Note that "Other" and "Don't know" labels have been added in order to avoid noise collection. Doing so, respondents are not forced to choose one of the seven basic expressions, the choice set is exhaustive. When labelling tasks are finished, respondents can use directly login and password to continue video labelling later. The data collection permits has permitted 612 video labels

from 40 respondents, it is available since August 2008.

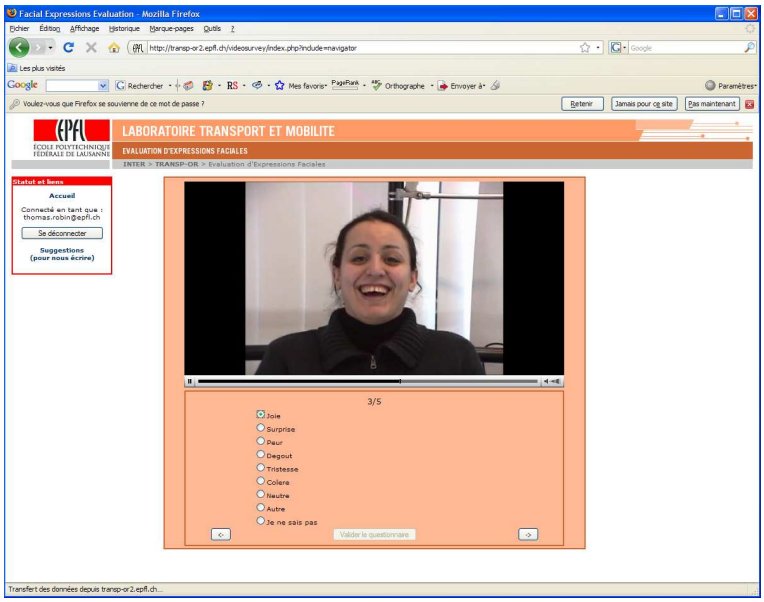

Figure 5: Snapshot of internet survey screen

The ambiguity of human expression perception is accounted for since several labels are collected for a single video. So in case of ambiguous expressions, the observed distribution among expressions will reflect the ambiguous perception of a single respondent.

5 Model Specification

This section deals with the specification of the model presented in section 2. We discuss the specification of the two combined DCMs. The first one concerns the expression perception within a frame, denoted by $P_n(i/t, o)$, the second one is related to the frame choice, denoted by $P_n(t/o, \xi_n)$, which capture when the respondent decides to label the video with the chosen expression. Both models are latent in the sense that only respondents videos labels are observed. In the following, both models specifications are described.

The expression perception model is similar to the one proposed by M.Sorci *et al.* (2008) in a static context. As mentioned in section 1, they have developed a DCM based on FACS, EDU, and C parameters. We drew our inspiration from their proposed specifications. In our case, the number of estimated parameters has to be limited because for the moment the size of the behavioral database is quite narrow. So we decide to start the model specification with the simplest final model developed by M.Sorci *et al.* (2008). It is the model with only FACS attributes. The expressions utilities are linearly specified in the parameters,

according to the FACS. Indeed utilities contain only attributes relative to AUs characterizing the expressions. This model contains already 93 parameters. In our specification, few C parameters are added in order to account for global expression perception. Attributes are constant over alternatives, since they are all relative to the face. So for one attribute, specific parameters are added in all utilities, except at least one. The neutral expression is taken as reference, due to its special status. This is expressed at the beginning of each video, and corresponds to a fully relaxed face. The " Don't know" alternative has also a special status, no attribute is related to this expression, associated utility contains only one Alternative Specific Constant (ASC). Other utilities contain also ASCs, except the neutral one being the reference, in order to absorb perception errors means. The specification of the error term stays simple for identification purpose, consequently a simple logit model is chosen. This is convenient and coherent with the static model proposed by M.Sorci *et al.* (2008). They did not find any significant correlation structure between expressions. Even if the model is of static nature, it is possible to include dynamic modelling as shown in equation 4. Some $a_{i,n}$ parameters will be estimated, considering that it is independent from respondents, i.e. $a_{i,n} = a_i$.

