Modelling human perception of static facial expressions

M.Sorci,J.Ph.Thiran
Electrical Engineering Institute,EPFL
Station 11, CH-1015, Lausanne
{matteo.sorci,JP.Thiran}@epfl.ch

J.Cruz,T.Robin,M.Bierlaire
Transport and Mobility Laboratory,EPFL
Station 11, CH-1015, Lausanne
{javier.cruz,thomas.robin,michel.bierlaire}@epfl.ch

G.Antonini
IBM Zurich Lab
Saumerstrasse 4 ,Ruschlikon
gan@zurich.ibm.com

B.Cerretani
University of Siena
DII,Siena
barbara.cerretani@gmail.com

Abstract

Data collected through a recent web-based survey show that the perception (i.e. labeling) of a human facial expression by a human observer is a subjective process, which results in a lack of a unique ground-truth, as intended in the standard classification framework. In this paper we propose the use of Discrete Choice Models(DCM) for human perception of static facial expressions. Random utility functions are defined in order to capture the attractiveness, perceived by the human observer for an expression class, when asked to assign a label to an actual expression image. The utilities represent a natural way for the modeler to formalize her prior knowledge on the process. Starting with a model based on Facial Action Coding Systems (FACS), we subsequently defines two other models by adding two new sets of explanatory variables. The model parameters are learned through maximum likelihood estimation and a cross-validation procedure is used for validation purposes.

1. Introduction

Facial expressions are probably the most visual method to convey emotions and one of the most powerful means to relate to each other. In order to move toward real interacting human-computer systems, where algorithms written by humans should be able to capture, mimic and reproduce human perceptions, facial expressions play surely a central role. The dominant challenge in building such an automatic system, even if narrowed down to the 'only' facial expression perception task, arises from the fact that such a perception (performed by human beings in the real world) is absolutely subjective and strongly related to contextual information. Most of the available literature on the subject proposes a two step procedure in order to make the problem operational: first, a representation of the expression, learned from a set of pre-selected meaningful features, is computed. Such a step is necessary and when properly done provides reduced 'objects', bringing the information that matter. In [12] optical flow analysis is proposed, in order to model dynamically muscle activities and estimating the displacements of salient points. Gabor wavelet based filters have been used in [18], in order to build templates for facial expressions, over multiple scales and different orientations. Data driven methods, based on statistical generative models (PCA, ICA) are used in [16] and [2], in order to capture meaningful statistics of face images. Recent years have seen the increasing use of feature geometrical analysis ([6, 11]). The Active Appearance Model (AAM, see [6]) is one of these techniques which elegantly combines shape and texture models, in a statistical framework, providing as output a mask of face landmarks. We use AAM to produce part of our feature set.

The second step in the procedure consists in defining a decision/classification rule which associates the feature-based representation with the correct facial expression. Previous works used for this step several well-know methods: HMM-based classifier [12], template matching [18], SVM [15], Dynamic Bayesian Networks [9]. The standard hard classification approach associates any two examples having the same features to the same corresponding class. Moreover, one of the main assumptions is that the facial expression labels reported in the training set represent the true expressions. This background does not hold in our case. Facial expressions are ambiguous and different people perceive differently the same expression. This fact is even more accentuated in a static context, where the lack of transitions between following expressions deprives the observer

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

of an important source of information. A soft approach is more suitable in our case, where a probabilistic model assigns a probability value to an example to be perceived as a particular expression. First, flexibility is added when two equal feature vectors are not necessarily assigned to the same class while at the same time the respective probability values do depend on the feature vectors. Second, the soft approach could relate the computed discrete probability distribution over the different expressions to the heterogeneity in the human observer population, potentially explaining such a difference in perceptions among different individuals. Based on the previous considerations, we believe that Random Utility Theory (RUM) and more specifically Discrete Choice Models (DCM) well fit our needs and they represent a reasonable and theoretically grounded modeling framework.
The paper starts describing the available data in Section 2, followed by the description of the feature set we used in our model and the relative methods. To make the paper as self-contained as possible, we give in Section 3 a short overview of RUM and DCM theoretical principles, while in Section 4,5,6 we go into the details, respectively, of the model specification, the estimation of the related parameters and the model validation procedure. We end in Section 7 with final remarks and some idea for future works.

2. Problem requirements

There are two crucial steps that a model maker has to deal with when modelling a certain behaviour: the choice and the collection of the data she needs to use in her study and the identification of the measures (explanatory variables) describing the phenomenon. In the following we focus on the choices we make for our problem.

