Travellers well-being measuring and dynamic facial expression recognition

Thomas Robin, Transport and Mobility Laboratory, Ecole Polytechnique Fédérale de Lausanne, CH-1015 Lausanne, Switzerland, thomas.robin@epfl.ch

Michel Bierlaire, Transport and Mobility Laboratory, Ecole Polytechnique Fédérale de Lausanne, CH-1015 Lausanne, Switzerland, michel.bierlaire@epfl.ch

Javier Cruz, Transport and Mobility Laboratory, Ecole Polytechnique Fédérale de Lausanne, CH-1015 Lausanne, Switzerland, javier.cruz@epfl.ch

Abstract

We propose a dynamic facial expression recognition framework based on discrete choice models (DCM). We model the choice of a person who has to label a video sequence representing a facial expression. The originality is based on the explicit modeling of causal effects between the facial features and the recognition of the expression. The model is composed of two parts. The first part captures the evaluation of the facial expression within each frame in the sequence. The second part determines which frame triggers the choice. The model is estimated using videos from the Facial Expressions and Emotions Database (FEED). Labeling data on the videos have been obtained using an internet survey available at *http://transp-or2.epfl.ch/videosurvey/*. The prediction capability of the model is studied in order to check its validity.

keywords: discrete choice, dynamic, latent model, facial expression recognition, video.

1 Introduction

Facial expressions are essential to convey emotions and represent a powerful way used by human beings to relate to each other. When developing human machine interfaces, where computers have to take into account human emotions, automatic recognition of facial expressions plays a central role. In this analysis, we propose a model predicting the evolution of a person who has to identify the expression of a human face on a video.

Some coding systems have been proposed to describe facial expressions. Ekman and Friesen (1978) have introduced the facial action coding system (FACS). They identify a list of fundamental expressions and associate groups of muscles tenseness or relaxations, called action units (AU) to each basic expression. A FACS expert can recognize AU activated on a face, and then deduct precisely the facial expression mixture. This is now the coding system of reference to characterize facial expressions.

The dynamic facial expression recognition (DFER) refers to the recognition of facial expressions in videos, whereas the static facial expression recognition (SFER) concerns the recognition of facial expressions in images. The DFER is an extension of the SFER. The DFER is a well known topic in computer vision. A great deal of research has been conducted in the field. 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 automatic segmentation and recognition of human facial expression from video sequences. Pantic and Patras (2006) present a dynamic system capable of recognizing facial AU and 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 the DFER. They focus particularly on the robustness in case of environmental changes.

There is a recent interest for quantifying facial expressions in different fields such as robotic, marketing or transportation. In the robotic field, Tojo et al. (2000) have implemented facial and body expressions on a conversational robot. With some experiments, they showed the added value of such a system in the communication between humans and the robot. Miwa et al. (2004) have also developed a humanoid robot able to reproduce human expressions and their associated human hand movements. In the marketing field, Weinberg and Gottwald (1982) have investigated human behavior characterizing impulse purchases. Emotions play a key role and facial expressions appeared to be one of their main indicators. Small and Verrochi (2009) studied how the victim faces displayed on advertisements for charities

affect both sympathy and giving.

The measuring of user emotions has become an important research topic in transportation behavior analysis. For instance, it may be used to analyse travelers satisfaction in public transportation. In the car context, it may allow to adapt the vehicle functionalities to the driver's mood for both well-being and safety reasons. Reimer et al. (2009) develop the concept of "awareness" of the vehicle in order to improve the mobility, performance and safety of older drivers. Information about driver general states, such as respiration, facial expression or concentration, are crucial to correctly apprehend the immediate driver capabilities and adapt the vehicle behavior to it. Moreover, some car manufacturers are currently working on the driver's mood recognition in order to warn the driver about possible dangers generated by other users. This aims at preventing road rages. Currently, the mood recognition is based only on the driver's voice. Facial expression recognition can also be used as a complementary source of information to determine the driver's mood. For routine trips, Abou-Zeid (2009) conducts experiments to measure the travel well-being for both public transportation and car modes. Collected data were employed to estimate mode choice models. Well-being measures are used as utility indicators, in addition to standard choice indicators. A system of facial expression recognition could be coupled to such models, in order to better capture the commuter emotional states. Another obvious application is security, for example in airports or train stations. More generally, the DFER models could be used in any human-machine interface.

In this paper, we propose the use of discrete choice models (DCM) as they are designed to describe the behavior of people in choice situations. We can consider a decision-maker who has to label a video sequence by choosing among a list of facial expressions. The list is composed of the seven basic expressions described by Keltner (2000): happiness, surprise, fear, disgust, sadness, anger, neutral. We have also added "Other" and "I don't know" , to avoid ambiguities. In the following, the expressions are respectively denoted by H, SU, F, D, SA, A, N, DK and O.

Contrarily to computer vision algorithms which are calibrated using a ground truth, our models are estimated using behavioral data. Computer vision algorithms can be often considered as a "black box", as their parameters are difficult to interpret. In our case, a specification is proposed where causal links between facial characteristics and expressions are explicitly modeled. The output of the model is a probability distribution among expressions. We have successfully applied the approach for SFER ((Sorci, Antonini, Cruz, Robin, Bierlaire and Thiran, 2010), and (Sorci, Robin, Cruz, Bierlaire, Thiran and Antonini, 2010)). We propose a logit model, with nine alternatives corresponding to the nine items cited above. Each utility is a function

of measures related to the AU associated to the expression, as defined by the FACS. Sorci, Antonini, Cruz, Robin, Bierlaire and Thiran (2010) have also introduced the concept of expression descriptive units (EDU), that capture interactions between AU. Moreover, some outputs of the computer vision algorithm used to extract measures on facial images, are also included in the utility, in order to account for the global facial perception.

The DFER does not fit into the usual discrete choice applications, so adjustments have to be done. We took inspiration from the work of Choudhury (2007) who uses a dynamic behavioral framework to model car lane changing. Three models are presented in this analysis. Different modeling assumptions have been tested and compared. We first present the behavioral data used to estimate the models. Then the specification of the proposed models and the estimation results are presented. We finally describe the validation and the applications of the models.

2 Data

The data is derived from a set of video sequences from the facial expressions and emotions database (FEED) collected by Wallhoff (2004). They have recorded students watching television. Different types of TV programs are presented to the subjects in order to generate a large spectrum of expressions. The database contains 95 sequences from 18 subjects. The collected videos last between 3 and 6 seconds. In each video, the subject starts with a neutral face (see example in Figure 1). Then, at some point the TV program triggers an expression.

We have selected 65 videos from 17 subjects. The videos of subject $\mathsf{N^o17}$ were removed because of the lack of variability in facial characteristics, or due to some discontinuities in the recording of the videos. The number of considered videos per subject is shown in Figure 3. We have no access to the type of expression that was meant to be triggered during the experiment.