The frame choice model is related to another part of the dynamic process. The choice set is composed of the considered face video sequence frames. Videos contain 25 frames per second. As exposed in section 2, for psychological reasons, considering all frames has no sense, indeed we assume that the human perception is evolving at regular time steps, that are around 1 second. In addition all frames consideration leads to heavy computational effort. For both reasons, videos are sampled, retaining only the first frame of each second. Consequently the choice set is varying from one video to another. Attributes retained to explain the frame choice are relative to frame dynamics. Naturally derivatives of features implied in the expression perception model are likely to explain the frame choice, both AUs and C parameters. The parameters number has to be limited and they have to be wisely chosen. In priority, derivatives of obvious directly perceived measures are included, such as "eyes heigh", "mouth's width" or "mouth's height". The model does not contain any ASCs, because a priori preference to a frame has no sense. As frames number is varying, structural correlations between them are difficult to identify, that's why a basic logit model is selected. In addition to ease the model estimation process $P_n(t/o, \xi_n)$ is simplified, indeed $P_n(t/o, \xi_n) = P_n(t/o)$, no panel data term is included.

6 Model Estimation and Results

For the model estimation only the labels corresponding to FEED videos will be used, due to their nature, in fact they are associated to longer and more natural videos than Cohn-Kanade

ones. 294 observations are used for estimation. Labels on Cohn-Kanade videos will be reserved for model validation. The model is estimated by likelihood maximization (equation 7). Practically this is done by using codes based on the BIOGEME software developed by Bierlaire (2003), it permits to estimate more complex model than DCM but as is not dedicated to special model class. The general estimation results are presented in table 1. Parameters values, associated standard errors and t-tests against 0 are presented in table 2.

Nb of observations:	294
Nb of parameters:	44
Null log-likelihood:	-645.98
Final log-likelihood:	-358.82
$\bar{\rho}^2$:	0.38

Table 1: General estimation results

Most of parameters are significant (t-tests against zero values higher than 1.96). If it is not the case, a likelihood ratio test has been conducted to check the significant improvement of the likelihood brought by a parameter. In addition to FACS attributes, both sub-models (expression choice model and frame choice model) contain outputs of the AAM. Output features of the AAM are denoted by "$C\_$", and the corresponding number (1 to 100). Parameters 1 to 32 are relative to the expression choice model. 1 to 8 are the ASCs, the letters after "ASC_" denotes the expression utility in which it is present (H: happiness, SU: surprise, F: fear, D: disgust, SA: sadness, A: anger, O: other, DK: don't know). 9 to 32 are parameters associated with attributes. Parameters names have a meaning, to understand it, the simplest is to make an example: if we take parameter "$b\_eye\_nose\_dist\_l\_A$","$b\_$" means that the parameter is associated with an attribute, "$eye\_nose\_dist\_l$" is the associated attribute, which is the distance between nose and left eye, "A" is the utility in which it is present, here anger utility. Parameters signs and values are in line with the static model.

Parameters 33 to 40 are present in the frame choice model, and are generic across the frames. As for the expression choice model the parameter name is interpretable: for example for name of parameter 39 ("$b\_FRAME\_mouth\_h\_deriv$"), "b_FRAME" means that the parameter is present in the frame choice model, "$mouth\_h\_deriv$" means that the parameter is associated with the derivative of the feature "$mouth\_h$", which stands for the height of the mouth. Only features derivatives are retained in this model. Parameters values are not obvious to explain, but in general they seem logical. For example parameter 39, associated with the mouth height is positive, meaning that the higher the mouth will be, the more probably the frame will be chosen to put the expression label on the video. It seems logical regarding to the surprise or fear expressions. Distances between features of the current frame and first frame were tested in