2.1. Data description

Modelling *human* perception of *facial expressions* implies the need for two sort of database: one for the facial expressions, a set of images of subjects performing expressions, and a database collecting how human perceives them. As for the images database we use the Cohn-Kanade Database [10]. The database consists of expression sequences of subjects, starting from a neutral expression and ending most of the time in the peak of the facial expression. The 104 subjects of the database are university students enrolled in introductory psychology classes. Six of the displays were based on descriptions of prototypic emotions (i.e, happiness, anger, fear, disgust, sadness and surprise). Before performing each display, an experimenter described and modelled the desired display. The choice for this database is twofold. Firstly, the Cohn-Kanade Database is one of the few available facial expressions databases and secondly this is the database used by Sorci *et al*. [17] in their facial expressions evaluation survey. In August 2006 Sorci *et al*. [17] published the internet facial expressions evaluation survey (http://lts5www.epfl.ch/face) in order to find a way to directly get humans' perception of facial expressions. The ultimate aim of the survey is to collect a dataset created by a population of real human observers, from all around the world, doing different jobs, having different cultural backgrounds, ages and gender, belonging to different ethnic groups, doing the survey from different places (work, home, on travel, etc.). The images used in the survey comes from the Cohn-Kanade Database. Over the 104 subjects in the database, only 11 of them gave the consent for publications. The subset of the Cohn-Kanade Database used in this survey consists of the 1274 images of these 11 subjects (9 women and 2 men). In the survey the participants are asked to annotate a certain number of images randomly chosen on the whole set. The annotation process consists in associating an expression label (among a set of available human expressions) to each of the images that will be presented to the survey's participant. In the list of the available expressions the authors included, in addition to the 6 prototypic expressions (happiness,surprise,fear,anger,disgust,sadness), the neutral one, the "I don't know" and "Other" options. The last two options have been introduced in order to deal with images extremely ambiguous to the participant. In the database 1780 participants have taken part to the survey for around 40000 annotated images. As far as we are aware, this is the only database collecting this kind of information.

2.2. Features: description and extraction

The survey, described in the previous paragraph, provides the raw data capturing the participants perception of facial expressions. This raw data consists on a set of facial expressions images (the Cohn-Kanade images) and the set of participants choices among the nine options. In order to exploit the information coming from both sources we need to identify and represent the facial visual cues describing an expression. Nowadays the Facial Action Coding Systems (FACS) [8] represents the leading standard for measuring facial expressions in behavioural science. For this reason we have decided to rely on the main measures suggested by this human observer system. In the rest of the paragraph we detail the Computer Vision tool used to represent a face, we describe the FACS system and its measures and we introduce a new and complementary set of visual measures improving the descriptiveness of the expressions.

Face Active Appearance Model Active appearance models (AAMs) are generative models commonly used to model faces which elegantly combines shape and texture models, in a statistical framework, providing as output a mask of face landmarks [6]. The appearance variability is modeled by applying the Principal Component Analysis

Emotional Category	Primary Visual Cues					Auxiliary Visual Cues					
	AU	AU	AU	AU	AU	AU	AU	AU	AU	AU	Transient Feature(s)
Happiness	6	12				25	26	16			Wrinkles on outer eye canthi, presence of nasolabial furrow
Sadness	1	15	17			4	7	25	26		
Disgust	9	10				17	25	26			Presence of nasolabial furrow
Surprise	5	26	27	1+2							Furrows on the forehead
Anger	2	4	7	23	24	17	25	26	16		Vertical furrows between brows
Fear	20	1+5	5+7			4	5	7	25	26	

Table 1. The association of six emotional expressions to AUs, AU combinations, and Transient Features (from [9])

(PCA) to the shape $\mathbf{s_i}$ and texture $\mathbf{g_i}$:

$$\mathbf{s_i} = \bar{\mathbf{s}} + \mathbf{\Phi_s b_{si}} \quad and \quad \mathbf{g_i} = \bar{\mathbf{g}} + \mathbf{\Phi_t b_{ti}} \tag{1}$$

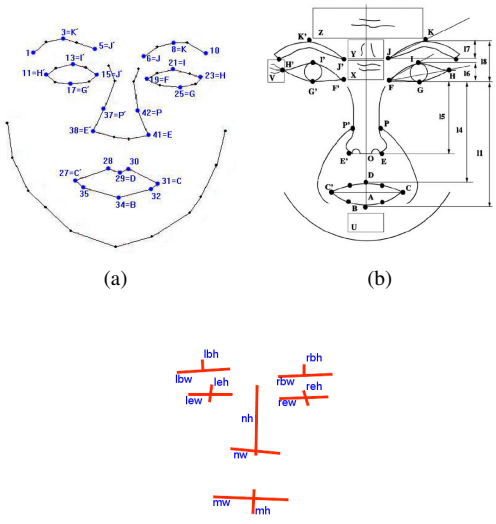
where Φ_s and Φ_t are the matrices describing the modes of variation derived from the training set, b_{si} and b_{ti} the mean shape and texture. The unification of the presented shape and texture models into one complete appearance model is obtained by concatenating the vectors b_{si} and b_{ti} and performing a further PCA:

