TRANSP-OR

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Dynamic facial expression recognition with a discrete choice model

Thomas Robin [*] Michel Bierlaire [*] Javier Cruz [*]

April 23, 2010

[*] Transport and Mobility Laboratory, Ecole Polytechnique Fédérale de Lausanne, CH-1015 Lausanne, Switzerland, {thomas.robin, michel.bierlaire, 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. Three models are proposed. The first assumes that only the last frame of the video triggers the choice of the expression. The second 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 third model is an extension of the second model. It assumes that the choice of the expression results from the average of expression perceptions within a group of frames. The models are estimated using videos from the Facial Expressions and Emotions Database (FEED). Labeling data on the videos has been obtained using an internet survey available at *http://transp-or2.epfl.ch/videosurvey/*. The prediction capability of the models is studied in order to check their validity. Finally the models are cross-validated using the estimation data.

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.

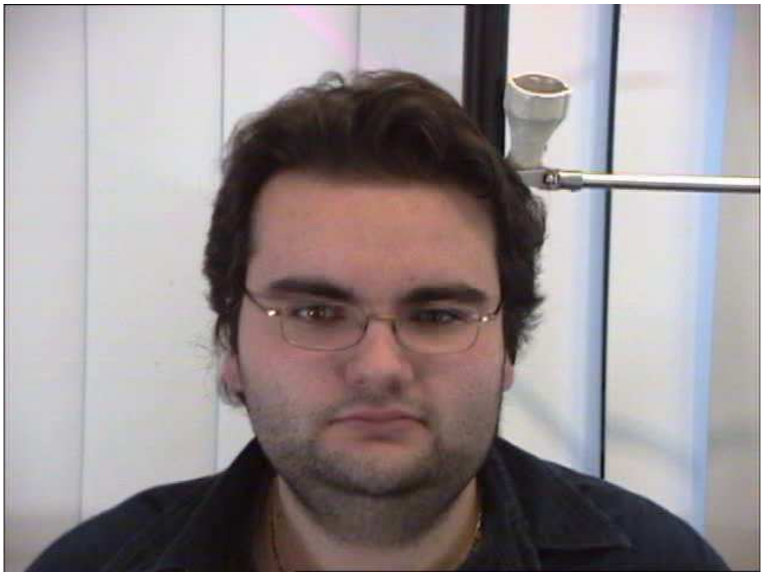


Figure 1: Snapshot of a FEED database video: neutral face (subject N°2)

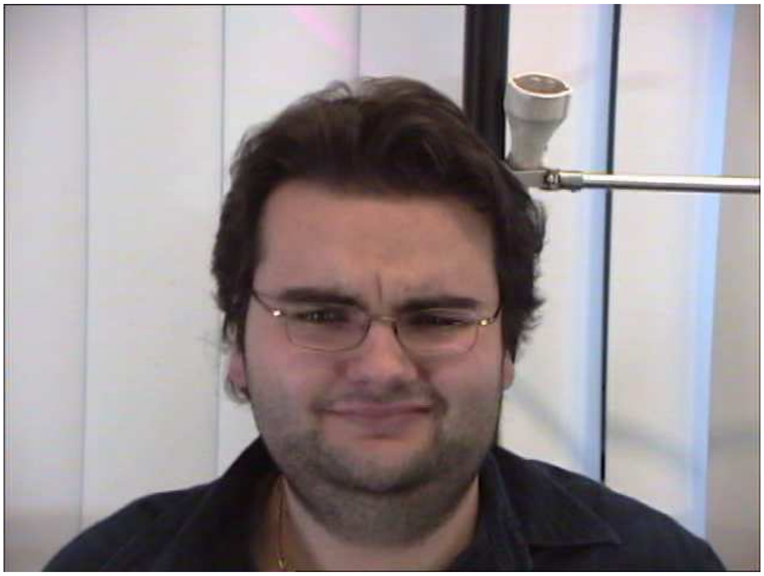


Figure 2: Snapshot of a FEED database video: expression produced by the TV program (subject N°2)

We have selected 65 videos from 17 subjects. The videos of subject N°17 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 (PCA) 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 AAM. A total of 88 variables capturing distances (number of pixels) and angles (radians), as well as 100 elements of the vector C, have 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

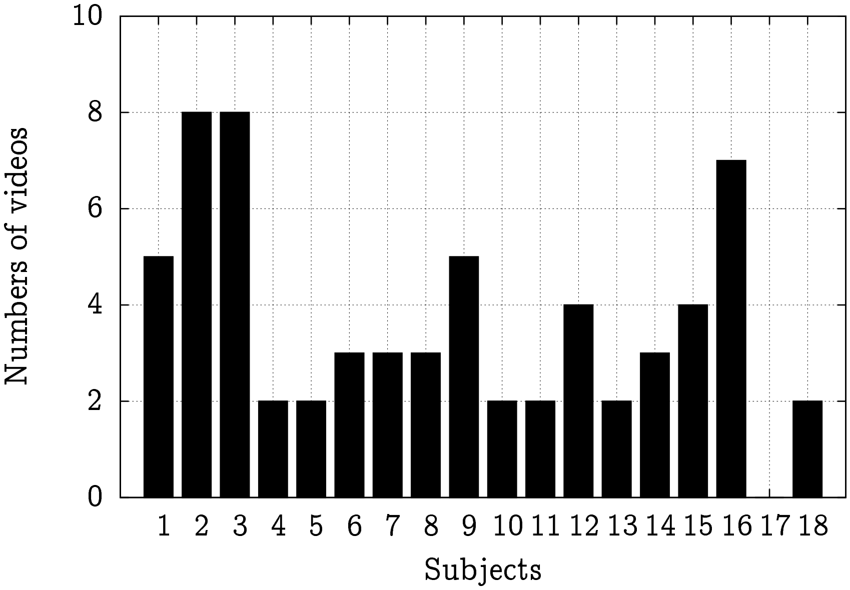


Figure 3: Numbers of considered videos per subject

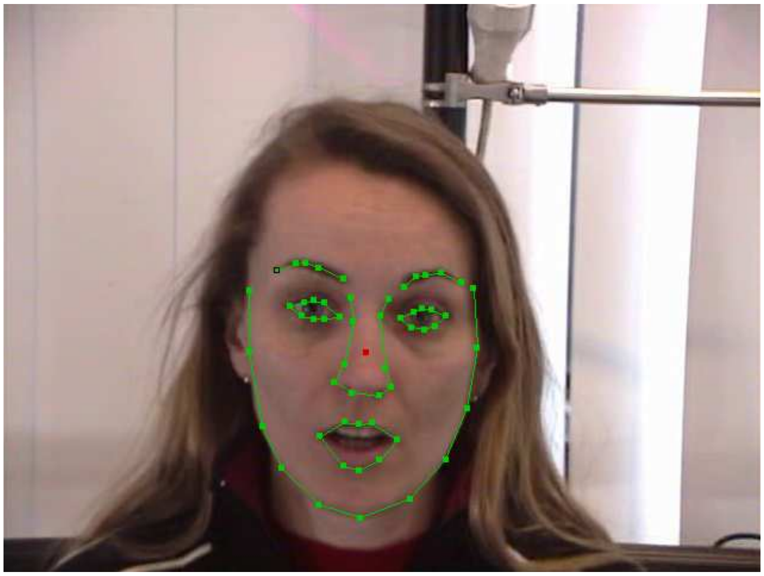


Figure 4: Mask tracked by AAM along a video sequence

of 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 $x_{k,t,o}$, $k = 1, \dots, 188$, we introduce the variable $y_{k,t,o}$ defined as

$$y_{k,t,o} = x_{k,t,o} - x_{k,t-1,o} \text{ for } t = 2, \dots, T_o, \tag{1}$$

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 differences. It quantifies the level of variation of the facial characteristics between two consecutive frames. Moreover, another variable $z_{k,t,o}$ is introduced for each $x_{k,t,o}$, $k = 1, \dots, 188$, and is defined as

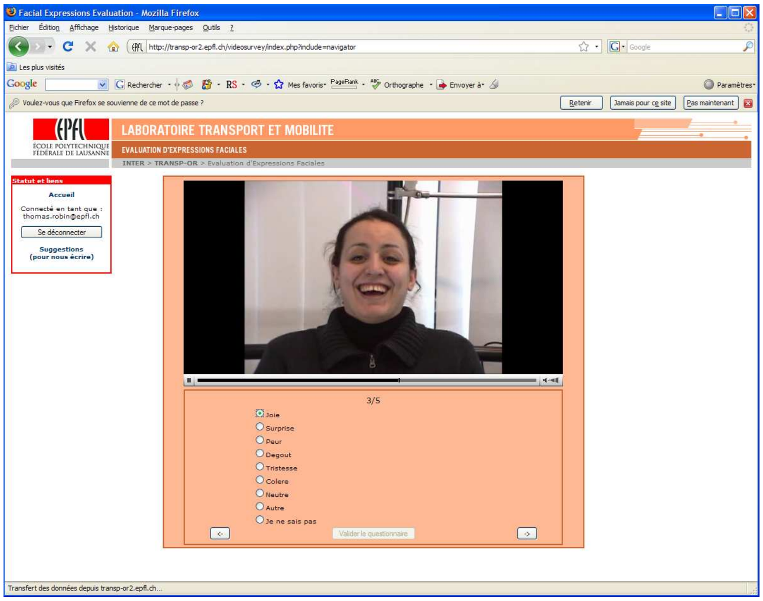
$$z_{k,t,o} = Var(x_{k,t,o}). \tag{2}$$

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

$$y_{k,1,o} = z_{k,1,o} = 0 \ \forall k, o \ , \tag{3}$$

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 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.