A video is a sequence of images. For each image, numerical data are extracted using an active appearance model (AAM, (Cootes et al., 2002)). It permits to extract facial distances and angles as well as facial texture information (such as levels of gray) from each image. This technique is based on several principal component analysis (AAM) performed on the image treated as an array of pixel values. The algorithm tracks a facial mask composed of 55 points (see Figure 4) used to measure various facial distances and angles. Another vector C of values capturing both the facial texture and shape is also generated by the PCA. A total of 88 variables capturing distances (number of pixels) and angles (radians), as well as 100 elements of the vector C, have

Figure 1: Snapshot of a FEED database video: neutral face (subject N°2)

Figure 2: Snapshot of a FEED database video: expression produced by the TV program (subject N°2)

been generated for each image in each video.

The video is discretized in groups of 25 images, each corresponding to one second of the video, *i.e.* the number of groups of images is equal to the duration in seconds of the video. The features associated with each group of

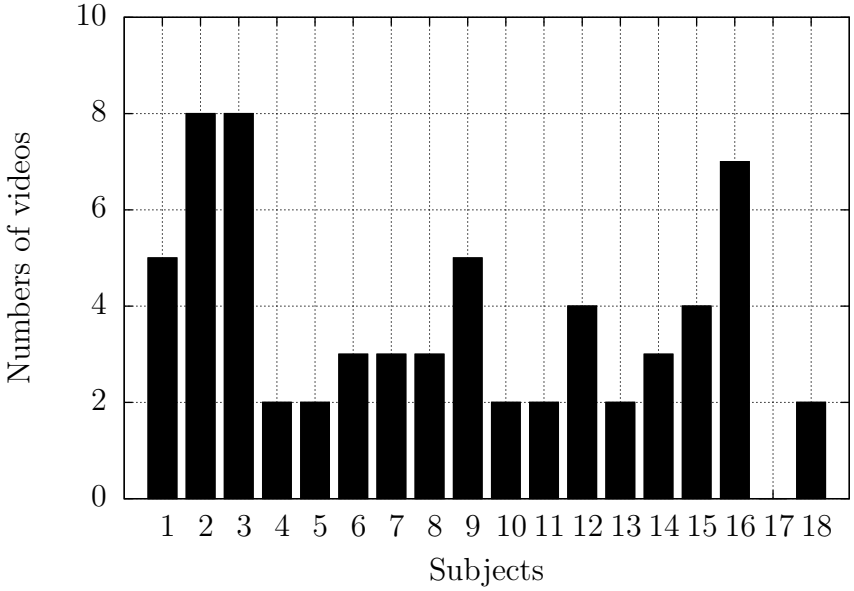

Figure 3: Numbers of considered videos per subject

Figure 4: Mask tracked by AAM along a video sequence

images are the features of the first image of the group. In the following, we use "frame" to refer to what is actually the first image of a group. The features of the 24 remaining images are used to compute variances (see Equation (2)).

For a given frame t and video o, three sets of variables are introduced: $\{x_{k,t,o}\}_{k=1,...,188}$, $\{y_{k,t,o}\}_{k=1,...,188}$, $\{z_{k,t,o}\}_{k=1,...,188}$. $\{x_{k,t,o}\}_{k=1,...,188}$ are the features extracted using the AAM. A complete description of these facial measurements is presented by Sorci, Antonini, Cruz, Robin, Bierlaire and Thiran (2010). In order to characterize the frame dynamics, some other variables are

calculated. For each variable $x_{k,t,o}$, $k = 1, \ldots, 188$, we introduce the variable $y_{k,t,o}$ defined as

$$y_{k,t,o} = x_{k,t,o} - x_{k,t-1,o} \text{ for } t = 2, \ldots, T_o, \quad (1)$$

where T_o is the number of frames in the video o. As each frame corresponds to one second, $y_{k,t,o}$ can be interpreted as the first derivative of $x_{k,t,o}$ with respect to time, approximated by finite difference. It quantifies the level of variation of the facial characteristic between two consecutive frames. Moreover, another variable $z_{k,t,o}$ is introduced for each $x_{k,t,o}$, $k = 1, \ldots, 188$, and is defined as

$$z_{k,t,o} = Var(x_{k,t,o}). \quad (2)$$

It is the variance of the features calculated over the 25 images preceding the frame t. It characterizes the short time variations of the facial characteristic $x_{k,t,o}$. For logical reasons, we have fixed

$$y_{k,1,o} = z_{k,1,o} = 0 \ \forall k, o \ , \quad (3)$$

meaning that the derivative and the variance of a variable in the first frame of all videos, is fixed to 0. We have a database of 564 ($= 188 \times 3$) variables for each frame t in each video o. The variables have been normalized in the interval $[-1, 1]$, in order to harmonize their scale: each variable has been divided by the maximum in absolute value between its observed maximum and its observed minimum over all frames and videos.

An internet survey has been conducted in order to obtain labels of FEED videos. It is available at *http://transp-or2.epfl.ch/videosurvey/* since august 2008. During the first session, respondents are asked to create an account and fill a socio-economic form. Once the account is created, they have to decide how many facial videos they want to label (5, 10 or 20). Videos are extracted randomly from the database. Then the expression labelling process can start. A screen snapshot is shown at Figure 5.

For this analysis, we have collected 369 labels from 40 respondents. The repartition of the observations among the expressions is displayed in Figure 6.

3 Models specification

The model proposed by Sorci, Antonini, Cruz, Robin, Bierlaire and Thiran (2010) is called **static model**. In this analysis, one model is developed. We suppose that the perception of the respondent starts at the first frame of the

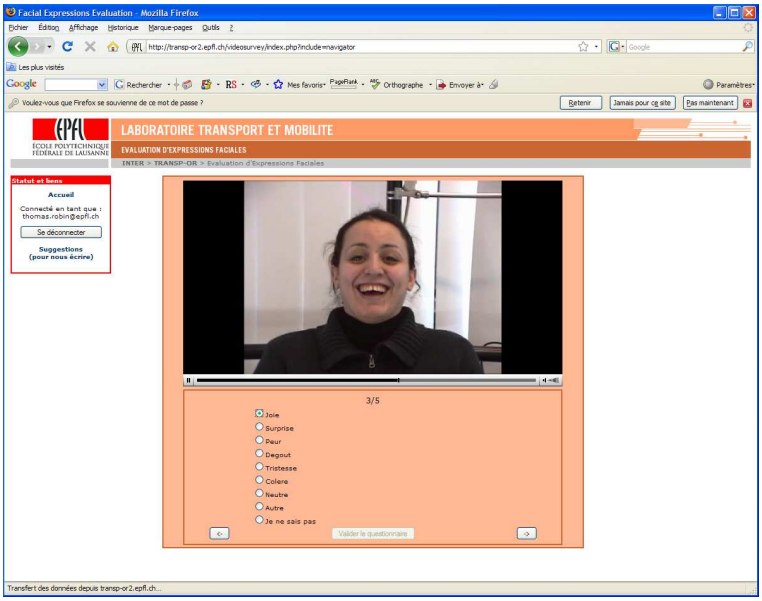

Figure 5: Snapshot of internet survey screen (subject N°15)