$$\mathbf{b_i} = \mathbf{\Phi_c c_i} \tag{2}$$

The vector of appearance parameters $\mathbf{c_i}$ allows to control simultaneously both shape and texture.
The statistical model is then given by:

$$\mathbf{s_i} = \bar{\mathbf{s}} + \mathbf{Q_s c_i} \quad and \quad \mathbf{g_i} = \bar{\mathbf{g}} + \mathbf{Q_t c_i} \tag{3}$$

where Q_s and Q_t are the matrices describing the principal modes of the combined variations.



(a) (b) (c)

Figure 1. a) Facial landmarks (55 points); b) the geometrical relationship of facial feature points, where the rectangles represent the regions of furrows and wrinkles; c) Featural descriptors used in the definition of the EDUs

FACS Facial expressions represent a visible consequence of facial muscle and autonomic nervous system actions: is it possible to describe and quantify every action the face can perform? Ekman and Friesen [8] provided an answer to this question with their Facial Action Coding System (FACS), by measuring all visible movements. Ideally, FACS would differentiate every change in muscular action, but it is limited to what a user can reliably discriminate. FACS measurement units are called "action units" (AUs) and represent the muscular activity that produces momentary changes in facial appearance. A facial expression is indeed the combination of AUs. In particular, the six basic emotions (happiness, anger, disgust, fear,surprise and sadness) have been postulated by Ekman [7] as having a distinctive content together with a unique facial expressions. Zhang *et al.* [9] group AUs of facial expressions as primary AUs and auxiliary AUs, see Table 1. The primary AUs refer to those AUs or AU combinations that univocally describe one of the 6 expressions. The auxiliary AUs provide an additional support to the expression description. Additionally, changes in facial transient features, such as wrinkles and furrows, also provide support cues to infer certain expressions. In order to transform the AUs in a set of quantitatively measures Zhang *et al.* [9] translate these appearance changes descriptors in a set of geometrical relationships of some facial feature points. In our work we use the same geometrical relationships, but with the different goal of modelling the human perception of expressions based on the response of the heterogeneous group of participants. We use the AAM to measure the set of angles and distances reported in Table 2. Indeed the application of the AAM model on each image of the survey provides a mask as the one in Figure 1(a). Figure 1(b) shows the relations between the features points suggested by Zhang *et al.* [9] and the landmarks automatically extracted by AAM. The presence of furrows and wrinkles on a face image can be determined by edge feature analysis in the areas where transient features appear. The regions of facial wrinkles and furrows are indicated by rectangles in Fig. 1(b). The change of wrinkles in the region $\square X$ is directly related to AU9 (Nose Wrinkler). The furrows in the regions $\square Z$, $\square Y$, $\square V$, $\square U$ provide diagnostic information for the identification of AU2 (Outer Brow Raiser), AU4 (Brow Lowerer), AU6 (Cheek Raiser), and AU17 (Chin Raiser), respectively. In order to detect these

features, the edge detection with embedded confidence, proposed by Meer and Georgescu [14], is used.

AUs	Facial Visual Cues
AU1	$\angle FHJ$, $\overline{JF}$ increased OR $\overline{JF}$ increased, $l8$ nonincreased
AU2	$l8$ increased and $\overline{JF}$ nonincreased furrow in $\square Z$ increased
AU4	$l8, \overline{FJ}, \overline{JJ'}, \overline{FP}, \overline{F'P'}$ decreased, $\angle HFI$ increased and wrinkle in $\square Y$
AU5	$l6$, $\overline{JF}$ and $\overline{JJ'}$
AU6	nasolabial furrow presence and wrinkle in $\square V$
AU7	$\angle HFI$ nonincreased and $\angle HGF$ increased
AU9	wrinkle increased in $\square X$ nasolabial furrow presence OR $\overline{PF}, \overline{FJ}$ decreased
AU10	$l4$ decreased and $\lvert\overline{FC} - \overline{F'C'}\rvert$ increased, nasolabial presence OR $\overline{OD}$ decreased, $\overline{DB}, \overline{C'C}$ increased
AU12	$\overline{FC}, \overline{F'C'}$ decreased, $\overline{CC'}$ increased, $\overline{GI}$ nonincreased
AU15	$\overline{FC}, \overline{F'C'}, \overline{CC'}$ increased
AU16	$\overline{OD}$ nonchange, $\overline{DB}$ decreased
AU17	$\overline{OB}$ decreased and wrinkle in $\square U$ presence
AU20	$\overline{CC'}$ increased and $\overline{FC}, \overline{F'C'}$ nonchange
AU23	$\overline{DB}, \overline{CC'}$ decreased
AU24	$\overline{DB}$ decreased, $\overline{CC'}$ nonchange
AU25	$\overline{DB}$ increased, $\overline{DB} < T_1, \overline{CC'}$ nonincreased
AU26	$T_1 < \overline{DB} < T_2, \overline{CC'}$ nonincreased
AU27	$\overline{DB} > T_2, \overline{CC'}$ nonincreased