Facial Expressions Evaluation - Mozilla Firefox
Fichier Édition Affichage Historique Marque-pages Outils ?
http://transp-or2.epfl.ch/videosurvey/index.php?include=navigator
Google
Les plus visités
Google Rechercher RS Mes favoris PageRank Orthographe Envoyer à Paramètres
Retenir
Jamais pour ce site
Pas maintenant
ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE
LABORATOIRE TRANSPORT ET MOBILITE
EVALUATION D'EXPRESSIONS FACIALES
INTER > TRANSP-OR > Evaluation d'Expressions Faciales
Statut et liens
Accueil
Se déconnecter
Suggestions (pour nous écrire)
<-
Valider le questionnaire
->

Voulez-vous que Firefox se souvienne de ce mot de passe ?

Connecté en tant que : thomas.robin@epfl.ch

3/5

● Joie
○ Surprise
○ Peur
○ Degout
○ Tristesse
○ Colere
○ Neutre
○ Autre
○ Je ne sais pas

Transfert des données depuis transp-or2.epfl.ch...

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

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

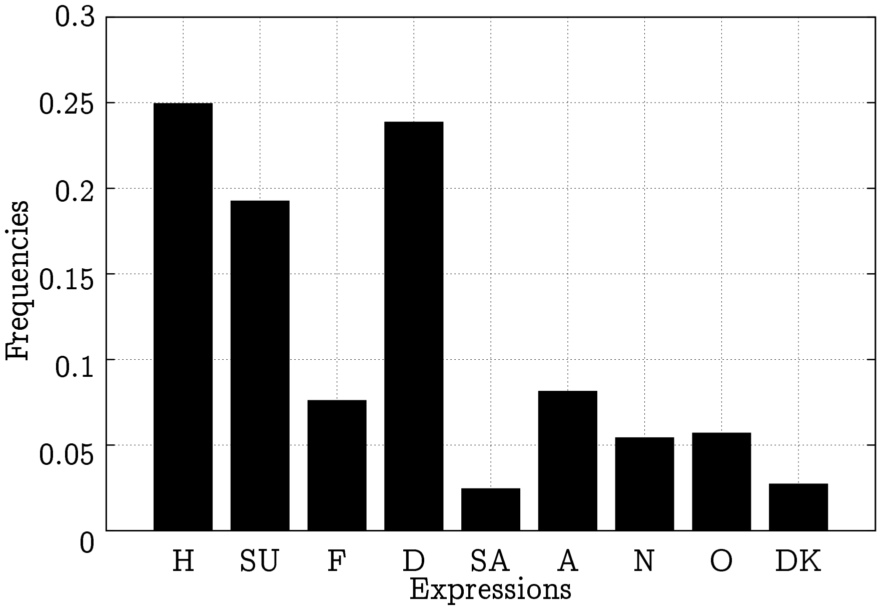


Figure 6: Distribution of the collected labels among the expressions

3 Models specification

The model proposed by Sorci, Antonini, Cruz, Robin, Bierlaire and Thiran (2010) is called **static model**. In this analysis, three models based on different assumptions have been developed. We suppose that the perception of the respondent starts at the first frame of the video. Then, we assume that the respondent updates her perception every second, which corresponds to every frame (see Section 2). In the first model we hypothesize that only the last frame of the video influences the observed choice of label. This is the simplest model presented in this analysis because it does not include dynamic aspects and it will be considered as a reference for comparison. This model is called **reduced model**. In the second model, only the most impressive frame is supposed to be influential on the choice of label. It is called **latent model**. Finally in the third model, we hypothesize that it is the average perception of a group of consecutive frames which generates the choice of label. This is called **smoothed model**. The theoretical details and specification of each model are described in Sections 3.1, 3.2 and 3.3. Due to the small number of respondents, their characteristics have not been included in the models.

3.1 The reduced model

In this model, only the perception of the last frame of a video is considered to be important for generating the observed choice of label. This assumption comes from the structure of a video. The filmed subject starts with a neutral face and evolves toward a certain expression which is triggered by the TV program that she is watching. Logically the subject's face on the last frame should be expressive. The model is a direct adaptation of **static model**.

The model associated to the perception of expressions is denoted by $P_{M_1}(i|o, \theta_{M_1})$. It is the probability for an individual to label the video o with the expression i, given the vector of unknown parameters θ_{M_1}. The last frame is supposed to be the only information used by the respondent to label the video o. The utility function associated with each expression is defined in Equation (4).

$$\begin{aligned}
V_{M_1}(H|T_o, o, \theta_{M_1}) &= ASC_H + \sum_{j=1}^{K_{M_1}} I_{M_1,H,j}\theta_{M_1,j} \sum_{k=1}^{188} I_{M_1,j,k}x_{k,T_o,o} \ , \\
V_{M_1}(SU|T_o, o, \theta_{M_1}) &= ASC_{SU} + \sum_{j=1}^{K_{M_1}} I_{M_1,SU,j}\theta_{M_1,j} \sum_{k=1}^{188} I_{M_1,j,k}x_{k,T_o,o} \ , \\
V_{M_1}(F|T_o, o, \theta_{M_1}) &= ASC_F + \sum_{j=1}^{K_{M_1}} I_{M_1,F,j}\theta_{M_1,j} \sum_{k=1}^{188} I_{M_1,j,k}x_{k,T_o,o} \ , \\
V_{M_1}(D|T_o, o, \theta_{M_1}) &= ASC_D + \sum_{j=1}^{K_{M_1}} I_{M_1,D,j}\theta_{M_1,j} \sum_{k=1}^{188} I_{M_1,j,k}x_{k,T_o,o} \ , \\
V_{M_1}(SA|T_o, o, \theta_{M_1}) &= ASC_{SA} + \sum_{j=1}^{K_{M_1}} I_{M_1,SA,j}\theta_{M_1,j} \sum_{k=1}^{188} I_{M_1,j,k}x_{k,T_o,o} \ , \\
V_{M_1}(A|T_o, o, \theta_{M_1}) &= ASC_A + \sum_{j=1}^{K_{M_1}} I_{M_1,A,j}\theta_{M_1,j} \sum_{k=1}^{188} I_{M_1,j,k}x_{k,T_o,o} \ , \\
V_{M_1}(N|T_o, o, \theta_{M_1}) &= 0 \ , \\
V_{M_1}(O|T_o, o, \theta_{M_1}) &= ASC_O + \sum_{j=1}^{K_{M_1}} I_{M_1,O,j}\theta_{M_1,j} \sum_{k=1}^{188} I_{M_1,j,k}x_{k,T_o,o} \ , \\
V_{M_1}(O|T_o, o, \theta_{M_1}) &= ASC_{DK} \ , \qquad (4)
\end{aligned}$$

where T_o denotes the length of the video o in seconds, which is also the index of the last frame of the video o. K_{M_1} is the total number of parameters related to facial measurements $\{x_{k,t,o}\}$ in **reduced model**. $I_{M_1,i,j}$ is an indicator equal to 1 if the parameter j is present in the utility of expression i, 0 otherwise. $I_{M_1,j,k}$ is an indicator equal to 1 if the parameter j is related to the facial measurement $x_{k,T_o,o}$ collected in the last frame of the video o, 0 otherwise. We have

$$\sum_{k=1}^{188} I_{M_1,j,k} = 1 \ \forall j \ , \qquad (5)$$

meaning that a parameter $\theta_{M_1,j}$ is related to only one facial measurement $x_{k,T_o,o}$. $\{x_{k,T_o,o}\}$ are introduced in Section 2. Each utility contains an alternative specific constant ASC_i except the neutral, which is taken as the reference, and its utility is fixed to 0. Note that there is no expression specific attributes, as the facial characteristics do not vary across the expressions. The details of the utility specifications are presented in Tables 4 and 5. For each parameter $\theta_{M_1,j}$, if $I_{M_1,i,j}$ is equal to 1, there is a "×" in the column of the corresponding expression i. If $I_{M_1,j,k}$ is equal to 1, the relative facial characteristic $x_{k,T_o,o}$ is indicated. The model is a logit, so the probability is

$$P_{M_1}(i|o,\theta_{M_1}) = \frac{e^{V_{M_1}(i|T_o,o,\theta_{M_1})}}{\sum_{j=1}^{9} e^{V_{M_1}(j|T_o,o,\theta_{M_1})}}. \tag{6}$$

Then the log-likelihood is

$$\mathcal{L}(\theta_{M_1}) = \sum_{o=1}^{O}\sum_{i=1}^{9} w_{i,o} \log(P_{M_1}(i|o,\theta_{M_1})), \tag{7}$$

where $w_{i,o}$ is a weight, corresponding to the number of times the expression i has been chosen for the video o in the collected database of annotations (see Section 2).