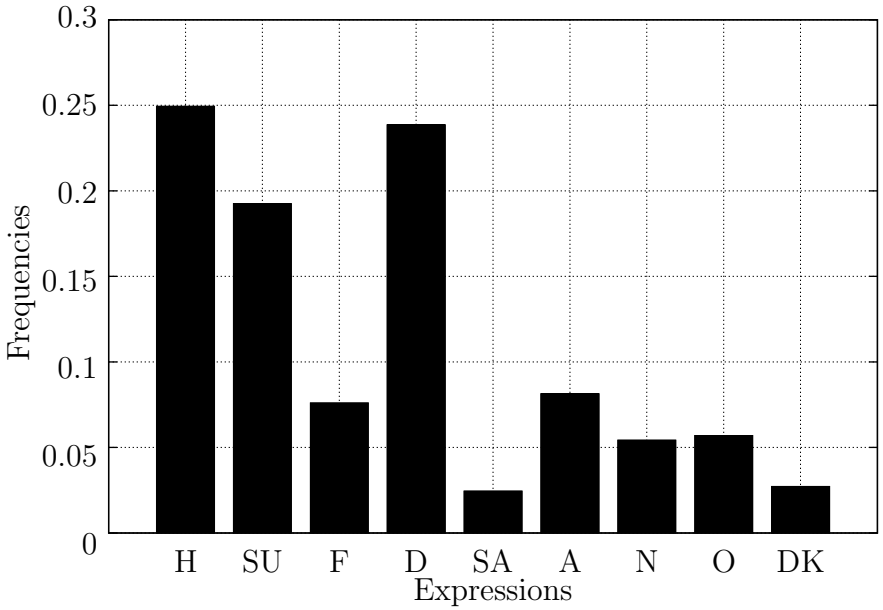

Figure 6: Distribution of the collected labels among the expressions

video. Then, we assume that the respondent updates her perception every second, which corresponds to every frame (see Section 2). We suppose that only the most impressive frame is influential on the choice of label. The

model is called **latent model**. Due to the small number of respondents, their characteristics have not been included in the model.

The model consists of a combination of two models. The first model quantifies the perception of expressions in a given frame. The second model predicts which frame has influenced the chosen label. It is a latent choice model where the choice set is composed of all frames in the video. The instantaneous perception of expressions and the most influential frame are not observed. Only the final choice of label for the video is observed.

The first model provides the probability for a respondent to choose the expression i when exposed to the frame t of the video sequence o, and is written $P_{M_2}(i|t, o, \theta_{M_2,1}, \alpha)$. The second model provides the probability for the frame t of video o to trigger the choice, and is denoted by $P_{M_2}(t|o, \theta_{M_2,2})$. The probability for a respondent to label the video o with expression i, is denoted by $P_{M_2}(i|o, \theta_{M_2}, \alpha)$, which is observable. $\theta_{M_2,1}$ and $\theta_{M_2,2}$ are the vectors of unknown parameters to be estimated, merged into the vector θ_{M_2}. α is a vector of parameters capturing the memory effects, which will be introduced in Equation (7), and has to be estimated ($\alpha = \{\alpha_i\}_{i=H,SU,F,D,SA,A,O}$). We obtain

$$P_{M_2}(i|o, \theta_{M_2}, \alpha) = \sum_{t=1}^{T_o} P_{M_2}(i|t, o, \theta_{M_2,1}, \alpha) P_{M_2}(t|o, \theta_{M_2,2}). \qquad (4)$$

For specifying the model $P_{M_2}(i|t, o, \theta_{M_2,1}, \alpha)$, we need to define a utility function associated to each expression. We hypothesize that the perception of an expression i in frame t depends on the instantaneous perceptions of this expression i in the frames t and $t-1$. $V_{M_2}(i|t, o, \theta_{M_2,1}, \alpha_i)$ is a utility reflecting the perception of the expression i in frame t for the video o. We decompose it into two parts. First $V^s_{M_2}(i|t, o, \theta_{M_2,1})$ concerns the instantaneous perception of the frame t in the video o. Second, $V^s_{M_2}(i|t-1, o, \theta_{M_2,1})$ concerns the instantaneous perception of the frame $t-1$ in the video o. This is designed to capture the dynamic nature of the decision making process, as illustrated in Figure 7. In this figure, the facial measurements $\{x_{k,t,o}\}$ and $\{z_{k,t,o}\}$ (introduced in Equation (2)) are observed, they are enclosed in rectangles and their influences are represented by plain arrows; whereas the utilities are latent, they are enclosed in ellipses and their influences are marked by dashed arrows. $\{x_{k,t,o}\}$ and $\{z_{k,t,o}\}$ influence $V^s_{M_2}(i|t, o, \theta_{M_2,1})$, while $V_{M_2}(i|t, o, \theta_{M_2,1}, \alpha_i)$ is only function of $V^s_{M_2}(i|t, o, \theta_{M_2,1})$ and $V^s_{M_2}(i|t-1, o, \theta_{M_2,1})$.

The specification of $\{V^s_{M_2}(i|t, o, \theta_{M_2,1})\}$ is presented in Equation (5)