Table 2. *Linguistic description of the AUs of Figure 1 (from [9])*

Expressions Descriptive Units In the visual perception community there is a general agreement on the fact that face recognition is the result of two main sources of information: the featural one coming from individual facial features (mouth, nose, etc.) and the configural one related to the facial layout and configuration of the previous features [5]. The measures extrapolated by the FACS give information about isolated components in a face, providing a featural contribution to face representation. In order to exploit the combination of these two useful sources we have decided to add a group of measures encoding the interactions among the featural descriptors showed in Figure 1(c). The new set of measures, called Expression Descriptive Unit (EDU) and reported in Table 3, has been introduced by Antonini *et al*. [1]. The first 5 EDUs represent, respectively, the eccentricity of eyes, left and right eyebrows, mouth and nose. The EDUs from 7 to 9 represent the eyes interactions with mouth and nose, while the 10th EDU is the nose-mouth relational unit. The last 4 EDUs relate the eyebrows to mouth and nose. The EDUs can be intuitively interpreted. For example, in a face displaying a surprise expression, the eyes and the mouth are usually opened and this can be captured by EDU7 ($eye_{height}/mouth_{height}$).

EDU1	$\frac{lew+rew}{leh+reh}$	**EDU8**	$\frac{leh+reh}{lbh+rbh}$
EDU2	$\frac{lbw}{lbh}$	**EDU9**	$\frac{lew}{nw}$
EDU3	$\frac{rbw}{rbh}$	**EDU10**	$\frac{nw}{mw}$
EDU4	$\frac{mw}{mh}$	**EDU11**	EDU2 / EDU4
EDU5	$\frac{nh}{nw}$	**EDU12**	EDU3 / EDU4
EDU6	$\frac{lew}{mw}$	**EDU13**	EDU2 / EDU10
EDU7	$\frac{leh}{mh}$	**EDU14**	EDU3 / EDU10

Table 3. Expressions Descriptive Units

The appearance parameters FACS and EDU provide measures of local facial features or areas that are prone to change with facial expressions, but they do not provide a description of a face as a global entity. This information can be obtained considering the appearance vector c matching the face in the processed image. Figure 2 shows the effect of varying the first appearance model parameter, showing changes in identity and expression.

Figure 2. Examples of synthesized faces obtained varying the first c parameter from the mean face ($\pm 3std$).

3. Discrete Choice Models

Discrete choice models are known in econometrics since the late 50's. They are defined to describe the behavior of people in choice situations, when the set of available alternatives is finite and discrete (choice set). They are based on the concept of *utility maximization* in economics, where the decision maker is assumed to be *rational*, performing a choice in order to maximize the utilities she perceives from the alternatives. The alternatives are supposed to be mutually exclusive and collectively exhaustive, while the rationality of the decision maker implies transitive and coherent preferences. The utility is a *latent* construct, which is not directly observed by the modeler, and is treated as a random variable. The discrete choice paradigm well matches the labelling assignment process of the participants in the survey. This approach can be interpreted as an attempt to model the decision process performed by an hypothetical human observer during the labelling procedure for the facial expressions. Given a population of N individuals, the (random) utility function U_{in} perceived by individual n from alternative i, given a choice set C_n, is defined as follows:

$$U_{in} = V_{in} + \varepsilon_{in} \tag{4}$$

It is composed by the sum of a deterministic term V_{in}, capturing the systematic behaviour (features extracted from a face), and a random term ε_{in}, capturing the uncertainty. This random term captures the uncertainty on unobserved attributes, unobserved individual characteristics, measurement errors and instrumental variables. We actually do not observe the real values of the utilities as perceived by the participant and we need a framework to deal with this uncertainty. Under the utility maximization assumption, the output of the model is represented by the choice probability that individual n will choose alternative i, given the choice set C_n. It is given by:

$$P_n(i|C_n) = P_n(U_{in} \geq U_{jn}, \forall j \in C_n, j \neq i) = \int_{\varepsilon_n} I(\varepsilon_n < V_{in} - V_{jn}, \forall j \in C_n, j \neq i) f(\varepsilon_n) d\varepsilon_n \quad (5)$$

