Robin, Thomas; Bierlairey, Michel; Cruz, Javier

**Article**

# Dynamic facial expression recognition with a discrete choice model

Journal of Choice Modelling

**Provided in Cooperation with:**
Journal of Choice Modelling

*Suggested Citation:* Robin, Thomas; Bierlairey, Michel; Cruz, Javier (2011) : Dynamic facial expression recognition with a discrete choice model, Journal of Choice Modelling, ISSN 1755-5345, University of Leeds, Institute for Transport Studies, Leeds, Vol. 4, Iss. 2, pp. 95-148

*Journal of Choice Modelling*, 2(1), pp. 95-148
www.jocm.org.uk

# Dynamic facial expression recognition with a discrete choice model

Thomas Robin* Michel Bierlaire† Javier Cruz‡

Transport and Mobility Laboratory, Ecole Polytechnique Fédérale de Lausanne, CH-1015 Lausanne, Switzerland

Received 23 April 2010, revised version received 4 May 2011, accepted 4 August 2011

**Abstract**

A generation of new models has been proposed to handle some complex human behaviors. These models account for the data ambiguity, and therefore extend the application field of the discrete choice modeling. The facial expression recognition (FER) is highly relevant in this context. We develop a dynamic facial expression recognition (DFER) framework based on discrete choice models (DCM). The DFER consists in modeling the choice of a person who has to label a video sequence representing a facial expression. The originality is based on the the analysis of videos with discrete choice models as well as the explicit modeling of causal effects between the facial features and the recognition of the expression. Five models are proposed. The first assumes that only the last frame of the video triggers the choice of the expression. The second model has two components. The first captures the perception of the facial expression within each frame in the sequence, while the second determines which frame triggers the choice. The third model is an extension of the second model and assumes that the choice of the expression results from the average of perceptions within a group of frames. The fourth and fifth models integrate the panel effect inherent to the estimation data and are respectively extensing the first and second models. 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 by cross-validation using the estimation data.

*Keywords:* video analysis; dynamic facial expression analysis; latent class models; modeling of ambiguity; collection of facial expression data; FACS

*T: +41 (0) 21 693 24 35, F: +41 (0) 21 693 80 60, thomas.robin@epfl.ch
†T: +41 (0) 21 693 25 37, F: +41 (0) 21 693 80 60, michel.bierlaire@epfl.ch
‡T: +41 (0) 21 693 24 35, F: +41 (0) 21 693 80 60, javier.cruz@epfl.ch



ISSN 1755-5345

# 1 Introduction

A new generation of models has been proposed to account for the complexity of the human behavior. Hybrid choice models allow handling of several behavioral aspects, such as discrete and continuous choices, or attitudes (Ben-Akiva et al., 2002). Models have also been proposed to capture the dynamics of phenomena, such as the integration of planning and action in discrete choice models (Ben-Akiva, 2010).

Regarding data, new technologies allow to collect a huge amount of detailed information, coming from several sources. Data is often dynamic and ambiguous. The ambiguity comes from the difficulty to identify real decisions, decisional contexts, and the inaccuracy of sensors. The dynamics comes from the observation of single individual behaviors across time. The underlying behavioral phenomena need complex models to be addressed, therefore DCM appears to be relevant. In this context, some attempts have been made to model the route choice using ambiguous GPS data provided by smart phones (Bierlaire et al., 2010), or by GPS devices (Bierlaire and Frejinger, 2008). The ambiguity of the data is directly taken into account while developing models. Another emerging way for collecting data is the video. Video devices are nowadays cheap and easy to install. They allow to collect detailed information about behaviors. Despite this quality, these data remains difficult to exploit due to their complexity.

In this paper, we propose a complete methodology for analyzing videos using DCM. Rare literature has been reported on this subject. We perform a detailed analysis, which underlines the added-value of the modeling method and provides a complete modeling framework, starting from the data collection and ending to the model validation. We focus on the facial expression recognition (FER). A facial expression is a mixture of several pure expressions, and a face is described by a large number of variables. This leads to ambiguity. In addition, the associated data is dynamic when facial videos are used. Facial expressions 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 addition, the emotion is essential in many choice processes (Lerner and Keltner, 2000; Mellers and McGraw, 2001) and the facial expression is one of its main indicators.