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 proposed model contains 32 parameters against 135 for the **static model**.

3.2 The latent model

The assumption supporting this model is that one frame in the video has influenced the observed choice of label, but the analyst does not know which one. The DFER model consists of a combination of two models.

The first model quantifies the perception of expressions in a given frame. It is similar to **reduced model** presented in Section 3.1. 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 (11), and has to be estimated ($\alpha = \{\alpha_i\}_{i=H,SU,F,D,SA,A,O}$). We obtain

$$P_{M_2}(i|o, \theta_{M_2}, \alpha) = \sum_{t=1}^{T_o} P_{M_2}(i|t, o, \theta_{M_2,1}, \alpha) P_{M_2}(t|o, \theta_{M_2,2}). \tag{8}$$

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 (9)

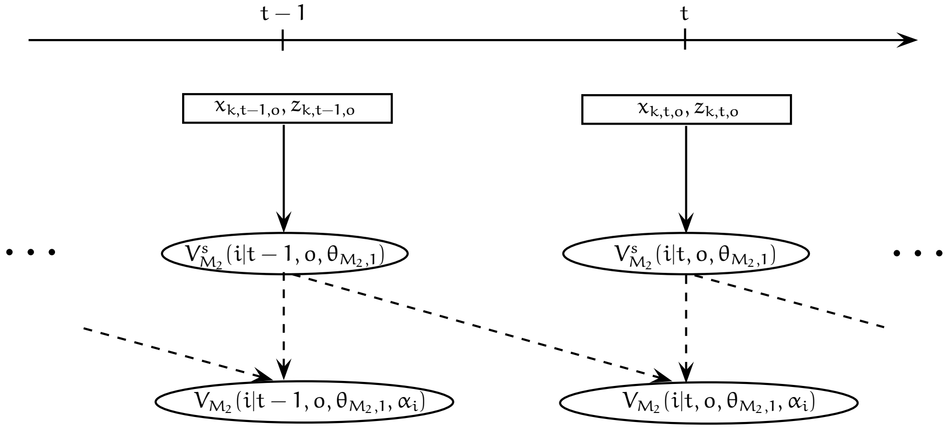

$t-1$
t
$x_{k,t-1,o}, z_{k,t-1,o}$
$x_{k,t,o}, z_{k,t,o}$
$\cdots$
$V^s_{M_2}(i|t-1,o,\theta_{M_2,1})$
$V^s_{M_2}(i|t,o,\theta_{M_2,1})$
$\cdots$
$V_{M_2}(i|t-1,o,\theta_{M_2,1},\alpha_i)$
$V_{M_2}(i|t,o,\theta_{M_2,1},\alpha_i)$

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

$$
\begin{aligned}
V^s_{M_2}(H|t,o,\theta_{M_2,1}) &= ASC_H + \sum_{j=1}^{K_{M_2}} I_{M_2,1,H,j}\theta_{M_2,1,j} \sum_{k=1}^{188} I_{M_2,j,k}x_{k,t,o} \ , \\
V^s_{M_2}(SU|t,o,\theta_{M_2,1}) &= ASC_{SU} + \sum_{j=1}^{K_{M_2}} I_{M_2,1,SU,j}\theta_{M_2,1,j} \sum_{k=1}^{188} I_{M_2,j,k}x_{k,t,o} \\
&+ \sum_{j=1}^{K^z_{M_2}} I^z_{M_2,SU,j}\theta^z_{M_2,1,j} \sum_{k=1}^{188} I^z_{M_2,j,k}z_{k,t,o} \ , \\
V^s_{M_2}(F|t,o,\theta_{M_2,1}) &= ASC_F + \sum_{j=1}^{K_{M_2}} I_{M_2,F,j}\theta_{M_2,1,j} \sum_{k=1}^{188} I_{M_2,j,k}x_{k,t,o} \ , \\
V^s_{M_2}(D|t,o,\theta_{M_2,1}) &= ASC_D + \sum_{j=1}^{K_{M_2}} I_{M_2,D,j}\theta_{M_2,1,j} \sum_{k=1}^{188} I_{M_2,j,k}x_{k,t,o} \ , \\
V^s_{M_2}(SA|t,o,\theta_{M_2,1}) &= ASC_{SA} + \sum_{j=1}^{K_{M_2}} I_{M_2,SA,j}\theta_{M_2,1,j} \sum_{k=1}^{188} I_{M_2,j,k}x_{k,t,o} \ , \\
V^s_{M_2}(A|t,o,\theta_{M_2,1}) &= ASC_A + \sum_{j=1}^{K_{M_2}} I_{M_2,A,j}\theta_{M_2,1,j} \sum_{k=1}^{188} I_{M_2,j,k}x_{k,t,o} \ , \\
V^s_{M_2}(N|t,o,\theta_{M_2,1}) &= 0 \ , \\
V^s_{M_2}(O|t,o,\theta_{M_2,1}) &= ASC_O + \sum_{j=1}^{K_{M_2}} I_{M_2,O,j}\theta_{M_2,1,j} \sum_{k=1}^{188} I_{M_2,j,k}x_{k,t,o} \ , \\
V^s_{M_2}(O|t,o,\theta_{M_2,1}) &= ASC_{DK} \ , \qquad (9)
\end{aligned}
$$

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 $\{z_{k,t,o}\}$. The indicators are similar to those introduced in Section 3.1. $I_{M_2,i,j}$ 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

$$\sum_{k=1}^{188} I_{M_2,j,k} = 1 \ \forall j \ , \tag{10}$$

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, whose 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 6 and 7. The reading of the tables is exactly the same as for Table 4 described in Section 3.1.

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 (11).

$$
\begin{aligned}
V_{M_2}(H|t, o, \theta_{M_2,1}, \alpha_H) &= V^s_{M_2}(H|t, o, \theta_{M_2,1}) \\
&+ \alpha_H V^s_{M_2}(H|t-1, o, \theta_{M_2,1}), \\
V_{M_2}(SU|t, o, \theta_{M_2,1}, \alpha_{SU}) &= V^s_{M_2}(SU|t, o, \theta_{M_2,1}), \\
V_{M_2}(F|t, o, \theta_{M_2,1}, \alpha_F) &= V^s_{M_2}(F|t, o, \theta_{M_2,1}) \\
&+ \alpha_F V^s_{M_2}(F|t-1, o, \theta_{M_2,1}), \\
V_{M_2}(D|t, o, \theta_{M_2,1}, \alpha_D) &= V^s_{M_2}(D|t, o, \theta_{M_2,1}), \\
V_{M_2}(SA|t, o, \theta_{M_2,1}, \alpha_{SA}) &= V^s_{M_2}(SA|t, o, \theta_{M_2,1}) \\
&+ \alpha_{SA} V^s_{M_2}(SA|t, o, \theta_{M_2,1}), \\
V_{M_2}(A|t, o, \theta_{M_2,1}, \alpha_A) &= V^s_{M_2}(A|t, o, \theta_{M_2,1}), \\
V_{M_2}(N|t, o, \theta_{M_2,1}, \alpha_N) &= V^s_{M_2}(N|t, o, \theta_{M_2,1}) = 0, \\
V_{M_2}(O|t, o, \theta_{M_2,1}, \alpha_O) &= V^s_{M_2}(O|t, o, \theta_{M_2,1}) \\
&+ \alpha_O V^s_{M_2}(O|t, o, \theta_{M_2,1}), \\
V_{M_2}(DK|t, o, \theta_{M_2,1}, \alpha_{DK}) &= V^s_{M_2}(DK|t, o, \theta_{M_2,1}). \qquad (11)
\end{aligned}
$$

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 8). $\{\alpha_i\}_{i=H,F,SA,O}$ 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

$$
P_{M_2}(i|t, o, \theta_{M_2,1}, \alpha_i) = \frac{e^{V_{M_2}(i|t,o,\theta_{M_2,1},\alpha_i)}}{\sum_j e^{V_{M_2}(j|t,o,\theta_{M_2,1},\alpha_j)}}. \qquad (12)
$$

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 $t = 2, \dots, T_o$,

$$\begin{aligned} V_{M_2}(t|o, \theta_{M_2,2}) &= \sum_{j=1}^{K^y_{M_2,2}} \theta^y_{M_2,2,j} \sum_{k=1}^{188} I^y_{M_2,2,j,k} y_{k,t,o} \\ &+ \sum_{j=1}^{K^z_{M_2,2}} \theta^z_{M_2,2,j} \sum_{k=1}^{188} I^z_{M_2,2,j,k} z_{k,t,o} \ , \end{aligned} \tag{13}$$

and

$$P_{M_2}(t|o; \theta_{M_2,2}) = \frac{e^{V_{M_2}(t|o,\theta_{M_2,2})}}{\sum_{\ell=1}^{T_o} e^{V_{M_2}(\ell|o,\theta_{M_2,2})}}. \tag{14}$$

$K^y_{M_2,2}$ 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

$$\sum_{k=1}^{188} I^y_{M_2,2,j,k} = 1 \ \forall j \ , \tag{15}$$

$I^z_{M_2,2,j,k}$ 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 9 (same reading as for Table 4 described in Section 3.1). Finally, the log-likelihood function is

$$\begin{aligned} \mathcal{L}(\theta_{M_2}, \alpha) &= \sum_{o=1}^{O} \sum_{i=1}^{9} w_{i,o} \log P_{M_2}(i|o, \theta_{M_2}, \alpha) \\ &= \sum_{o=1}^{O} \sum_{i=1}^{9} w_{i,o} \log(\sum_{t=1}^{T_o} P_{M_2}(i|t, o, \theta_{M_2,1}, \alpha_i) P_{M_2}(t|o, \theta_{M_2,2})). \end{aligned} \tag{16}$$

3.3 The smoothed model

In this model, we hypothesize that the behavior of the respondent is composed of two consecutive phases, when watching a video. First the respondent is waiting for information, no perception of expressions is influencing the observed choice of label. This is the first phase. At a certain point in time, the respondent starts to use the information of the frames to make her choice of label. This consideration of information is continued until the end of the video. It constitutes the second phase. The model combines a model related to the perception of expressions and a model which detects the changing of phase. The observed choice of label is supposed to be the average across the frames of the perception of expressions in the second phase. Both models are latent as only the choice of label is observed.