where $\tilde{\varepsilon} = \varepsilon_{jn} - \varepsilon_{in}$. Based on Equation 5, in order to define the choice probability, only the difference between the utilities matters. The specification of the utility functions represents the modeler's mean to add her prior knowledge on the choice process. Different models are obtained making different assumptions on the ε_{in} term. A family of models widely used in literature are the GEV (Generalized Extreme Value) models, introduced by [13]. GEV models provide a closed form solution for the choice probability integral in 5, allowing at the same time for a certain flexibility in designing the variance/covariance structure of the problem at hand (i.e., several correlation patterns between the alternatives can be explicitly captured by these models). Assuming the error terms being multivariate type I extreme value distributed the general expression of the GEV choice probability for a given individual to choose alternative i, given a choice set C with J alternatives, is as follows:

$$P(i|C) = \frac{e^{V_i + logG_i(y_1,...,y_J)}}{\sum_{j=1}^{J} e^{V_j + logG_j(y_1,...,y_J)}} \quad (6)$$

where $y_i = e^{V_i}$ and $G_i = \frac{\partial G}{\partial y_i}$. The function G is called *generating function* and it captures the correlation patterns between the alternatives. Details about the mathematical properties of G are reported in [13] (differentiable and homogeneous of degree $\mu > 0$, among the others). Several GEV models can be derived from Equation 6, through different specifications of the generating function. In this paper we use a Multinomial Logit Model (MNL), which is largely the simplest and most used discrete choice model in literature. It is obtained assuming the following G function, which implies no correlations between the alternatives:

$$G(y_1, ..., y_J) = \sum_{j \in C} y_j^{\mu} \quad (7)$$

where μ is a positive scale parameter. Under these assumptions, the MNL choice probability is given by the following expression

$$P_n(i|C_n) = \frac{e^{\mu V_{in}}}{\sum\limits_{j \in C_n} e^{\mu V_{jn}}} \quad (8)$$

In this work the choice set C_n is represented by the 9 survey alternatives ("happiness", "surprise", "fear", "disgust", "sadness", "anger", "neutral", "other" and "I don't know").

4. Model Specifications

In this paragraph we focus on the deterministic part V_i of the random utility function (see equation 4). Any alternative i can be described in terms of a combination of a certain number of attributes (explanatory variables) EV_i reflecting reasonable hypotheses about the effects of these variables on the corresponding utility. The model maker's a priori knowledge of the phenomenon plays a key role in the specification of V_i. This knowledge is reflected in a set of specific assumptions about the relationships between the set of explanatory variables observed by the model maker. For our problem we can rely on the FACS system as a valid a priori theory providing us the first set of explanatory variables to use in the construction of the model. In line with the considerations made in the previous section, we improve the descriptiveness of the model by adding the measures induced by the EDUs and by the appearance parameters c. The model design process has been done, indeed, in three consecutive steps. Each step defines a model by adding a new set of explanatory variables to the "cleaned" model of the previous stage. Cleaning a model consists in evaluating the model, retaining those parameters that are statistically significant (*t*-test statistic against the zero value) and combining those that are correlated. In the first step, the local measures of the face, coming from the AUs defined by the FACS, are used to define the systematic utility functions of the MNL model, *Model F* in eq.9. In the second step the local interactions between facial features provided by the EDUs are included, *Model FE* in eq.9. In the last model the c appearance parameters, encoding global measures about the face, are finally added to the two previous sets of measures, *Model FEC* in eq.9. For all the models the utility functions are specified using a linear-in-parameters form, combining the explanatory variables chosen by the model maker. The choice of a linear form is based purely on simplicity considerations, in order to reduce the number of parameters in the estimation process. The following equation summarizes the form of the utility for each of the three developed models:

$$\begin{aligned} V_j = ASC_j &+ \textstyle\sum_{k=1}^{K_F} I_{kj}^F \beta_{kj}^F EV_k^F \quad \text{, Model F} \\ &+ \textstyle\sum_{h=1}^{K_E} I_{hj}^E \beta_{hj}^E EV_h^E \quad \text{, Model FE} \\ &+ \textstyle\sum_{l=1}^{K_C} I_{kl}^C \beta_{lj}^C EV_l^C \quad \text{, Model FEC} \end{aligned} \quad (9)$$