Id	Parameter name	Value	Std-error	t-test 0
1	ASC_A	-7.83	5.31	-1.47
2	ASC_D	6.90	4.10	1.68
3	ASC_DK	-0.54	0.39	-1.40
4	ASC_F	-31.90	8.89	-3.59
5	ASC_H	24.23	5.31	4.56
6	ASC_O	5.31	3.14	1.69
7	ASC_SA	8.70	6.85	1.27
8	ASC_SU	-13.33	2.97	-4.48
9	$b\_broweye\_l2\_SA$	570.49	134.49	4.24
10	$b\_broweye\_l3\_SU$	70.73	19.54	3.62
11	$b\_broweye\_r2\_A\_D\_F\_SA\_SU$	-99.24	26.90	-3.69
12	$b\_eye\_angle\_below\_l\_F$	6.24	2.46	2.54
13	$b\_eye\_angle\_l\_F\_SA$	17.33	5.01	3.46
14	$b\_eye\_angle\_r\_F\_SA$	-10.17	3.08	-3.30
15	$b\_eye\_brow\_angle\_l\_SA$	-16.58	3.95	-4.20
16	$b\_eye\_mouth\_dist\_l2\_D$	-49.54	24.91	-1.99
17	$b\_eye\_mouth\_dist\_l\_H\_O\_SA$	-97.20	36.70	-2.65
18	$b\_eye\_nose\_dist\_l\_A$	248.02	36.42	6.81
19	$b\_eye\_nose\_dist\_l\_D\_F\_O\_SA$	101.16	22.25	4.55
20	$b\_eye\_nose\_dist\_r\_D\_F\_O\_SA\_A$	-131.09	19.88	-6.59
21	$b\_leye\_h\_F$	660.84	145.33	4.55
22	$b\_leye\_h\_SU$	340.57	62.41	5.46
23	$b\_mouth\_h\_A\_D\_H\_SA\_F\_SU$	79.71	25.32	3.15
24	$b\_mouth\_nose\_dist2\_A\_SA$	-283.30	56.02	-5.06
25	$b\_mouth\_nose\_dist\_H$	-324.71	52.45	-6.19
26	$b\_mouth\_w\_A\_D\_F\_H\_O$	36.40	15.42	2.36
27	$b\_C\_1\_SU$	90.35	20.63	4.38
28	$b\_C\_1\_F$	153.47	28.59	5.37
29	$b\_C\_1\_D$	115.28	19.57	5.89
30	$b\_C\_1\_A$	170.99	27.17	6.29
31	$b\_C\_2\_H$	23.23	10.47	2.22
32	$b\_C\_2\_SU$	33.94	13.28	2.56
33	$b\_FRAME\_C\_1\_deriv$	-45.46	25.41	-1.79
34	$b\_FRAME\_C\_2\_deriv$	-224.99	72.18	-3.12
35	$b\_FRAME\_C\_3\_deriv$	240.01	79.08	3.04
36	b_FRAME_C_5_deriv	-73.34	27.28	-2.69
37	b_FRAME_eye_h_deriv	-805.69	226.21	-3.56
38	b_FRAME_eye_brow_angle_deriv	43.97	14.33	3.07
39	b_FRAME_mouth_h_deriv	1309.91	399.85	3.28
40	b_FRAME_mouth_w_deriv	-184.44	56.81	-3.25
41	A_H	-0.70	0.13	-5.25
42	A_D	-0.15	0.10	-1.49
43	A_SA	-0.49	0.11	-4.28
44	A_A	-0.15	0.09	-1.58

Table 2: Estimated parameters of the model

the model, but even if it improves the model fit, it deteriorates the model prediction power, in terms of outliers percentage (see 7).

Parameters 41 to 44 capture the memory effect (described in equation 4) in the expression choice model, they are denoted by a "$A\_$", and the following letter stands for the expression utility to which it is associated. Note that they were not significant for all expressions, and they are less than one in absolute value, according more importance to the present frame than the previous one, which seems logical.

Another model (called ASC model) with only ASCs in the expressions utilities has been estimated. The frames utilities are fixed to zero, meaning that each frame has the same probability to be chosen to make the expression choice. The model does not contain any attribute, no causal effect is captured. This is a very simple model, which is used to compare the proposed specification , and show the significant improvement brought by the addition of new parameters. The main property of the ASC model is to reproduce the aggregated expressions shares of the estimation dataset, when using it for prediction. The general estimation results are shown in table 3, and ASCs values, standard errors and t-tests against zero in table 4.