The first model provides the probability for a respondent to choose the expression i when exposed to frame ℓ of the video sequence o, and is written $P_{M_3}(i|l, o, \theta_{M_3,1})$. The second model $P_{M_3}(t|o, \theta_{M_3,2})$ provides the probability for a respondent to enter in her second phase when being exposed to the frame t. The probability for a respondent to label the video o with expression i, is denoted by $P_{M_3}(i|o, \theta_{M_3})$, which is observable. $\theta_{M_3,1}$ and $\theta_{M_3,2}$ are the vectors of unknown parameters to be estimated within each of the two models, merged into the vector θ_{M_3}. $P_{M_3}(i|o, \theta_{M_3})$ is the average of $\{P_{M_3}(i|l, o, \theta_{M_3,1})\}_{l=t\ldots T_o}$, weighted by $P_{M_3,n}(t|o, \theta_{M_3,2})$, sum up over all the possibilities for t, which are in $\{1 \ldots T_o\}$. We obtain

$$P_{M_3}(i|o, \theta_{M_3}) = \sum_{t=1}^{T_o} P_{M_3}(t|o, \theta_{M_3,2}) \frac{1}{T_o - t + 1} \sum_{l=t}^{T_o} P_{M_3}(i|l, o, \theta_{M_3,1}). \tag{17}$$

For $P_{M_3}(i|t, o, \theta_{M_3,1})$, a utility $V_{M_3}(i|t, o, \theta_{M_3,1})$ is associated to each expression i. The specification of $\{V_{M_3}(i|t, o, \theta_{M_3,1})\}$ is defined in Equation (18).

$$
\begin{aligned}
V_{M_3}(H|t,o,\theta_{M_3,1}) &= ASC_H + \sum_{j=1}^{K_{M_3}} I_{M_3,1,H,j}\theta_{M_3,1,j} \sum_{k=1}^{188} I_{M_3,j,k}x_{k,t,o} \ , \\
V_{M_3}(SU|t,o,\theta_{M_3,1}) &= ASC_{SU} + \sum_{j=1}^{K_{M_3}} I_{M_3,1,SU,j}\theta_{M_3,1,j} \sum_{k=1}^{188} I_{M_3,j,k}x_{k,t,o} \\
&+ \sum_{j=1}^{K^z_{M_3}} I^z_{M_3,SU,j}\theta^z_{M_3,1,j} \sum_{k=1}^{188} I^z_{M_3,j,k}z_{k,t,o} \ , \\
V_{M_3}(F|t,o,\theta_{M_3,1}) &= ASC_F + \sum_{j=1}^{K_{M_3}} I_{M_3,F,j}\theta_{M_3,1,j} \sum_{k=1}^{188} I_{M_3,j,k}x_{k,t,o} \ , \\
V_{M_3}(D|t,o,\theta_{M_3,1}) &= ASC_D + \sum_{j=1}^{K_{M_3}} I_{M_3,D,j}\theta_{M_3,1,j} \sum_{k=1}^{188} I_{M_3,j,k}x_{k,t,o} \ , \\
V_{M_3}(SA|t,o,\theta_{M_3,1}) &= ASC_{SA} + \sum_{j=1}^{K_{M_3}} I_{M_3,SA,j}\theta_{M_3,1,j} \sum_{k=1}^{188} I_{M_3,j,k}x_{k,t,o} \ , \\
V_{M_3}(A|t,o,\theta_{M_3,1}) &= ASC_A + \sum_{j=1}^{K_{M_3}} I_{M_3,A,j}\theta_{M_3,1,j} \sum_{k=1}^{188} I_{M_3,j,k}x_{k,t,o} \ , \\
V_{M_3}(N|t,o,\theta_{M_3,1}) &= 0 \ , \\
V_{M_3}(O|t,o,\theta_{M_3,1}) &= ASC_O + \sum_{j=1}^{K_{M_3}} I_{M_3,O,j}\theta_{M_3,1,j} \sum_{k=1}^{188} I_{M_3,j,k}x_{k,t,o} \ , \\
V_{M_3}(O|t,o,\theta_{M_3,1}) &= ASC_{DK} \ .
\end{aligned} \tag{18}
$$

The general description of the utilities is exactly the same as for the utilities in Equation (9). The detailed specifications of $\{V_{M_3}(i|t,o,\theta_{M_3,1})\}$ are presented in Tables 10 and 11 (same reading as for Table 4 described in Section 3.1). A logit form is postulated for $P_{M_3}(i|t,o,\theta_{M_3,1})$

$$
P_{M_3}(i|t,o,\theta_{M_3,1}) = \frac{e^{V_{M_3}(i|t,o,\theta_{M_3,1})}}{\sum_j e^{V_{M_3}(j|t,o,\theta_{M_3,1})}}. \tag{19}
$$

The second model $P_{M_3}(t|o,\theta_{M_3,2})$ is capturing the changing of phases. A utility $V_{M_3}(t|o,\theta_{M_3,2})$ is associated to each frame t in the video o

$$V_{M_3}(t|o,\theta_{M_3,2}) = \sum_{k=1}^{K^y_{M_3,2}} \theta^y_{M_3,2,k} \sum_{k=1}^{188} I^y_{M_3,2,j,k} y_{k,t,o}, \tag{20}$$

where $K^y_{M_3,2}$ is the number of parameters associated to this model. The specification of $V_{M_3}(t|o,\theta_{M_3,2})$ is generic. $I^y_{M_3,2,j,k}$ is an indicator equal to 1 if $\theta^y_{M_3,2,k}$ is associated to $y_{k,t,o}$, 0 otherwise. $\theta^y_{M_3,2,k}$ is linked to only one $y_{k,t,o}$, we have

$$\sum_{k=1}^{188} I^y_{M_3,2,j,k} = 1 \ \forall j \ , \tag{21}$$

the model contains only $\{y_{k,t,o}\}$. $\{z_{k,t,o}\}$ have been tested but do not appear to be significant . The detailed specifications of the utilities are presented in Table 12 (same reading as for Table 4 described in Section 3.1). Finally, $P_{M_3}(t|o,\theta_{M_3,2})$ is a logit model

$$P_{M_3}(t|o,\theta_{M_3,2}) = \frac{e^{V_{M_3}(t|o,\theta_{M_3,2})}}{\sum_{\ell=1}^{T_o} e^{V_{M_3}(\ell|o,\theta_{M_3,2})}}, \tag{22}$$

and the log-likelihood function is

$$\begin{aligned}\mathcal{L}(\theta_{M_3}) &= \sum_{o=1}^{O}\sum_{i=1}^{9} w_{i,o}\log P_{M_3}(i|o,\theta_{M_3}) \\ &= \sum_{o=1}^{O}\sum_{i=1}^{9} w_{i,o}\log(\sum_{t=1}^{T_o} P_{M_3}(t|o,\theta_{M_3,2})\frac{1}{T_o - t + 1}\sum_{k=t}^{T_o} P_{M_3}(i|k,o,\theta_{M_3,1})).\end{aligned} \tag{23}$$

4 Estimations of the models

The models are estimated by maximum likelihood (see Equations (7), (16), and (23)) with codes based on the BIOGEME software developed by Bierlaire (2003) to do simultaneous estimation. Estimation results are presented in Table 1.

Reduced model is the simplest model because it only accounts for the influence of the last frame on the observed choice of label. The values of the 32 estimated parameters and associated t-tests are presented in Tables

4 and 5. Fourteen parameters are related to facial measurements characterizing AU (see Section 3.1). The signs are consistent with the work of Sorci, Antonini, Cruz, Robin, Bierlaire and Thiran (2010), and with the FACS (Ekman and Friesen, 1978). The asymmetry of the face is taken into account by associating different parameters to the left and right measurements of a same type. All parameters related to AU are significantly different from 0 (t-test ≥ 1.96). This is also the case for the five parameters related to EDU and for the five parameters associated to elements of the vector C. Their signs are coherent with the work of Sorci, Antonini, Cruz, Robin, Bierlaire and Thiran (2010). Some of the eight $\{ASC_i\}$ do not appear to be significant, which is a good feature because they are designed to absorb the unobserved perception of respondents.