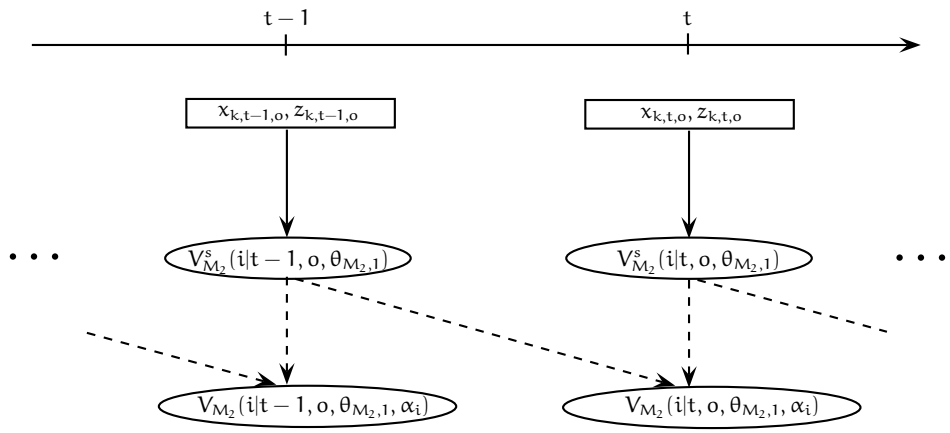

Figure 7: The dynamic process of **latent model**

$$
\begin{aligned}
V^s_{M_2}(H|t,o,\theta_{M_2,1}) &= ASC_H + \sum_{j=1}^{K_{M_2}} I_{M_2,1,H,j}\theta_{M_2,1,j} \sum_{k=1}^{188} I_{M_2,j,k}x_{k,t,o} \ , \\
V^s_{M_2}(SU|t,o,\theta_{M_2,1}) &= ASC_{SU} + \sum_{j=1}^{K_{M_2}} I_{M_2,1,SU,j}\theta_{M_2,1,j} \sum_{k=1}^{188} I_{M_2,j,k}x_{k,t,o} \\
&+ \sum_{j=1}^{K^z_{M_2}} I^z_{M_2,SU,j}\theta^z_{M_2,1,j} \sum_{k=1}^{188} I^z_{M_2,j,k}z_{k,t,o} \ , \\
V^s_{M_2}(F|t,o,\theta_{M_2,1}) &= ASC_F + \sum_{j=1}^{K_{M_2}} I_{M_2,F,j}\theta_{M_2,1,j} \sum_{k=1}^{188} I_{M_2,j,k}x_{k,t,o} \ , \\
V^s_{M_2}(D|t,o,\theta_{M_2,1}) &= ASC_D + \sum_{j=1}^{K_{M_2}} I_{M_2,D,j}\theta_{M_2,1,j} \sum_{k=1}^{188} I_{M_2,j,k}x_{k,t,o} \ , \\
V^s_{M_2}(SA|t,o,\theta_{M_2,1}) &= ASC_{SA} + \sum_{j=1}^{K_{M_2}} I_{M_2,SA,j}\theta_{M_2,1,j} \sum_{k=1}^{188} I_{M_2,j,k}x_{k,t,o} \ , \\
V^s_{M_2}(A|t,o,\theta_{M_2,1}) &= ASC_A + \sum_{j=1}^{K_{M_2}} I_{M_2,A,j}\theta_{M_2,1,j} \sum_{k=1}^{188} I_{M_2,j,k}x_{k,t,o} \ , \\
V^s_{M_2}(N|t,o,\theta_{M_2,1}) &= 0 \ , \\
V^s_{M_2}(O|t,o,\theta_{M_2,1}) &= ASC_O + \sum_{j=1}^{K_{M_2}} I_{M_2,O,j}\theta_{M_2,1,j} \sum_{k=1}^{188} I_{M_2,j,k}x_{k,t,o} \ , \\
V^s_{M_2}(O|t,o,\theta_{M_2,1}) &= ASC_{DK} \ , \qquad (5)
\end{aligned}
$$

where K_{M_2} is the total number of parameters related to $\{x_{k,t,o}\}$. $K^z_{M_2}$ is the total number of parameters related to $\{z_{k,t,o}\}$. $I_{M_2,i,j}$ is an indicator equal to 1 if the parameter j is included in the utility of expression i, 0 otherwise. $I_{M_2,j,k}$ is an indicator equal to 1 if the parameter j is related to the facial measurement $x_{k,t,o}$ collected in the frame t of the video o, 0 otherwise. We have

$$\sum_{k=1}^{188} I_{M_2,j,k} = 1 \ \forall j \ , \tag{6}$$

meaning that a parameter $\theta_{M_2,j}$ is related to only one $x_{k,t,o}$. $I^z_{M_2,su,j}$ and $I^z_{M_2,j,k}$ have exactly the same role as $I_{M_2,i,j}$ and $I_{M_2,j,k}$, but they concern the parameter $\theta^z_{M_2,j}$ which is related to $z_{k,t,o}$. Each utility contains a constant, except for the neutral expression. Its utility is the reference and is fixed to 0. The presence of $\{z_{k,t,o}\}$ (short time variations of facial characteristics) in the surprise utility accounts for the perception of suddenness. $\{z_{kto}\}$ are better than $\{y_{k,t,o}\}$ in this case, because they capture faster variations of facial characteristics. This does not lead necessarily to the surprise facial expression, but according to the collected data (see Section 2), fast variations of facial characteristics could be perceived as surprise by respondents. The detailed specification of $\{V^s_{M_2}(i|t,o,\theta_{M_2,1})\}$ is described in Tables 3 and 4. For each parameter $\theta_{M_2,1,j}$, if $I_{M_2,i,j}$ is equal to 1, there is a "$\times$ in the column of the corresponding expression i. If $I_{M_1,j,k}$ is equal to 1, the relative facial characteristic $x_{k,t,o}$ is indicated.

Sorci, Antonini, Cruz, Robin, Bierlaire and Thiran (2010) employed the database proposed by T.Kanade (2000) when collecting behavioral data. The estimated parameters of the **static model** cannot be used directly in our analysis due to problems of facial position and scale between this database and the FEED (see Section 2). The filmed subjects are further from the camera in the FEED, compared to the Cohn-Kanade. Consequently, the model has to be re-estimated. In addition, the specifications of the utilities have been adapted to this analysis because of the lower number of data available. We use 369 observations of labels against 38110 for the work of Sorci, Antonini, Cruz, Robin, Bierlaire and Thiran (2010). This implies the estimation of a lower number of parameters: the utility specifications have been simplified and parameters have been grouped together regarding their sign and interpretability.

The utility function $V_{M_2}(i|t,o,\theta_{M_2,1},\alpha_i)$ is supposed to be the sum of $V^s_{M_2}(i|t,o,\theta_{M_2,1})$ and $\{V^s_{M_2}(i|t-1,o,\theta_{M_2,1})$ weighted by α_i, the parameter of memory effect. The specification of $V_{M_2}(i|t,o,\theta_{M_2,1},\alpha_i)$ is defined in Equation (7).

$$
\begin{aligned}
V_{M_2}(H|t,o,\theta_{M_2,1},\alpha_H) &= V^s_{M_2}(H|t,o,\theta_{M_2,1}) \\
&+ \alpha_H V^s_{M_2}(H|t-1,o,\theta_{M_2,1}), \\
V_{M_2}(SU|t,o,\theta_{M_2,1},\alpha_{SU}) &= V^s_{M_2}(SU|t,o,\theta_{M_2,1}), \\
V_{M_2}(F|t,o,\theta_{M_2,1},\alpha_F) &= V^s_{M_2}(F|t,o,\theta_{M_2,1}) \\
&+ \alpha_F V^s_{M_2}(F|t-1,o,\theta_{M_2,1}), \\
V_{M_2}(D|t,o,\theta_{M_2,1},\alpha_D) &= V^s_{M_2}(D|t,o,\theta_{M_2,1}), \\
V_{M_2}(SA|t,o,\theta_{M_2,1},\alpha_{SA}) &= V^s_{M_2}(SA|t,o,\theta_{M_2,1}) \\
&+ \alpha_{SA} V^s_{M_2}(SA|t,o,\theta_{M_2,1}), \\
V_{M_2}(A|t,o,\theta_{M_2,1},\alpha_A) &= V^s_{M_2}(A|t,o,\theta_{M_2,1}), \\
V_{M_2}(N|t,o,\theta_{M_2,1},\alpha_N) &= V^s_{M_2}(N|t,o,\theta_{M_2,1}) = 0, \\
V_{M_2}(O|t,o,\theta_{M_2,1},\alpha_O) &= V^s_{M_2}(O|t,o,\theta_{M_2,1}) \\
&+ \alpha_O V^s_{M_2}(O|t,o,\theta_{M_2,1}), \\
V_{M_2}(DK|t,o,\theta_{M_2,1},\alpha_{DK}) &= V^s_{M_2}(DK|t,o,\theta_{M_2,1}). \qquad (7)
\end{aligned}
$$