Nb of observations:	294
Nb of parameters:	8
Null log-likelihood:	-645.98
Final log-likelihood:	-572.437
$\bar{\rho}^2$:	0.10

Table 3: Estimation results of the model with only ASCs

Id	Parameter name	Value	Std-error	t-test 0
1	ASC_H	1.43	0.27	5.29
2	ASC_SU	1.17	0.28	4.23
3	ASC_F	0.34	0.32	1.09
4	ASC_D	1.42	0.27	5.23
5	ASC_SA	-0.27	0.37	-0.73
6	ASC_A	0.21	0.33	0.65
7	ASC_O	-0.27	0.37	-0.73
8	ASC_DK	-0.53	0.40	-1.33

Table 4: Estimated parameters of the model with only ASCs

7 Model predictions

Even if the model fit seems to be good, the model prediction power has to be tested. Due to the high number of parameters and the little number of collected observations, the model can easily over-fit the dataset. The dataset used in this section is the same that the one used in the section 6. We proceed in two steps: the first one consists of comparing the percentage of outliers of the proposed model, and the one of the ASC model described in section 6; in a second step we study the proposed model predictions at a more disaggregated level, looking in details to the two sub-models predictions for a couple of observations.

In figure 6 and 7, the choice probability distributions are plotted for both models: proposed model and ASC model. Theoretically, for a perfect model, the choice probabilities are equal to one, the corresponding histogram should be completely displaced to the right. In reality it is of course not the case, but the prediction goodness can be measured, for example with the outliers percentage. This is defined as the percentage of observations predicted with a probability less or equal to $\frac{1}{E}$, E being the number of expressions, which is 9. In other terms, an outlier is an observation with a predicted choice probability less or equal to the hazard threshold. On both figure outliers thresholds are represented by black lines. Qualitatively the proposed model is better than the ASC one, because the choice probability distribution is widely spread on the right. The percentage of outliers for the proposed model is 16.33% against 33.33% for the ASC model. This shows the significant improvement of the explanatory variables addition in the model, in terms of prediction.

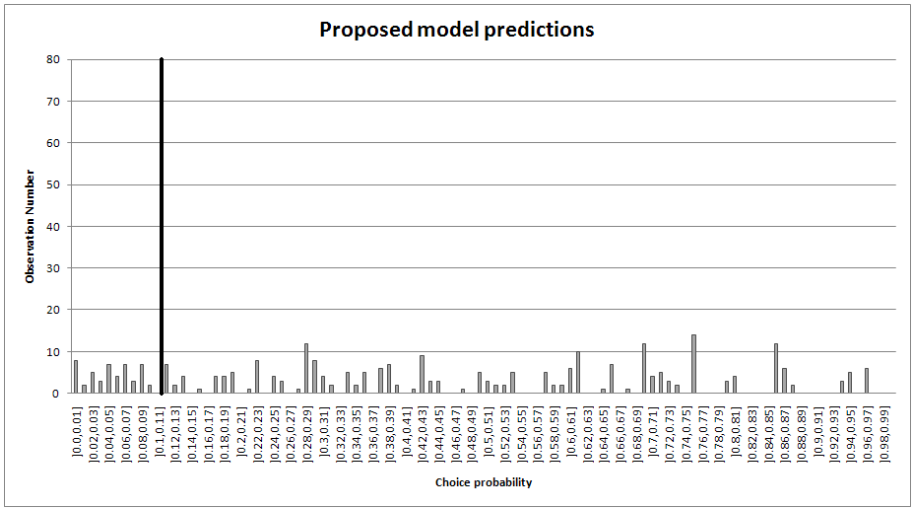

Figure 6: Distribution of the choice probabilities for the proposed model

We looked at the model prediction power over the estimation dataset. The study of some particular observations permit to go in the details of the sub-model predictions. Some models