For **latent model**, 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 6 and 7. Signs and significance of parameters related to AU, EDU and elements of the vector C are correct and consistent with the estimated parameters obtained for **reduced model**. In addition, the model contains two more parameters. The parameter $\theta_{M_2,1,22}$ associated to the height of the mouth (*"mouth_h"*), appears to be significant, while it was not the case for **reduced model**. This is due to the fact that **reduced model** accounts only for the perception of the last frame in a video, compared to all the frames here. So the **reduced model** could not be as precisely specified as this model. $\theta^z_{M_2,1,1}$ 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 8). 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 9). 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.

For **smoothed model**, the model dealing with the perception of the expressions contains 36 parameters (see Tables 10 and 11). Signs and significance of parameters related to AU, EDU and C parameters are the same than for **reduced model**. The model contains 4 more parameters. $\theta_{M_3,1,4}$ and $\theta_{M_3,1,12}$ are respectively attached to the EDU corresponding to the fraction between the height of the eyebrows and their width ("*RAP_brow*"), and to the fifth element of the vector C ("*C_5*"). Both are in the utility of disgust. Compared to **reduced model**, they appear to be significant due to the fact that we now account for the total number of frames. $\theta^z_{M_3,1,1}$ and $\theta^z_{M_3,1,2}$ are respectively related to the variance of the height of the mouth ("*mouth_h*") and the variance of the height of the left eye ("*leye_h*") and are included in the utility of surprise in order to capture the perception of suddenness. They are positive as expected, meaning that the higher $z_{1,t,o}$ and $z_{3,t,o}$, the more the surprise is favored, which is logical. The model designed to detect the first frame of the relevant group of frames contains 8 parameters (see Table 12). They are all linked with $\{y_{k,t,o}\}$. None of the parameters attached to $\{z_{k,t,o}\}$ appeared to be significant. The perception of the short time variations of facial characteristics is not relevant for activating the second phase of behavior, which seems logical. The changing in the facial characteristics should be more drastic, that's why $\{y_{k,t,o}\}$ are better adapted. As for **reduced model** , the magnitude of the parameters is larger compared to the model handling with the perception of the expressions. The interpretation remains the same as for **latent model**.

The final log-likelihood is improved between **reduced** and **latent models**, and **reduced** and **smoothed models**. The three models can not be compared using likelihood ratio-tests. We use $\bar{\rho}^2$ as a goodness of fit to identify the best model. Looking at Table 1, **latent model** appears to be the best model, closely followed by **smoothed model**. The improvement brought by the dynamic modeling is substantial.

The magnitude of the parameter values and signs are the same for the three models. For example, $\theta_{M_1,4}$, $\theta_{M_2,1,4}$ and $\theta_{M_3,1,5}$ are 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*"). They are present in the utilities of surprise and fear. The

	Reduced model	**Latent model**	**Smoothed model**
Nb of observations	369	369	369
Nb of parameters	32	45	44
Null log-likelihood	-810.78	-810.78	-810.78
Final log-likelihood	-475.79	-441.28	-447.67
$\bar{\rho}^2$	0.374	0.400	0.394

Table 1: General estimation results

associated parameters are all positive, showing the stability of the models. Their positive sign is logical because when a person has the mouth opened, the perceived facial expression is more likely to be fear or surprise.

The specifications of the model related to the detection of the most impressive frame in **latent model**, and to the detection of the first frame of the relevant group of frames in **smoothed model**, are very similar. For **latent model**, it contains parameters associated with both $\{y_{k,t,o}\}$ and $\{z_{k,t,o}\}$ and for **smoothed model**, only associated with $\{y_{k,t,o}\}$. For example, $y_{2,t,o}$ is present in both models and 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 parameters associated to $y_{2,t,o}$ ($\theta_{M_2,2,6}$ and $\theta_{M_3,2,8}$) is positive for both **latent** and **smoothed models**, 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 $y_{2,t,o}$, frame 3 seems to be the most important.

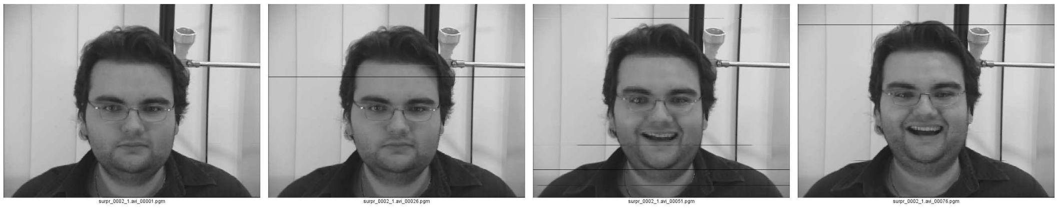

surpr_0002_1.avi_00001.pgm
surpr_0002_1.avi_00026.pgm
surpr_0002_1.avi_00051.pgm
surpr_0002_1.avi_00076.pgm

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

In conclusion, the parameters of the models are significant and interpretable. Moreover, the addition of a dynamic part in the model signifi-

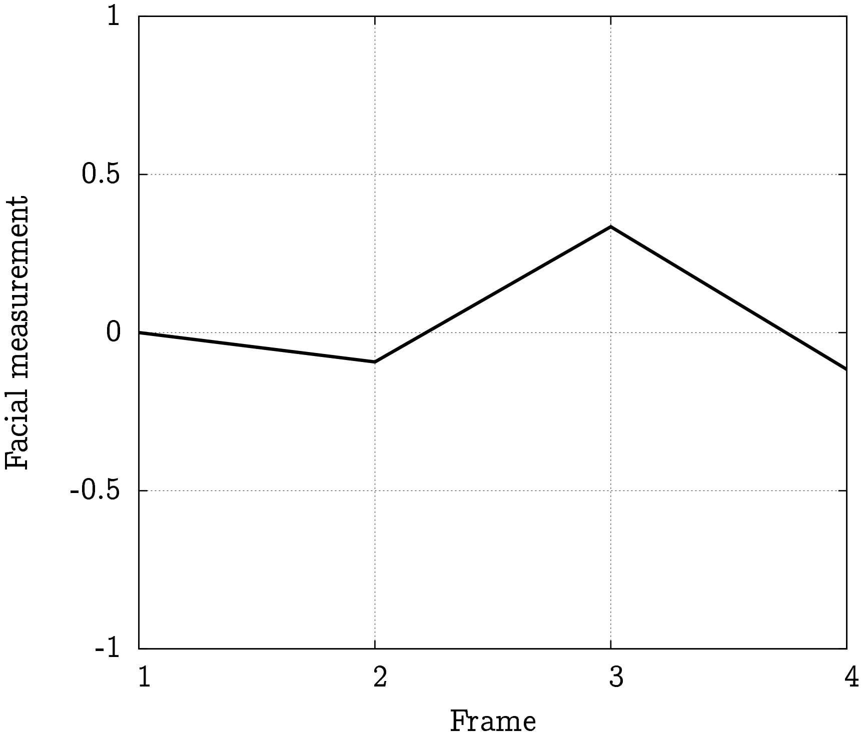


Figure 9: Variations of $y_{2,t,o}$, related to the height of the mouth ("*mouth_h*") for the video presented in Figure 8

cantly improves the fit.

5 Prediction capability

The prediction capability is tested in order to ensure the quality of the models. The dataset used in this section is the same as the one used in Section 4. We proceed in three steps: the first one consists of comparing the percentages of badly predicted observations for the proposed models. In a second step, the models are validated using the method of cross-validation. In the third step, we study the predictions of the proposed models at a more disaggregated level. This consists of picking a certain video and analysing the predictions of the models in detail.

5.1 Aggregate prediction

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. Table 2 summarizes the percentages of badly predicted observations per model. The percentages are consistent with the fitting results presented in Section 4, which is a good sign. The percentage of badly predicted observations is already low for **reduced model**. The improvement brought by **latent** and **smoothed models** compared to **reduced model** is minor in terms of prediction. This can be explained by the structure of the considered facial videos. As the “peak” emotion is often observed at the end of the video, there are few observations where the dynamic models could do better. However **smoothed model** is the best.

Reduced model	**Latent model**	**Smoothed model**
17.89	17.34	15.45

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

The cumulative distributions of the choice probabilities predicted by the models are displayed in Figure 10. If the models were perfect, the curves should be flat with a pick for choice probabilities equal to one. This would mean that the models replicate exactly the observed choices of labels. Of course this is not the case. The three curves are close in the “badly predicted” interval (choice probabilities less than $\frac{1}{9} = 0.112$). This is consistent with the results shown in Table 2. Then, in the interval $[0.112, 0.680]$ the **latent** and **smoothed models** are better than **reduced model**. In the last interval, **reduced model** appears to be better than **smoothed model**, but **latent model** is largely better than **reduced** and **smoothed models**, and predict the highest probabilities (its curve is the last to reach the level of one). These results show that **latent model** is always better than **reduced model**, and consequently demonstrate the added value of the dynamic modeling.