Note that this is not anymore a linear-in-parameter specification for happiness, fear, sadness and anger, since $\{\alpha_i\}$ are estimated. Five memory effects parameters $\{\alpha_i\}_{i=SU,D,A,N,DK}$ have been fixed to zero : for neutral because it is the referent alternative, so its utility is fixed to zero; and for "I don't know" because its utility contains only ASC_{DK}, which is invariant across the frames. For surprise, disgust and anger, they do not appeared to be significant in previous specifications of the model (see Section 4 and Table 5). $\{\alpha_i\}_{i=H,F,SA,O}$ are supposed to be in the interval $[-1,1]$ because we hypothesize that the instantaneous perception of expression i in the previous frame $t-1$ has less influence than the instantaneous perception of expression i in the frame t, on the perception of expression i at time t. The model for $P_{M_2}(i|t,o,\theta_{M_2,1},\alpha)$ is a logit model, that is

$$
P_{M_2}(i|t,o,\theta_{M_2,1},\alpha_i) = \frac{e^{V_{M_2}(i|t,o,\theta_{M_2,1},\alpha_i)}}{\sum_j e^{V_{M_2}(j|t,o,\theta_{M_2,1},\alpha_j)}}. \qquad (8)
$$

The model $P_{M_2}(t|o,\theta_{M_2,2})$ is also specified as a logit model. Note that we decide to ignore here the potential correlation between error terms of successive frames. A utility $V_{M_2}(t|o,\theta_{M_2,2})$ is associated to each frame t in the video o. The utility depends on variables $\{y_{k,t,o}\}$ which capture the levels of variation of the facial measurements between two consecutive frames (see Equation (1)), and $\{z_{k,t,o}\}$ which capture the short time changes of the facial measurements (see Equation (2)). We define $V_{M_2}(1|o,\theta_{M_2,2}) = 0$ and, for $t = 2,\ldots,T_o$,

$$\begin{aligned} V_{M_2}(t|o,\theta_{M_2,2}) &= \sum_{j=1}^{K^y_{M_2,2}} \theta^y_{M_2,2,j} \sum_{k=1}^{188} I^y_{M_2,2,j,k} y_{k,t,o} \\ &+ \sum_{j=1}^{K^z_{M_2,2}} \theta^z_{M_2,2,j} \sum_{k=1}^{188} I^z_{M_2,2,j,k} z_{k,t,o} \ , \end{aligned} \tag{9}$$

and

$$P_{M_2}(t|o;\theta_{M_2,2}) = \frac{e^{V_{M_2}(t|o,\theta_{M_2,2})}}{\sum_{\ell=1}^{T_o} e^{V_{M_2}(\ell|o,\theta_{M_2,2})}}. \tag{10}$$

$K^y_{M_2,2}$ and $K^z_{M_2,2}$ are the numbers of parameters associated to $\{y_{k,t,o}\}$, and $\{z_{k,t,o}\}$ respectively, in the utility related to each frame. $I^y_{M_2,2,j,k}$ is an indicator equal to 1 if the parameter $\theta^y_{M_2,2,j}$ is associated to $y_{k,t,o}$, 0 otherwise. As for the other indicators, it is related to only one $y_{k,t,o}$, we have

$$\sum_{k=1}^{188} I^y_{M_2,2,j,k} = 1 \ \forall j \ , \tag{11}$$

$I^z_{M_2,2,j,k}$ is similar to $I^y_{M_2,2,j,k}$, but is associated to $z_{k,t,o}$. The vector of parameters $\theta_{M_2,2}$ is described in Table 6 (same reading as for Table 3). Finally, the log-likelihood function is

$$\begin{aligned} \mathcal{L}(\theta_{M_2},\alpha) &= \sum_{o=1}^{O} \sum_{i=1}^{9} w_{i,o} \log P_{M_2}(i|o,\theta_{M_2},\alpha) \\ &= \sum_{o=1}^{O} \sum_{i=1}^{9} w_{i,o} \log\Big(\sum_{t=1}^{T_o} P_{M_2}(i|t,o,\theta_{M_2,1},\alpha_i) P_{M_2}(t|o,\theta_{M_2,2})\Big). \end{aligned} \tag{12}$$

4 Estimations of the models

The model is estimated by maximum likelihood (see Equation (12)) with codes based on the BIOGEME software developed by Bierlaire (2003) to do simultaneous estimation. Estimation results are presented in Table 1.

The values and associated t-tests of the 34 parameters related to the model handling with the perception of the expressions are presented in Tables 3 and 4. Signs and significance of parameters related to AU, EDU and elements of the vector C are correct and consistent with the estimated parameters obtained by Sorci, Antonini, Cruz, Robin, Bierlaire and Thiran (2010).

In addition, the model contains one more parameter. $\theta^z_{M_2,1,1}$ is related to the variance of the height of the mouth ("*mouth_h*"). It is positive meaning that the more variations in the height of the mouth there are within the previous second, the more the surprise will be favored, which is logical. Four parameters of memory effect ($\alpha_H, \alpha_F, \alpha_{SA}, \alpha_O$) appear to be significantly different from zero (see Table 5). They have the same magnitude. Without any constraint, their estimated values are in $[-1, 1]$ meaning that the present perception is predominant, as expected. Seven parameters related to the model characterizing the influence of the frames are estimated significantly different from zero (see Table 6). Six are associated to $\{y_{k,t,o}\}$ and one to $z_{2,t,o}$ which is the variance of the distance between eyebrows ("*brow_dist*"). Their magnitude is larger than for the parameters associated to the model of perception of the expressions. This means that the model is sensitive to small variations of features and tends to produce a sharp probability distribution among the frames. The signs of the parameters are logical, for example $\theta_{M_2,2,5}$ is attached to the height of the eyes ("*eye_h*") and is negative. This means that the more a subject has the eye closed on a frame, the more the frame has influence on the observed choice of label.

	Latent model
Nb of observations	369
Nb of parameters	45
Null log-likelihood	−810.78
Final log-likelihood	−441.28
$\bar{\rho}^2$	0.400

Table 1: General estimation results

Another example is $\theta_{M_2,1,4}$ which is related to the opening of the mouth ("RAP_mouth"), defined as the fraction between the height of the mouth ("*mouth_h*") and the width of the mouth ("*mouth_w*"). It is present in the utilities of surprise and fear. The associated parameter is positive, which is logical because when a person has the mouth opened, the perceived facial expression is more likely to be fear or surprise.

The parameters values of the model related to the detection of the most impressive frame are also interpretable. For example, $y_{2,t,o}$ is related to the height of the mouth ("*mouth_h*"). Figure 9 displays the variation of this feature among frames of a video. The frames of the considered video are shown in Figure 8. The sign of the parameter associated to $y_{2,t,o}$ ($\theta_{M_2,2,6}$) is positive which is logical. The higher the difference of mouth height between two consecutive frames, the more important the second frame is. In that special

case and regarding only $\mathtt{y}_{2,t,o}$, frame 3 seems to be the most important.