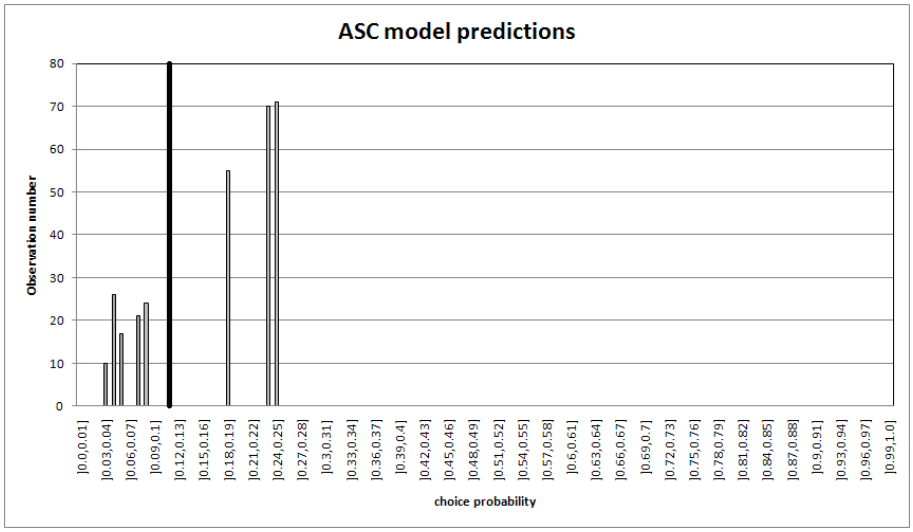

Figure 7: Distribution of the choice probabilities for the ASC model

predictions are shown in figure 8, 9, 10. On a picture, each column is relative to video frame, except the extreme right one. The first line contains the considered video frames. Each frame is the first one of a video second (each video second has 25 frames). The second line is relative to the expression choice model predictions. For each frame, the probability distribution among the expressions is presented. The order of the histogram is the following: Happiness, Surprise, Fear, Disgust, Sadness, Anger, Neutral, Other and Don't know. The third line corresponds to the frame choice model, for each frame the probability of choosing it, is displayed. Finally in the extreme right column, you find on the first line the video name; on the second one, there is the expression distribution predicted by the complete model; finally on the last line, the observed expression distribution is displayed. Concerning one video, the observed expressions distribution shows expression labels of the web survey respondents.

At the video beginning, the face tends to be more or less neutral, and then evolves toward a different expression. On figure 8, the subject face evolves toward a combination of fear and surprise. On the third frame the expression choice model predicts more fear than surprise, and the contrary for the last frame. The frame choice model predicts high choice probabilities for the two last frames, which is logical, due to their distances from the neutral expression. Finally the model predictions are very similar to the collected web survey labels (last column), which is a good point. On figure 9, the final expression predicted by the model is a mixture of surprise, fear, disgust, sadness and anger. Note that the expression choice model predictions are logical except for the first frame, which is usually the case. It seems to be hard for the model to recognize the neutral expression, probably because there are very few labels about it in the collected database (5.82% of the observations). This seems logical due to the video nature, evolving from neutral expression to another one. Respondents choose the other expression instead of the neutral one. But the frame choice model weight the first frame very