F MODEL			FE MODEL			FEC MODEL		
β^F_{ki}	**estimate**	***t* test 0**	β^{FE}_{ki}	**estimate**	***t* test 0**	β^{FEC}_{ki}	**estimate**	***t* test 0**
β^F_{17H}	+ 103	+ 56.81	β^{FE}_{17H}	+ 34	+ 4.98	β^{FEC}_{17H}	+ 105	+ 37.67
			β^{FE}_{31SU}	+ 8.12	+ 48.3	β^{FEC}_{31SU}	+ 6.89	+ 39.59
						β^{FEC}_{46A}	- 9.67	- 11.13
β^F_{17H}=mouth width Happiness, β^{FE}_{31SU}=EDU4 Surprise, β^{FEC}_{46A}=C5 Anger								
Sample size = 38110			Sample size = 38110			Sample size = 38110		
Nb. of estimated parameters = 93			Nb. of estimated parameters = 120			Nb. of estimated parameters = 139		
Null log-likelihood = - 83736.229			Null log-likelihood = - 83736.229			Null log-likelihood = - 83736.229		
Final log-likelihood = - 57072.872			Final log-likelihood = - 55027.381			Final log-likelihood = - 53474.271		
Likelihood ratio test = 53326.712			Likelihood ratio test = 57417.695			Likelihood ratio test = 60523.915		
$\bar{\rho}^2$ = 0.317			$\bar{\rho}^2$ = 0.341			$\bar{\rho}^2$ = 0.360		

Table 4. **MNL** Part of the estimation results

(a) Expressions decreasing C5 (b) Neutral (c) Expressions increasing C5

Figure 3. Interpretation of the β^{FEC}_{46A} parameter. The central image refers to the neutral expression. The images on the right correspond to an increase of the c5 parameter, those on the left to a decrease on the c5 parameter

where $j \in$ {"happiness", "surprise", "fear", "disgust", "sadness", "anger", "neutral", "other", "I don't know"}, $\{F, E, C\}$ refer respectively to the FACS, EDUs and the appearance parameters c, $EV^{\{F,E,C\}}_{\{k,h,l\}}$ refers to $\{k, h, l\}$-th explanatory variable of one of the used sets, $K_{\{F,E,C\}}$ is the total number of the explanatory variables for each set, $I^{\{F,E,C\}}_{kj}$ is an indicator function equal to 1 if the k-th explanatory variable is included in the utility for the alternative j and 0 otherwise, $\beta^{\{F,E,C\}}_{kj}$ is the weight for the k-th EV in alternative j and ASC_j is an alternative specific constant. The ASC_j coefficients captures the average effect on the utility of all factors that are not included in the model. For identification purposes the absolute values of the constants must be normalized. This normalization is obtained by fixing one of the constant to zero. In our case the neutral alternative is considered as the reference alternative and its ASC is set to zero.

5. Model Estimation

The parameters of the three models, introduced in the previous paragraph, have been estimated by maximum likelihood estimation through a sequential quadratic programming algorithm, using the Biogeme package [4]. Note that such nonlinear programming algorithms identify local maxima of the likelihood function. We performed various runs, with different starting points (a trivial model with all parameters to zero, and the estimated value of several intermediary models). They all converged to the same solution. In Table 4 we report the final coefficients estimates for some β for the three models. In the first half of the table, each row relates each particular β for a specific model to its estimated coefficient and its associated *t*-statistic values. The second half of the table shows summary statistics for the entire estimation run for each of the three models. The interpretation of the model estimation outputs is performed, at first, considering a given structure of the model and analysing the significance of the explanatory variables in the utility functions and then comparing the goodness of fit among the different specifications. The most basic test, in the given structure analysis case, consists in the examination of the values of the coefficient estimates. The goal of this informal test is to check if the signs of the estimated coefficients reflects our a priori expectations. In the three cases, the learned parameters show important consistencies with the common reading of facial expressions in terms of facial component modifications. For space reasons, we report in Table 4 only a subset of β_{ki} estimates. A parameter is considered significant if the norm of the t-test against 0 is bigger than 1.96. β^F_{17H} represents the coefficient of the mouth width measure in the happiness expression. It is a FACS parameter and it is included in all the specifications. Its positive value shows a positive impact on the respective utility. This means that an increase of the mouth width with respect to the neutral expression (the reference one in our model) corresponds to higher utilities for the happiness alternative. The β^F_{17H} estimate is inline with the FACS expectations for the happiness expression. The first row in Table 1 describes the FACS