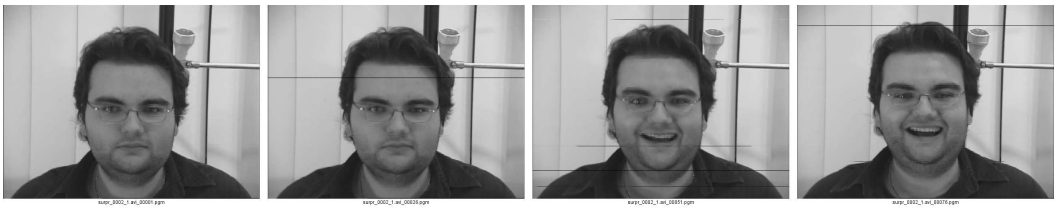

Figure 8: Frames of the considered video which is used for studying variations of $\mathtt{y}_{2,t,o}$, in Figure 9

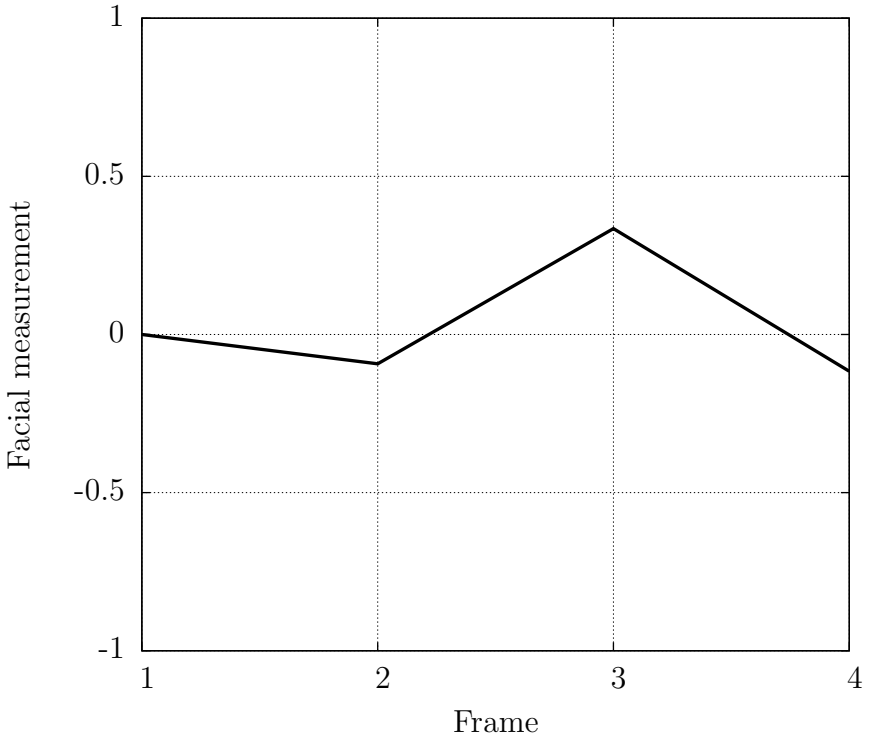

Figure 9: Variations of $\mathtt{y}_{2,t,o}$, related to the height of the mouth ("*mouth_h*") for the video presented in Figure 8

5 Prediction capability

The prediction capability is tested in order to ensure the quality of the model. The dataset used in this section is the same as the one used in Section 4.

We proceed in two steps: the first one consists of studying the percentage of badly predicted observations. In the second step, we study the predictions of the proposed model at a more disaggregated level. This consists of picking a certain video and analysing the predictions of the model in detail.

An observation is considered as badly predicted, if its forecasted choice probability is less than $\frac{1}{9}$, which corresponds to the probability predicted by a uniform probability on the number of alternatives. The percentage of badly predicted observations is displayed in Table 2.

Latent model
17.34

Table 2: Percentages of badly predicted observations on the estimation data

The cumulative distribution of the choice probabilities predicted by the model is displayed in Figure 10. If the model was perfect, the curve should be flat with a pick for choice probabilities equal to one. This would mean that the model replicate exactly the observed choices of labels. Of course this is not the case. However, the curve reaches slowly the threshold of one, which is a good feature.

We looked at the power of prediction over the estimation dataset, at the aggregate level. The study of a particular video allows to go in details of the predictions of the model. The video is the same than the one considered in Figure 8. The detailed predictions of the model is shown in Figure 11. On this figure, each column is related to a frame, except the extreme right. The first line displays the considered frames. As mentioned in Section 2, each frame is the first of a group of images corresponding to one second in a video. The second line concerns the predictions of the model associated to the perception of the expressions. For each frame, the probability distribution among the expressions is presented. The third line shows the influence of the frames. The contributions of the frames sum up to one.

On the first frame of the considered video (see Figure 11), the face tends to be neutral, and then evolves toward a different expression. Seven respondents have labeled this video: three gave the label happiness, three gave the label surprise, and one the label anger. Anger does not seem to be appropriate for this video, but it has been kept because there was no proof of mistakes made by the respondent. In addition the subject on the two first frames of the video could be considered angry. The observed distribution of the collected labels is displayed at the bottom right of the figure.

The model predicts 24% of happiness, 58% of surprise, 18% of disgust and 0% for anger. The model has selected frame 3 as being the most impressive

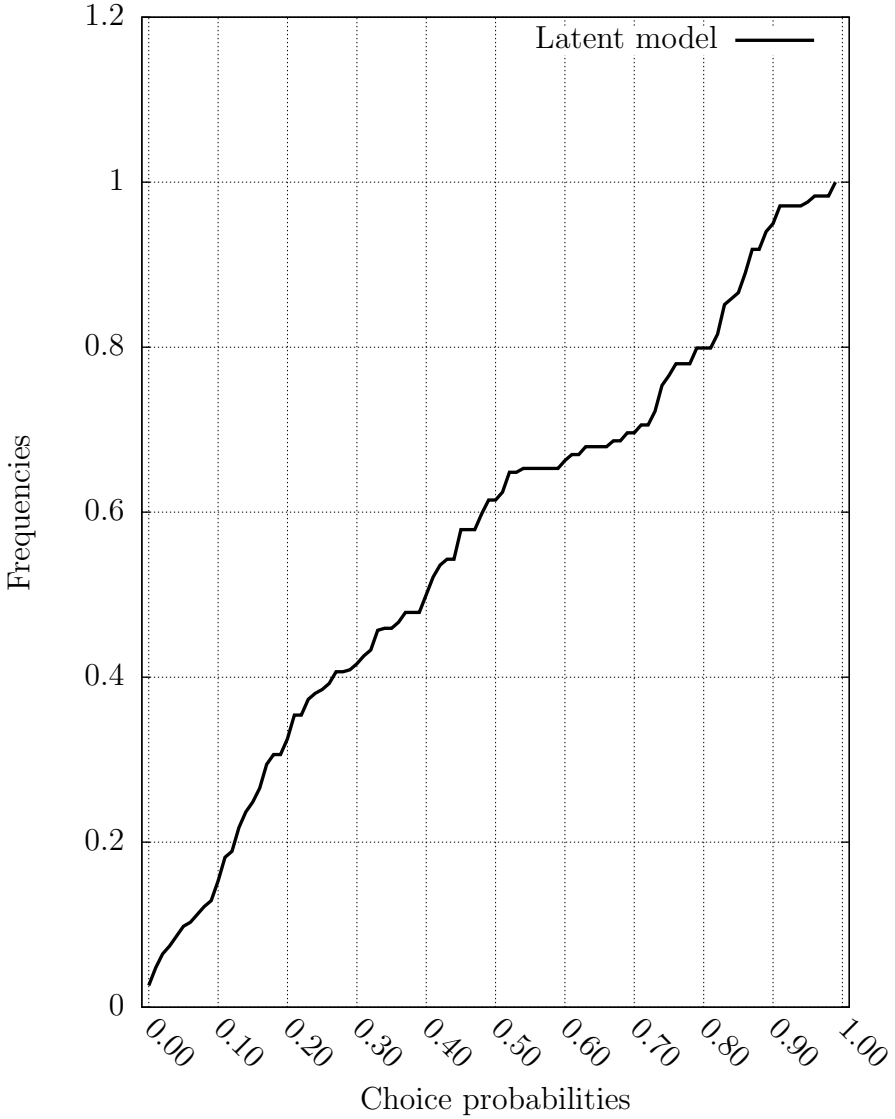

Figure 10: Cumulative distribution of the choice probabilities predicted by the proposed model, on the estimation data