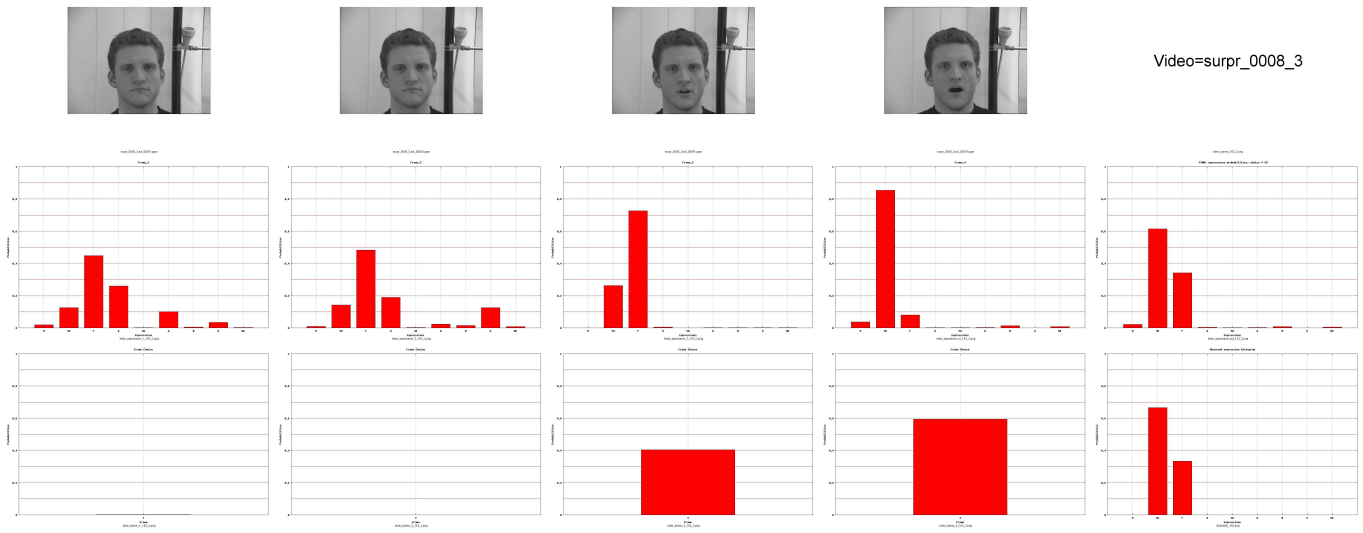

Figure 8: First example of detailed predictions for one observation

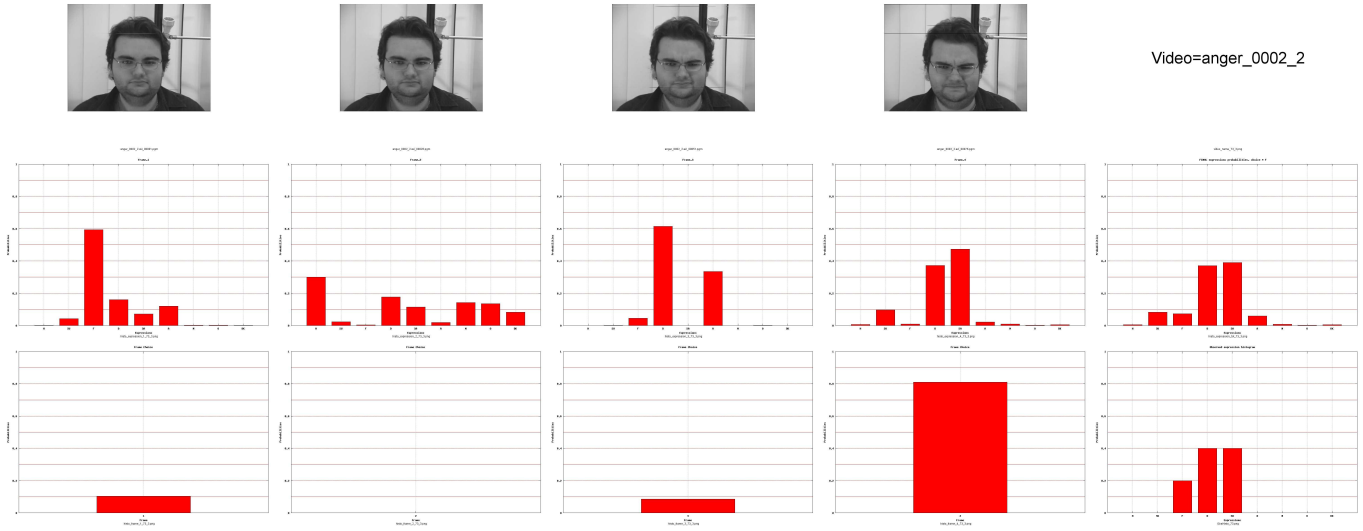

Figure 9: Second example of detailed predictions for one observation

low, in favour of last ones, compensating this problem. The model predictions and observed labels match well, even if it puts little probabilities on surprise and anger. Figure 10 deals with a non-ambiguous video, indeed all respondents put the happiness label on it. In that case, the sub-models predictions are very good, indeed the expressions distribution are logical for each frame, and the frame choice model detects well the last frame, when the subject starts to smile.