happiness encoding in terms of the primary action units 6 and 12. During an AU12 a stretching of the mouth's lip corners is expected. This corresponds indeed to an increase of the measure $\overline{CC'}$ associated to the estimated parameter β^{F}_{17H} and representing the mouth width. β^{FE}_{31SU} is the parameter related to EDU4 (Table 3) describing the mouth eccentricity in the surprise alternative. Its positive sign explains the expected behaviour of the mouth in subjects performing a surprise expression, where the mouth movement leads to a lower mouth's height and a higher mouth's width, with respect to the reference alternative. The third parameter β^{FEC}_{46A} is the coefficient related to the fifth appearance parameters c for the anger utility. The bigger this coefficient is the more negative is the impact on the anger utility. We can visually interpret this result by looking at Figure 3. Considering the neutral c5 value as the reference value, we can notice how increasing this parameter (leaving unchanged the others) we move towards an happiness-like expression, whereas an anger-like face corresponds to values of c5 smaller than the reference one. The statistics concerning the goodness of fit for the three different models are reported in the second half of Table 4. It can be observed that for the second model the fitting is better than for the first one (higher log-likelihood and $\bar{\rho}^2$) and the same for the third model with respect to the second one. The proposed models have been built in a nested way. This means that the first model is a restricted version of the second one and the latest a restriction of the third one. The restrictions imply that the restricted model can be obtained as a special case of the unrestricted one. In this case, a *likelihood ratio test* [3] can be used to verify if the additional variables of the unrestricted model add a significant explanatory power to the model and compensate for the degrees of freedom used by the fuller specification. The null hypothesis for this test states that the restricted and unrestricted models are equivalent. The statistic to compute the test is

$$-2(\mathcal{L}(\hat{\beta_R}) - \mathcal{L}(\hat{\beta_U})) \sim \chi^2_{K_U - K_R} \qquad (10)$$

where K_i is the number of parameters of the model i and χ^2_j is a χ^2 distribution with j degrees of freedom. Usually, a significance level of 95% is taken, and then the null hypothesis is rejected if the test value is above the threshold provided by the χ^2 distribution corresponding to the j degrees of freedom. The results for this test are reported in Table 5. The performed tests refer to the two possible *(restriced,unrestricted)* models couples. The first test shows that the inclusion of new parameters makes the unrestricted FE model significantly different from its restricted counterpart, the F model. This result justifies the second test comparing the the most complex model (FEC) with its restricted version (FE), showing that the model considering the whole set of 3 different explanatory variables can be considered and retained as the final model that best fit our data.

Performed test	Degrees of freedom	Test value	χ^2 Th.
F vs FE	27	4090.98	40.11
FE vs FEC	19	3106.22	30.14

Table 5. Summary of the different performed likelihood ratio tests

6. Model Validation

Model	JD	CB	NC
FEC	0.23 ± 0.0221	0.60 ± 0.0346	0.76 ± 0.0496
FE	0.25 ± 0.0239	0.59 ± 0.0376	0.72 ± 0.0582
F	0.26 ± 0.0231	0.57 ± 0.0350	0.72 ± 0.0528

Table 6. Cross-validation results.

Models validation has been performed by means of a subject based cross-validation. The observations from the survey are split into eleven groups corresponding to the the number of subjects of the survey. Based on this partition we perform a 11-fold cross-validation. Each fold contains all the survey observations for all the images of a single subject. A subset of the data is, in turn, held out and used as a validating set; the model is fit on the remaining data (calibration set) and used to predict for the validation set. For each image in the validation set, the predicted probability distribution over the expressions is compared with the observed one. The comparison is performed using several measures useful to compare probability distribution. The aforementioned measures belong to the category of the bin-by-bin similarity measures, in which only pairs of bins with the same index in the two compared histograms are matched. The similarity between two histograms is a combination of all the pairwise comparisons. In the following, the nine bins in the histograms $O^i = \{o^i_b\}$ and $P^i = \{p^i_b\}$ represent the mass of the distribution that falls into the corresponding bin (survey alternative *b*) for the image *i*. In particular, $O^i = \{o^i_b\}$ refers to the participants responses distribution for the image *i* and it is considered as the groundtruth, whereas $P^i = \{p^i_b\}$ represents the predicted distribution of the estimated model for the same image.The used measures are: the Jeffrey Divergence, the City-Block metric and the Normalized Correlation. The empirically derived Jeffrey Divergence is a modification of the KL divergence that is symmetric and numerically stable when comparing two empirical distributions. It is a measure of the inefficiency of assuming that the distribution is P^i when the true distribution is O^i. Although the JD measure is always non negative and is zero if and only if $o^i = p^i$, it is not symmetric and does not satisfy the triangle inequality. It is defined as:

$$d_J(O^i, P^i) = \sum_b \left(o^i_b \log \frac{o^i_b}{m^i_b} + p^i_b \log \frac{p^i_b}{m^i_b} \right) \qquad (11)$$

where $m^i_b = \frac{o^i_b + p^i_b}{2}$. The JD is defined in the interval $[0, +\infty)$. The closer their value to zero the more similar the

two distributions should be. The city-block metric (CB):

$$d_{\cap}(O^i, P^i) = 1 - \sum_b \left| o_b^i - p_b^i \right| \qquad (12)$$

provides a value in the interval $[0, 1]$. The closer the value to one the more similar distributions should be. The Normalized Correlation (NC) is a widely used measure to describe the similarity between two vectors, in pattern classification and signal processing problems. It is defined as:

$$d_{NC}(O^i, P^i) = 1 - \frac{\sum_i o_b^i p_b^i}{\sqrt{\sum_b o_b^i o_b^i} \sqrt{\sum_b p_b^i p_b^i}} \qquad (13)$$