frame, with a probability almost equal to one, so the predictions of the model results only from the perception of this frame. This is logical because the utilities of the frames contain both $\{y_{k,t,o}\}$ and $\{z_{k,t,o}\}$ (see Section 3), and they appear to be very high for frame 3 (see Figure 9 for the height of the mouth). For this frame, the predicted probability of surprise is very high. This is logical, because the utility of surprise contains $\{z_{k,t,o}\}$ (see Equation (5)), which account for the perception of suddeness. For this frame, the high

probability for happiness is also intuitive due to the facial characteristics. The prediction of disgust does not seem to be appriopriate.

6 Conclusions and Perspectives

We propose a new approach for the recognition of dynamic facial expressions. The estimation of the model is based on labels collected through respondents to an internet survey. The developed model capture up causal effects between facial characteristics and expressions. Statistical tests and model predictions have proved the quality of the model. Finally, some qualitative analysis of the model predictions allow to check the modeler's intuition about the facial video.

As such the model can be used directly for applications. The major difficulty concerns the computation of the variables. The quality of the considered videos should be very high, in terms of definition and size of the face. The applications in the field of transportation cited in the introduction could be considered. The videos of the FEED database are not dedicated to transportation (the stimuli used to generate the facial expressions of the subjects 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 the model applicability to transportation ((Denis, 2009)). 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 models. In a second time, we can dedicate the model to transportation, by estimating it on data related to the field. Instead of FEED videos, some facial videos of transportation users in special situations could be employed. The video collection could consist in acquiring some facial videos of drivers, when placed in simulators. Typical driving situations could be displayed as stimuli, to generate drivers expressions. Note that the experimental design of the video collection has to be closely linked to the application. Finally in the context of "Aware" vehicles, the proposed model could be incorporated in global emotion recognition systems, including other elements of recognition, such as the intonation of the voice or the concentration.

Even if this new modeling framework is meaningful, some improvements could be done. The model has been estimated on a small dataset. More observations would be useful. The number and type of videos is also a critical aspect, feature variabilities are quite low and should be increased. This

could allow to have more complete specifications. In addition, more complex structures could be tested for the choice models, such as MEV or mixtures of logit. This allows to account for correlation between alternatives. Moreover, the specificities of respondents could be taken into account in the model by specifying an error component capturing unobserved heterogeneity. A validation should be done on another dataset. Finally a comparison with a state of the art machine learning method, such as neural networks (NN) would be interesting.

Acknowledgments

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

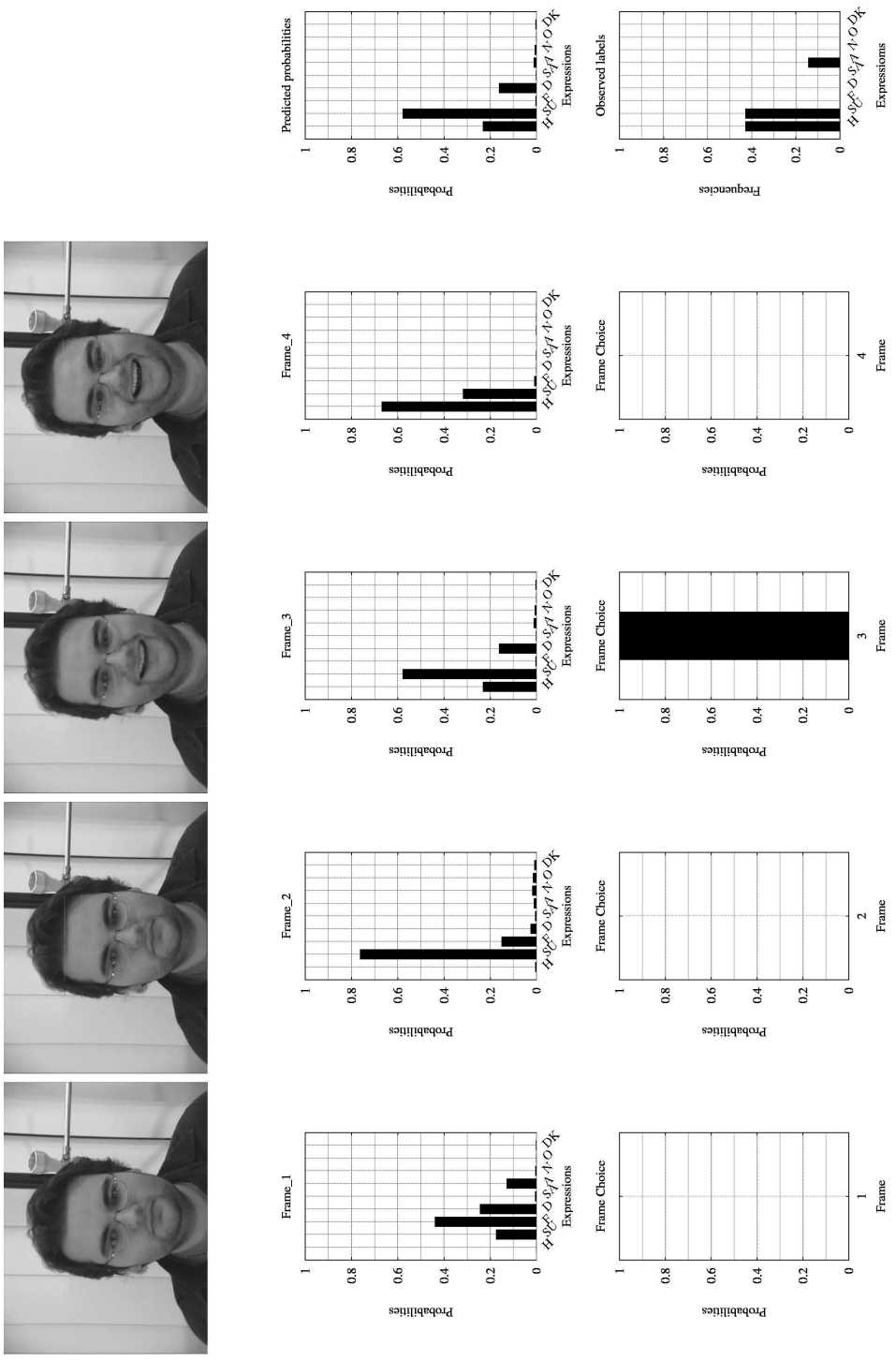

Figure 11: Example of detailed prediction of **latent model**

References

Abou-Zeid, M. (2009). *Measuring and Modeling Travel and Activity Well-Being*, PhD thesis, Massachusetts Institute of Technology.