We conclude this section by underlying the superiority of the proposed model on the ASC model, showing the gain brought by adding explanatory variables. The quality of the model predictions is also good. For each video, it reproduces well the observed distribution of the expressions labels, collected with the internet survey. Finally, both sub-models predictions

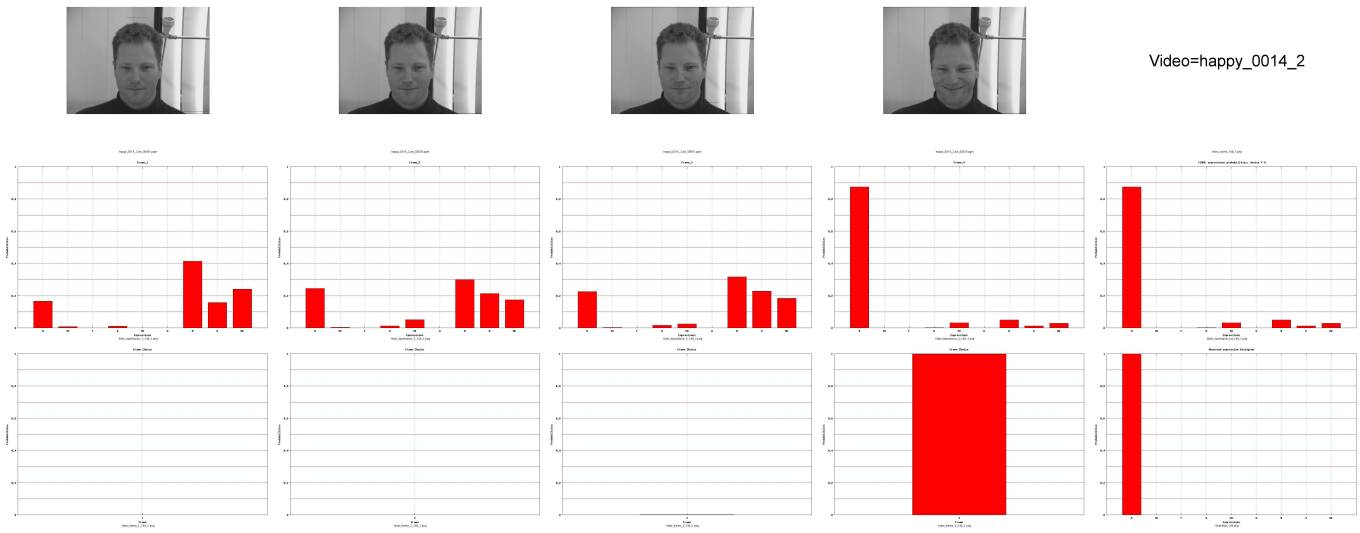

Figure 10: Third example of detailed predictions for one observation

have sense, proving the adequacy between the theoretical modelling framework and the dynamic facial expression recognition.

8 Conclusions and Perspectives

Compared to machine learning methods, we propose a new approach of dynamic facial expression recognition. The model estimation is not based on a single human ground truth, but on expressions labels collected beside internet survey respondents. In addition, the developed model point up causal effects between expressions and facial characteristics. Statistical tests and model predictions study have proved the model quality, compared to a simpler model, called ASC model. Finally qualitative exams of the proposed model predictions permit to check modeller intuition concerning the face video.

Even if this new model framework is meaningful, many improvements could be done. We saw in section 6 that the model has been estimated on a little dataset. Ideally the more observations we have, the better the model will be, so the model can be re-estimated with a higher number of observations. The number of videos is also a critical aspect, features variabilities are quite low and should be increased. This could permit to have more complex and complete specifications for both sub-models. In addition the panel data effect discussed in section 2 is not yet implemented and will allow to account for respondents specificities. In this work we have tested the model fit quality and prediction goodness on the estimation dataset. In order to prove the generality of such model, a validation step should be done on another dataset, not

implied in the estimation process, for example on the observations relative to Cohn-Kanade videos. Finally a comparison with a state of the art machine learning method, such as neural networks (NN) could demonstrate the superiority of the approach.