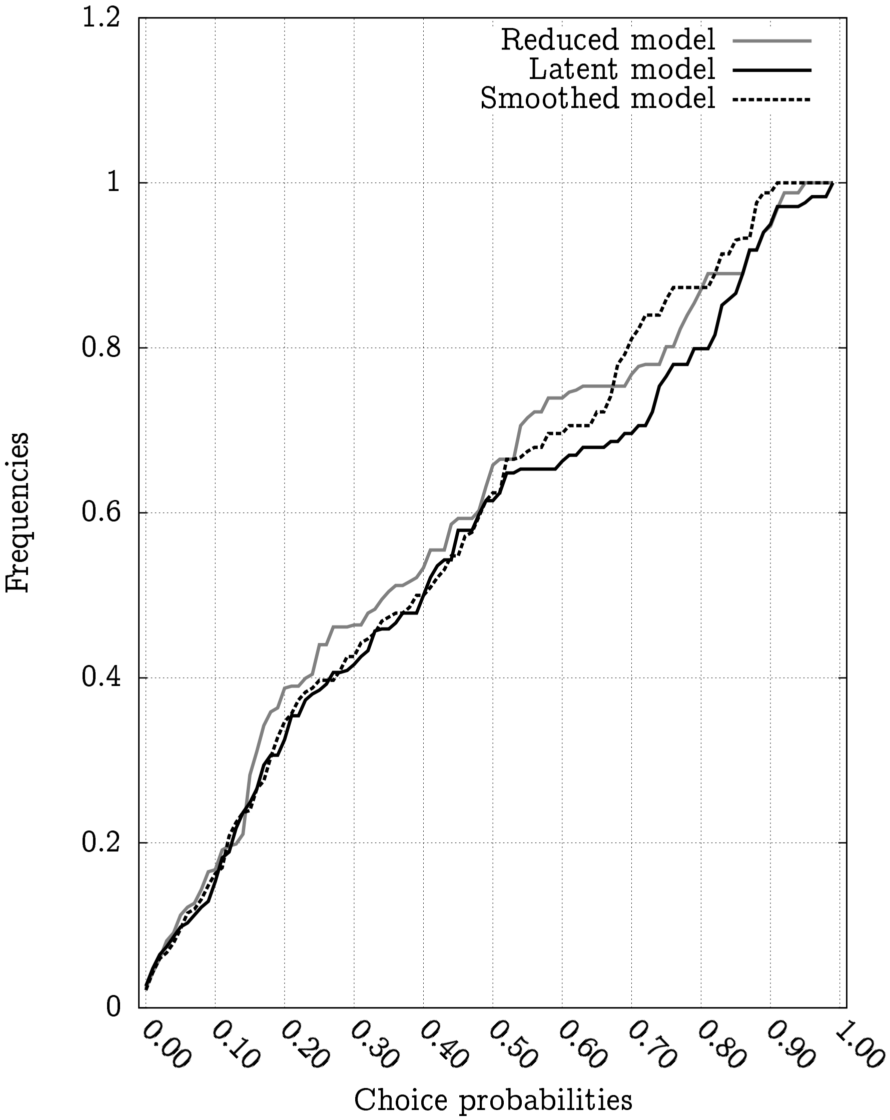


Figure 10: Cumulative distributions of the choice probabilities predicted by the three proposed models, on the estimation data

5.2 Cross-validation

The study of the badly predicted observations, described in Section 5.1 is done on the estimation data presented in Section 2. The finality of the models is to be used on some data not involved in the estimation process, for prediction. Consequently the quality of the model should be tested on some new data, but we do not have such data. In this situation, the cross-validation allows to validate the models. The methodology is inspired from the work of Robin et al. (2009) who successfully cross-validate a model of pedestrian behavior. The dataset is split into an estimation subset and a validation subset. The models are estimated on the estimation data, and are applied on the validation data. The dataset is randomly split across the videos, in five subsets. Each subset contains twenty percent of the videos. In the data, there are 65 videos, so each subset contains the collected labels related to 13 videos. Four subsets are combined into the estimation dataset. After estimation, the model is applied on the remaining subset. The operation is repeated five times. The percentages of badly predicted observations, calculated over the validation subsets are presented in Table 3.

Validation subsets	1	2	3	4	5
Reduced model	28.74	26.15	21.31	21.87	28.26
Latent model	24.14	13.85	11.48	17.19	21.74
Smoothed model	20.69	16.92	18.03	15.63	10.87

Table 3: Percentages of badly predicted observations calculated over the validation subsets, obtained when cross-validating the models

Looking at Table 3, the two dynamic models (**latent** and **smoothed models**) are always better than **reduced model**. In addition, the percentages of badly predicted observations are close from those obtained on the entire estimation data (see Table 2) for **latent** and **smoothed models**, not **reduced model**. The dynamic models appear to be much more robust than **reduced model**. This justifies the goodness of the approach and the validity of the dynamic models.

5.3 Disaggregate prediction

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 three models. The video is the same than the one considered in Figure 8. The detailed predictions of the models are shown in Figure 11 for **reduced model**, Figure 12 for **latent model**, and Figure 13 for **smoothed model**. On those figures, 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. For **reduced model**, only the last frame is considered relevant, so the peak is logically on this last frame. For **latent model**, it shows the influence of each frame on the final expression choice. For **smoothed model**, the peak measures the contribution of the average perception of the following group of frames (until the end of the video), including the frame of the peak. Finally in the extreme right column, you find on the second row the final probability distribution among the expressions, which is predicted by the model, and on the third row, the distribution of the collected labels for the video.

On the first frame of the considered video (see Figures 11, 12 and 13), the face tends to be neutral, and then evolves toward a different expression. Seven respondents have labelled 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 figures. **Reduced model** predicts 65% of happiness, 35% of surprise, and 0% for anger. The prediction seems logical regarding only the facial characteristics in the last frame.

Latent model predicts 24% of happiness, 58% of surprise, 18% of disgust and 0% for anger. This is further away from the distribution of the collected labels, compared to **reduced model**. The model has selected frame 3 as being the most impressive 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 3.2), 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 (9)), 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.

Smoothed model predicts 58% of happiness, 38% of surprise, 4% of disgust and 0% of anger. The prediction is well adapted to the observed distribution of labels. The model detects frame 3 as being the first frame of the relevant group of frames. As for **latent model**, this is due to the presence of $\{y_{k,t,o}\}$ in the utilities of the frames (see Section 3.3), and $\{y_{k,t,o}\}$ are higher for this frame (see Figure 9). The model handling with the perception of the expressions predicts more surprise than happiness for frame 3, and the contrary for frame 4. This is logical due to the perception of suddenness in frame 3 (see the utility of surprise in Equation (18)). The facial characteristics are stabilized in frame 4 and lead to the expression happiness, which is coherent. The final prediction of the model is the average of the perception of expressions among the frames of the relevant group (frames 3 and 4), which explains the balanced share between happiness and surprise.

The predictions of the three models are explainable. **Smoothed model** seems to be the most interpretable and predicts the closest distribution of probability across the expressions, from the collected labels.

6 Conclusions and Perspectives

We propose a new approach for the recognition of dynamic facial expressions. The estimation of the models is based on labels collected through respondents to an internet survey. The developed models capture up causal effects between facial characteristics and expressions. Statistical tests and model predictions have proved the quality of the models, and the added value of the dynamic formulation (**latent** and **smoothed models** compared to **reduced model**). The models have been cross-validated on the estimation data, **latent** and **smoothed models** appear to be more robust than **reduced model**. Finally, some qualitative analysis of the model predictions allow to confirm 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.