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

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

Choudhury, C. F. (2007). *Model Driving Decisions with Latent Plans*, PhD thesis, Massachusetts institute of technology.

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

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

Denis, C. (2009). Facial expression recognition project: Collect a database, *Technical report*, Transport and Mobility Laboratory (TRANSP-OR), EPFL, EPFL ENAC INTER TRANSP-OR, Station 18, CH-1015 Lausanne, Switzerland.

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

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

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

Miwa, H., Itoh, K., Matsumoto, M., Zecca, M., Takanobu, S., Rocella, S., Carrozza, P., Dario, A. and A., T. (2004). Effective emotional expressions with emotion expression humanoid robot we-4rii - integration

of humanoid robot hand rch-1, *International Conference on Intelligent Robots and Systems*, Vol. 3, pp. 2203–2208.

Pantic, M. and Patras, I. (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.

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

Small, D. and Verrochi, N. (2009). The face of need: facial emotion expression on charity advertisements, *journal of marketing research* **XLVI**: 777 – 787.

Sorci, M., Antonini, G., Cruz, J., Robin, T., Bierlaire, M. and Thiran, J.-P. (2010). Modelling human perception of static facial expressions, *Image and Vision Computing* **28**(5): 790–806.

Sorci, M., Robin, T., Cruz, J., Bierlaire, M., Thiran, J.-P. and Antonini, G. (2010). Capturing human perception of facial expressions by discrete choice modelling, *in* S. Hess and A. Daly (eds), *Choice Modelling: The State-of-the-Art and the State-of-Practice*, Emerald Group Publishing Limited, pp. 101–136. ISBN:978-1-84950-772-1.

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

Tojo, T., Matsusaka, Y., Ishii, T. and Kobayashi, T. (2000). A conversational robot utilizing facial and body expressions, *Systems, Man, and Cybernetics, 2000 IEEE International Conference on*, Vol. 2, pp. 858–863.

Wallhoff, F. (2004). Fgnet-facial expression and emotion database, *Technical report*, Technische Universitt Mnchen.
URL: *http://www.mmk.ei.tum.de/ waf/fgnet/feedtum.html*

Weinberg, P. and Gottwald, W. (1982). Impulsive consumer buying as a result of emotions, *Journal of Business Research* **10**(1): 43 – 57.
URL: *http://www.sciencedirect.com/science/article/B6V7S-45JWVJH-2W/2/c5ad26cf95e71a37ca1cdbc072a7254b*

parameter	H	SU	F	D	SA	A	N	O	DK	$x_{k,t,o}$	value	t-test 0
ASC_A						×				1	-5.86	-1.31
ASC_D				×						1	22.73	4.48
ASC_{DK}									×	1	-0.71	-1.83
ASC_F			×							1	-4.55	-1.13
ASC_H	×									1	3.02	0.22
ASC_O								×		1	14.44	4.22
ASC_{SA}					×					1	8.54	1.57
ASC_{SU}		×								1	-25.69	-7.08

Table 3: Estimation results of the constants for **latent model**, associated the expression perception model

parameter	H	SU	F	D	SA	A	N	O	DK	$x_{k,t,o}$	value	t-test 0
$\theta_{M_2,1,1}$				×						EDU_6	-6.92	-3.37
$\theta_{M_2,1,2}$				×						EDU_8	-3.92	-5.42
$\theta_{M_2,1,3}$		×				×				RAP_brow	7.84	4.45
$\theta_{M_2,1,4}$		×	×							RAP_mouth	4.93	3.42
$\theta_{M_2,1,5}$	×									RAP_mouth	12.74	2.54
$\theta_{M_2,1,6}$	×									C_1	-38.18	-5.27
$\theta_{M_2,1,7}$						×				C_2	40.99	4.81
$\theta_{M_2,1,8}$				×						C_2	45.77	7.12
$\theta_{M_2,1,9}$	×									C_3	23.96	3.71
$\theta_{M_2,1,10}$		×								C_3	24.46	4.11
$\theta_{M_2,1,11}$					×					broweye_l2	240.75	4.11
$\theta_{M_2,1,12}$		×								broweye_l3	104.09	4.61
$\theta_{M_2,1,13}$		×	×	×	×	×				broweye_r2	-41.76	-2.93
$\theta_{M_2,1,14}$			×		×					eye_angle_l	44.95	2.58
$\theta_{M_2,1,15}$					×					eye_brow_angle_l	-199.01	-6.04
$\theta_{M_2,1,16}$				×						eye_mouth_dist_l2	-73.15	-2.72
$\theta_{M_2,1,17}$	×				×			×		eye_mouth_dist_l	-84.03	-3.83
$\theta_{M_2,1,18}$						×				eye_nose_dist_l	217.99	3.69
$\theta_{M_2,1,19}$			×	×	×			×		eye_nose_dist_l	80.02	2.09
$\theta_{M_2,1,20}$			×	×	×	×		×		eye_nose_dist_r	-211.73	-4.45
$\theta_{M_2,1,21}$		×	×							leye_h	51.35	4.12
$\theta_{M_2,1,22}$	×	×	×	×	×	×				mouth_h	98.27	3.27
$\theta_{M_2,1,23}$					×	×				mouth_nose_dist2	-92.34	-2.04
$\theta_{M_2,1,24}$	×									mouth_nose_dist	-412.5	-5
$\theta_{M_2,1,25}$	×									mouth_w	158.29	2.13

parameter	H	SU	F	D	SA	A	N	O	DK	$x_{k,t,o}$	value	t-test 0
$\theta^{z}_{M_2,1,1}$										mouth_h, $z_{1,t,o}$	50.21	3.04

Table 4: Estimation results and description of the specification of **latent model**, associated to the expression perception model

parameter	value	t-test 0
α_H	-0.62	-8.18
α_F	-0.33	-2.73
α_{SA}	-0.46	-2.04
α_O	-0.70	-2.68

Table 5: Estimation results of **latent model**, associated to the memory effects parameters

parameter	$y_{k,t,o}$	value	t-test 0
$\theta^{y}_{M_2,2,1}$	C_2	-426.75	-1.83
$\theta^{y}_{M_2,2,2}$	eye_brow_angle	350.53	1.7
$\theta^{y}_{M_2,2,3}$	mouth_w	407.34	1.76
$\theta^{y}_{M_2,2,4}$	C_4	463.35	1.75
$\theta^{y}_{M_2,2,5}$	eye_h	-566.62	-1.79
$\theta^{y}_{M_2,2,6}$	mouth_h	104.51	1.84
$\theta^{z}_{M_2,2,1}$	brow_dist, $z_{4,t,o}$	261.65	1.84

Table 6: Estimation results and description of the specification of **latent model**, associated to the model which detects the most meaningful frame