This measure, as the JD divergence, is not a metric since it cannot satisfy the non-negativity and the triangle inequality. As for the CB, the NC is defined in $[0, 1]$ with expected values for similar mass distributions close to one. For the calibration of the model on each set, the Biogeme package described in Section 5 is used. The prediction is performed applying the estimated models on the validation set by means of the Biosim package (available at the same address of Biogeme). BioSim performs a sample enumeration on the validation data, providing for each of them the utilities and the choice probabilities for each alternative in the choice set. The results of the cross-validation are presented in Table 6. For each fold we compute the mean of the (O^i, P^i) similarities measures. The values reported in the table represents the average and the variance of the per-fold mean distance over the eleven fold for the three considered models. The measures are all coherent, showing the robustness of the specification of the FEC model.

7. Conclusions

In this paper we propose a new method for facial expressions modelling, based on discrete choice analysis. The data of the facial evaluation survey suggests that a subjective component biases the labeling process, requiring a detailed statistical analysis on the collected data. DCM paradigm well matches the human observer labeling procedure, allowing to capture and model the subjective perception of the choice makers. We showed how to improve the descriptiveness of the model by sequentially introducing complementary set of features. The parameters estimation of the three proposed models, has shown the correctness of the chosen sets of features, revealing the best fitting behavior of the third and most complex model. The use of a cross-validation has allowed to validate the proposed models. We are currently working on the comparison of our approach with some other techniques, especially with neural networks that have analogies with DCMs.

References

[1] G. Antonini, M. Sorci, M. Bierlaire, and J. Thiran. Discrete choice models for static facial expression recognition. In *8th International Conference on Advanced Concepts for Intelligent Vision Systems*, pages 710–721, 2006.

[2] M. Bartlett. Face image analysis by unsupervised learning and redundancy reduction, 1998.

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

[4] M. Bierlaire. BIOGEME: a free package for the estimation of discrete choice models. In *Proceedings of the 3rd Swiss Transportation Research Conference*, Ascona, Switzerland, 2003. www.strc.ch.

[5] R. Cabeza and T. Kato. Features are also important: Contributions of featural and configural processing to face recognition.

[6] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 23:681–685, June 2001.

[7] P. Ekman and W. V. Friesen. Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, (17):124–129, 1971.

[8] P. Ekman and W. V. Friesen. *Facial Action Coding System Investigator's Guide*. Consulting Psycologist Press, Palo Alto, CA, 1978.

[9] Y. Z. Q. Ji. Active and dynamic information fusion for facial expression understanding from image sequences. *Transactions on Pattern Analysis and Machine Intelligence*, 27(5):699–714, May 2005.

[10] T. Kanade, J. Cohn, and Y. L. Tian. Comprehensive database for facial expression analysis. In *Proceedings of the 4th IEEE International Conference on Automatic Face and Gesture Recognition (FG'00)*, pages 46 – 53, March 2000.

[11] A. Lanitis, C. J. Taylor, and T. F. Cootes. Automatic interpretation and coding of face images using flexible models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 19(7):743–756, 1997.

[12] J. Lien. Automatic recognition of facial expressions using hidden markov models and estimation of expression intensity, 1998.

[13] D. McFadden. Modelling the choice of residential location. In A. Karlquist *et al.*, editor, *Spatial interaction theory and residential location*, pages 75–96, Amsterdam, 1978. North-Holland.

[14] B. Meer, P.; Georgescu. Edge detection with embedded confidence. *Transactions on Pattern Analysis and Machine Intelligence*, 23(12):1351–1365, Dec 2001.

[15] P. Michel and R. E. Kaliouby. Real time facial expression recognition in video using support vector machines. In *ICMI '03: Proceedings of the 5th international conference on Multimodal interfaces*, pages 258–264, New York, NY, USA, 2003. ACM Press.

[16] C. Padgett and G. Cottrell. *Representing face images for emotion classification*. MIT Press, Cambridge, MA, 1997.

[17] M. Sorci, G. Antonini, J.-P. Thiran, and M. Bierlaire. Facial Expressions Evaluation Survey, 2007. ITS.

[18] J. Ye, Y. Zhan, and S. Song. Facial expression features extraction based on gabor wavelet transformation. In *IEEE International Conference on Systems, Man and Cybernetics*, pages 10–13, October 2004.