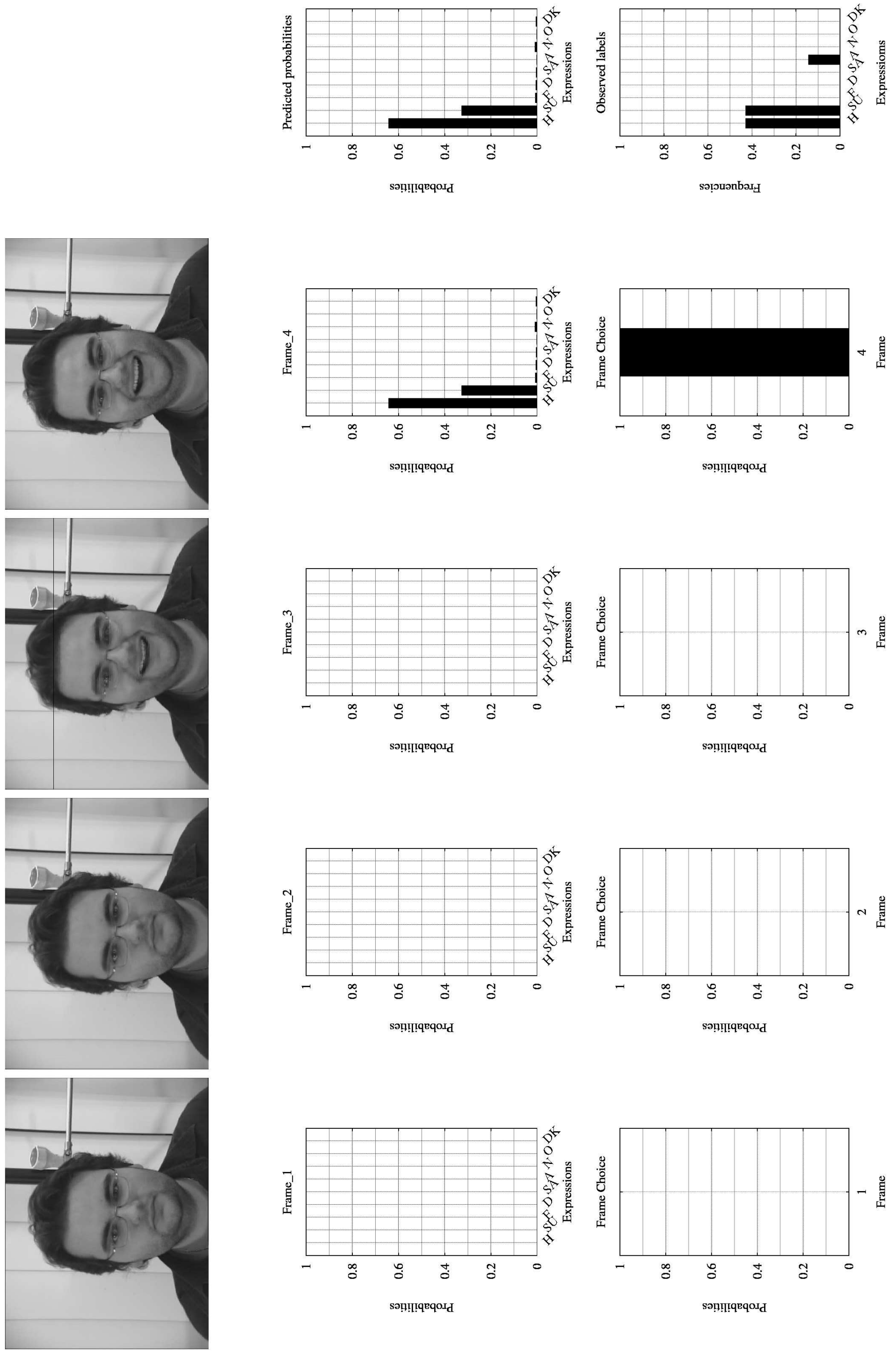


Figure 11: Example of a detailed prediction of **reduced model**

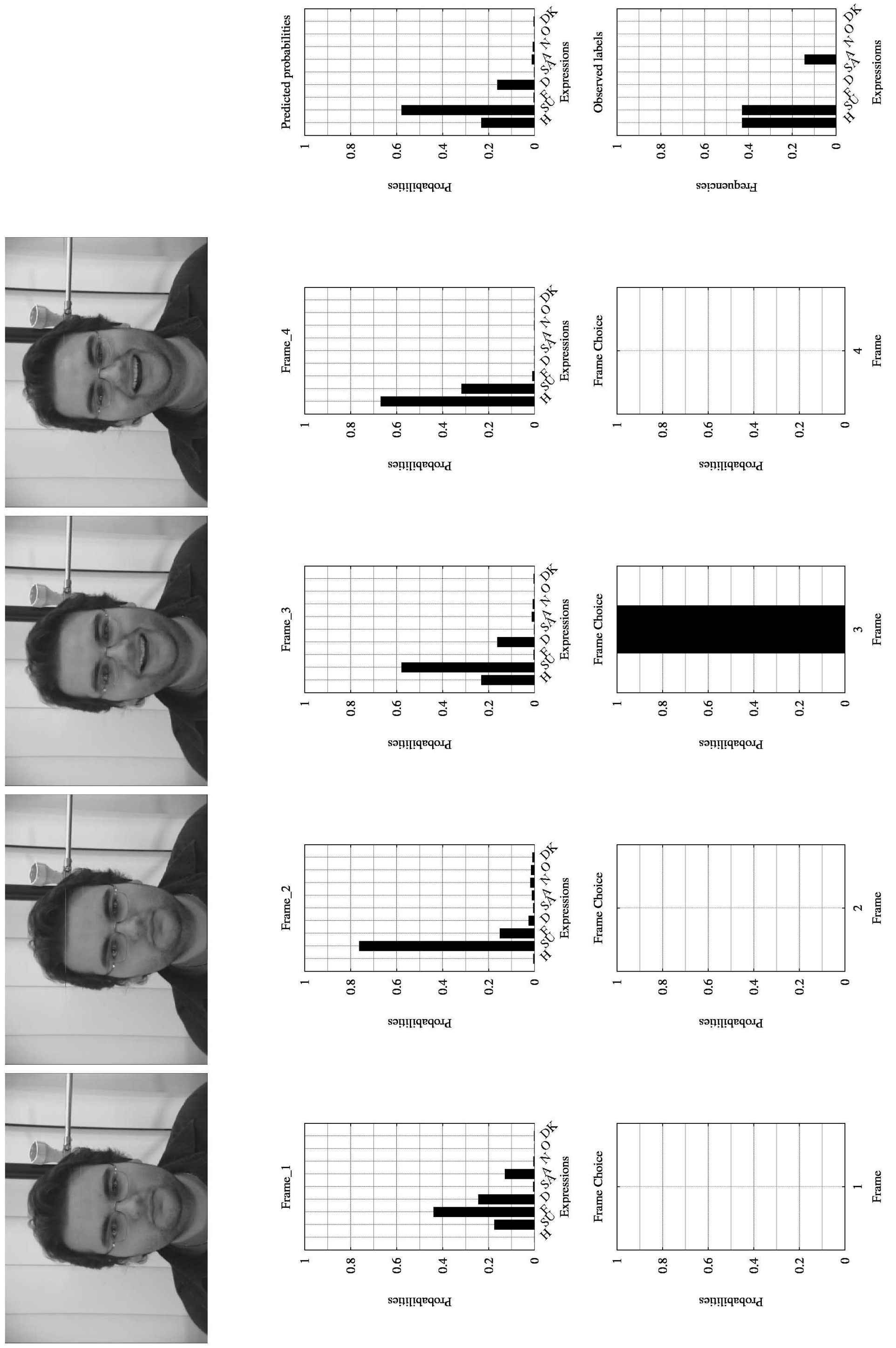


Figure 12: Example of detailed prediction of **latent model**

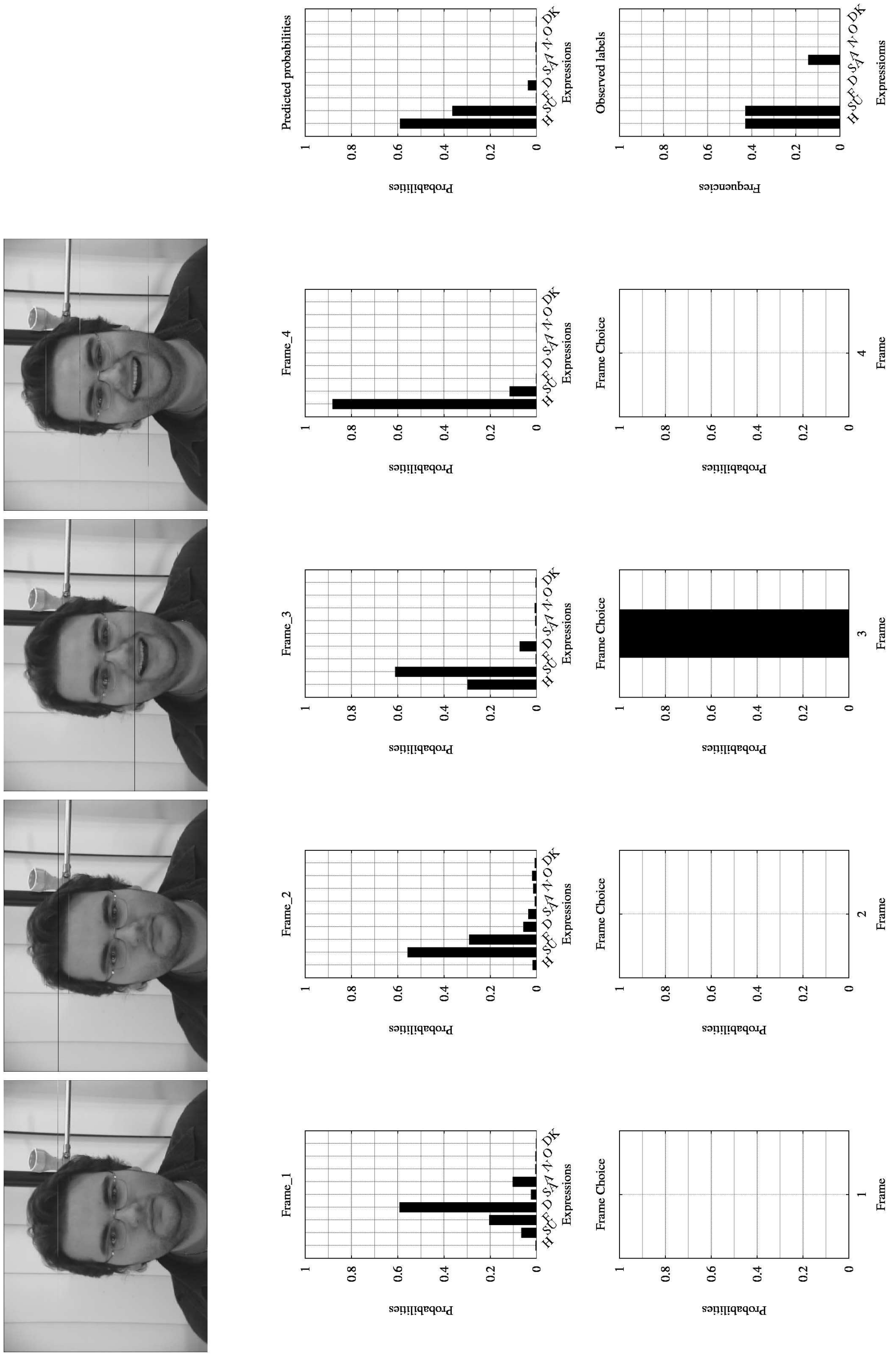


Figure 13: Example of detailed prediction of **smoothed model**

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.