As is, the model can be used directly in transportation applications cited in the introduction, even if videos of the FEED database are not dedicated to transportation (indeed stimuli used to generate facial expressions were not necessarily related to the field). In a first time, this is not an insurmountable problem, in the sense that FEED videos are quite general, and labels about all expressions have been collected. Some case studies can be conducted in order to completely prove model applicability to transportation. For immediate applications, we can install cameras in front of users (drivers, or public transportation users), couple cameras with facial tracking systems, for extracting facial features, and then determine users facial expressions by using the proposed model. In a second time, we can think to completely dedicate the model to transportation, by estimating it on data relative to the field. Instead of FEED videos, some facial videos of transportation users in special situations could be employed. Video collection could consist in acquiring facial videos of drivers, when placed in simulators. Typical driving situations could be displayed as stimuli, to generate drivers expressions. Note that video collection experimental design has to be closely link to the application. Finally in the context of “Aware” vehicles, the proposed model could be incorporated in global emotion recognition systems, including others elements recognition, such as voice intonation or concentration.

Acknowledgments

We are very grateful to Matteo Sorci who provided the necessary programs used to extract facial features using AAM.

References

Abou-Zeid, M. (2009) Measuring and modeling travel and activity well-being, Ph.D. Thesis, Massachusetts Institute of Technology.

Bartlett, M. S., G. Littlewort, I. Fasel and J. R. Movellan (2003) Real time face detection and facial expression recognition: Development and applications to human computer interaction., paper presented at *Computer Vision and Pattern Recognition Workshop, 2003. CVPRW '03. Conference on*, vol. 5, 53–53, June 2003, ISSN 1063-6919.

Ben-Akiva, M. E. and S. R. Lerman (1985) *Discrete Choice Analysis: Theory and Application to Travel Demand*, MIT Press, Cambridge, Ma.

Bierlaire, M. (2003) BIOGEME: a free package for the estimation of discrete choice models, paper presented at *Proceedings of the 3rd Swiss Transportation Research Conference*, Ascona, Switzerland. Www.strc.ch.

Choudhury, C. F. (2007) Model driving decisions with latent plans, Ph.D. Thesis, Massachusetts institute of technology.

Cohen, I., N. Sebe, A. Garg, L. S. Chen and T. S. Huang (2003) Facial expression recognition from video sequences: temporal and static modeling, *Computer Vision and Image Understanding*, **91** (1-2) 160 – 187, ISSN 1077-3142. Special Issue on Face Recognition.

Cootes, T. F., G. V. Wheeler, K. N. Walker and C. J. Taylor (2002) View-based active appearance models, *Image and Vision Computing*, **20** (9-10) 657 – 664, ISSN 0262-8856.

Ekman, P. and W. Friesen (1978) *Facial action coding system: A technique for the measurement of facial movement*, Consulting Psychologists Press, Palo Alto, California.

Fasel, B. and J. Luettin (2003) Automatic facial expression analysis: a survey, *Pattern Recognition*, **36** (1) 259 – 275, ISSN 0031-3203.

J.L.Walker (2001) Extended discrete choice models: Integrated framework, flexible error structures, and latent variables, Ph.D. Thesis, Massachusetts Institute of Technology.

Keltner, P., D. Ekman (2000) Facial expression of emotion, in *Handbooks of emotions*, 236–249, M.Lewis & J.M.Havilland.

M.Sorci, M.Bierlaire, J-P.Thiran, J.Cruz, Th.Robin and G.Antonini (2008) Modeling human perception of static facial expressions, paper presented at *8th IEEE Int'l Conference on Automatic Face and Gesture Recognition*.

Pantic, M. and I. Patras (2006) Dynamics of facial expression: recognition of facial actions and their temporal segments from face profile image sequences, *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, **36** (2) 433–449, April 2006, ISSN 1083-4419.

Reimer, B., J. Coughlin and B. Mehler (2009) Development of a driver aware vehicle for monitoring, managing & motivating older operator behavior, *Technical Report*, ITS America, June 2009.

T.Kanade, Y.-L., J.Cohn (2000) Comprehensive database for facial expression analysis, paper presented at *Proceedings of the 4th IEEE International Conference on Automatic Face and Gesture Recognition (FG'00)*, 46–53, March 2000.

Wallhoff, F. (2004) Fgnet-facial expression and emotion database, *Technical Report*, Technische Universität München, `http://www.mmk.ei.tum.de/~waf/fgnet/feedtum.html`.