Robin, T., Antonini, G., Bierlaire, M. and Cruz, J. (2009). Specification, estimation and validation of a pedestrian walking behavior model, *Transportation Research Part B: Methodological* **43**(1): 36–56.

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	$x_{k,T_o,o}$	value	t-test 0
ASC_A						×				1	0.95	0.28
ASC_D				×						1	25.38	7.88
ASC_{DK}									×	1	-0.69	-1.79
ASC_F			×							1	0.49	0.19
ASC_H	×									1	-3.14	-0.79
ASC_O								×		1	6.95	3.20
ASC_{SA}					×					1	10.80	2.54
ASC_{SU}		×								1	-11.27	-5.63

Table 4: Estimation results of the constants for **reduced model**

parameter	H	SU	F	D	SA	A	N	O	DK	$x_{k,T_o,o}$	value	t-test 0
$\theta_{M_1,1}$				×						EDU_6	-6.52	-3.63
$\theta_{M_1,2}$				×						EDU_8	-4.75	-6.18
$\theta_{M_1,3}$		×				×				RAP_brow	6.70	4.53
$\theta_{M_1,4}$		×	×							RAP_mouth	2.94	2.85
$\theta_{M_1,5}$	×									RAP_mouth	9.36	5.35
$\theta_{M_1,6}$	×									C_1	-16.30	-3.51
$\theta_{M_1,7}$						×				C_2	23.98	3.49
$\theta_{M_1,8}$				×						C_2	26.22	5.16
$\theta_{M_1,9}$	×									C_3	15.34	3.13
$\theta_{M_1,10}$		×								C_3	15.73	3.27
$\theta_{M_1,11}$					×					broweye_l2	153.91	3.17
$\theta_{M_1,12}$		×								broweye_l3	85.58	5.75
$\theta_{M_1,13}$		×	×	×	×	×				broweye_r2	-49.81	-4.30
$\theta_{M_1,14}$			×		×					eye_angle_l	58.55	3.43
$\theta_{M_1,15}$					×					eye_brow_angle_l	-140.87	-5.10
$\theta_{M_1,16}$				×						eye_mouth_dist_l2	-69.83	-3.42
$\theta_{M_1,17}$	×				×			×		eye_mouth_dist_l	-36.03	-2.89
$\theta_{M_1,18}$						×				eye_nose_dist_l	245.03	5.05
$\theta_{M_1,19}$			×	×	×			×		eye_nose_dist_l	147.67	4.89
$\theta_{M_1,20}$			×	×	×	×		×		eye_nose_dist_r	-213.93	-6.04
$\theta_{M_1,21}$		×	×							leye_h	20.97	2.09
$\theta_{M_1,22}$					×	×				mouth_nose_dist2	-90.97	-2.15

parameter	H	SU	F	D	SA	A	N	O	DK	$x_{k,T_o,o}$	value	t-test 0
$\theta_{M_1,23}$	×									mouth_nose_dist	-236.37	-5.65
$\theta_{M_1,24}$	×									mouth_w	188.42	4.90

Table 5: Estimation results and description of the specification of **reduced model**

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 6: 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

parameter	H	SU	F	D	SA	A	N	O	DK	$x_{k,t,o}$	value	t-test 0
$\theta_{M_2,1,16}$				×						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
$\theta^{z}_{M_2,1,1}$										mouth_h, $z_{1,t,o}$	50.21	3.04

Table 7: 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 8: 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

parameter	$y_{k,t,o}$	value	t-test 0

Table 9: Estimation results and description of the specification of **latent model**, associated to the model which detects the most meaningful frame

parameter	H	SU	F	D	SA	A	N	O	DK	$x_{k,t,o}$	value	t-test 0
ASC_A						×				1	-7.53	-1.63
ASC_D				×						1	20.28	4.03
ASC_{DK}									×	1	-0.69	-1.79
ASC_F			×							1	-0.35	-0.09
ASC_H	×									1	-7.66	-1.43
ASC_O								×		1	12.95	4.38
ASC_{SA}					×					1	4.17	1.04
ASC_{SU}		×								1	-29.15	-7.07

Table 10: Estimation results of the constants for **smoothed model**, associated to the expression perception model

parameter	H	SU	F	D	SA	A	N	O	DK	$x_{k,t,o}$	value	t-test 0
$\theta_{M_3,1,1}$				×						EDU_6	-9.19	-3.82
$\theta_{M_3,1,2}$				×						EDU_8	-4.18	-4.09
$\theta_{M_3,1,3}$		×				×				RAP_brow	12.6	5.69
$\theta_{M_3,1,4}$				×						RAP_brow	5.44	2
$\theta_{M_3,1,5}$		×	×							RAP_mouth	2.89	2
$\theta_{M_3,1,6}$	×									RAP_mouth	11.77	4.44
$\theta_{M_3,1,7}$	×									C_1	-23.36	-3.36
$\theta_{M_3,1,8}$						×				C_2	42.46	5.3
$\theta_{M_3,1,9}$				×						C_2	33.98	5.51
$\theta_{M_3,1,10}$	×									C_3	25.82	3.88
$\theta_{M_3,1,11}$		×								C_3	17.61	2.74
$\theta_{M_3,1,12}$				×						C_5	-16.4	-2.5
$\theta_{M_3,1,13}$					×					broweye_l2	149.31	3.15
$\theta_{M_3,1,14}$		×								broweye_l3	128.49	5.76
$\theta_{M_3,1,15}$		×	×	×	×	×				broweye_r2	-61.58	-4.31

parameter	H	SU	F	D	SA	A	N	O	DK	$x_{k,t,o}$	value	t-test 0
$\theta_{M_3,1,16}$			×		×					eye_angle_l	40.99	2.06
$\theta_{M_3,1,17}$					×					eye_brow_angle_l	-126.55	-4.59
$\theta_{M_3,1,18}$				×						eye_mouth_dist_l2	-50.07	-2.13
$\theta_{M_3,1,19}$	×				×			×		eye_mouth_dist_l	-32.09	-2.2
$\theta_{M_3,1,20}$						×				eye_nose_dist_l	163.49	3.75
$\theta_{M_3,1,21}$			×	×	×			×		eye_nose_dist_l	114.66	3.15
$\theta_{M_3,1,22}$			×	×	×	×		×		eye_nose_dist_r	-256.49	-5.39
$\theta_{M_3,1,23}$		×	×							leye_h	52.58	3.73
$\theta_{M_3,1,24}$	×	×	×	×	×	×				mouth_h	90.92	2.96
$\theta_{M_3,1,25}$	×									mouth_nose_dist	-342.14	-6.17
$\theta_{M_3,1,26}$	×									mouth_w	228.81	4.47
$\theta^{z}_{M_3,1,1}$		×								mouth_h, $z_{1,t,o}$	0.13	4.46
$\theta^{z}_{M_3,1,2}$		×	×							leye_h, $z_{3,t,o}$	0.04	2.39

Table 11: Estimation results and description of the specification of **smoothed model**, associated to the expression perception model

parameter	$y_{k,t,o}$	value	t-test 0
$\theta^{y}_{M_3,2,1}$	C_1	-234.75	-1.75
$\theta^{y}_{M_3,2,2}$	eye_brow_angle	548.34	1.76
$\theta^{y}_{M_3,2,3}$	mouth_w	23.29	1.81
$\theta^{y}_{M_3,2,4}$	C_2	101.9	1.85
$\theta^{y}_{M_3,2,5}$	C_3	-221.23	-1.57
$\theta^{y}_{M_3,2,6}$	C_5	529.64	1.91
$\theta^{y}_{M_3,2,7}$	eye_h	-122.15	-1.79
$\theta^{y}_{M_3,2,8}$	mouth_h	119.21	1.88

Table 12: Estimation results and description of the specification of **smoothed model**, associated to the model related to the detection of the first frame of the relevant group of frames