Some coding systems have been proposed to describe facial expressions. Ekman and Friesen (1978) introduced the facial action coding system (FACS). They identified a list of fundamental expressions and associated groups of muscles tenseness or relaxations, called action units (AU) to each basic expression. A FACS expert can recognize AU activated on a human face, and then deducts the facial expression mixture precisely. This is 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. 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 on the robustness in case of environmental changes.

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

Measuring user emotions has become an important research topic in transportation behavior analysis. 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 rage. Currently, the mood recognition is based only on the driver's voice. For routine trips, Abou-Zeid (2009) conducted experiments to measure the travel well-being for both public transportation and car modes. Collected data was employed to estimate mode choice models. Well-being measures were used as utility indicators, in addition to standard choice indicators.

Contrary to computer vision algorithms which are calibrated using a ground truth, the proposed 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 between facial expressions. We have successfully applied the approach for SFER (Sorci et al., 2010a,b). We proposed a logit model, with nine alternatives corresponding to the nine expressions that are considered. Each utility is a function of measures related to the AU associated to the expression, as defined by the FACS. Sorci et al. (2010a) 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, so that the global facial perception can be accounted for.

Since, the DFER does not fit into the usual discrete choice applications, certain adjustments have to be done. We took inspiration from the work of Choudhury (2007). They used a dynamic behavioral framework to model car lane changing, and more generally from the framework developed by Ben-Akiva (2010) for the concept of "planning and action". Five models are presented in this analysis. Different modeling assumptions are tested and compared. We first present the behavioral data used to estimate the models. Then, we present the specification of the proposed models and the associated estimation results. We finally describe the cross-validation and the predictions of the proposed models. In order to ease the understanding of the mentioned acronyms, Table 4 in Appendix A summarizes them and their definitions.

## 2 Data

The data is derived from a set of video sequences from the facial expressions and emotions database (FEED) collected by Wallhoff (2004). This collection has recordings of students watching television. Different types of TV programs are presented to the subjects in order to generate a large spectrum of facial expressions. The database contains 95 sequences from 18 subjects. The 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 a facial expression (see example in Figure 2).
We have selected 65 videos from 17 subjects. The videos of subject $N^o17$ were removed because of the lack of variability in facial characteristics, and due to

Figure 1: Snapshot of a FEED database video: neutral face (subject $N^o2$)

Figure 2: Snapshot of a FEED database video: expression produced by the TV program (subject $N^o2$)

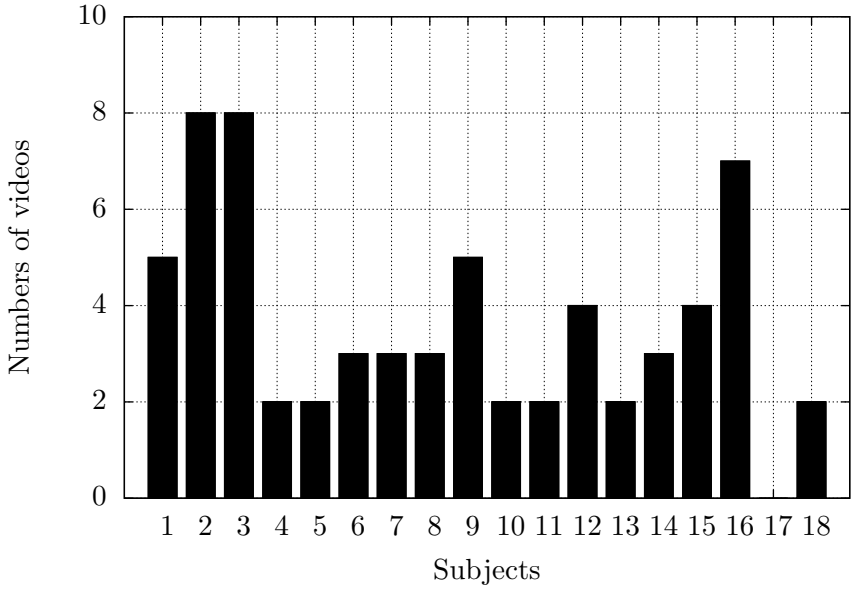



Figure 3: Numbers of considered videos per subject

some discontinuities in the recording. The number of selected 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 is extracted

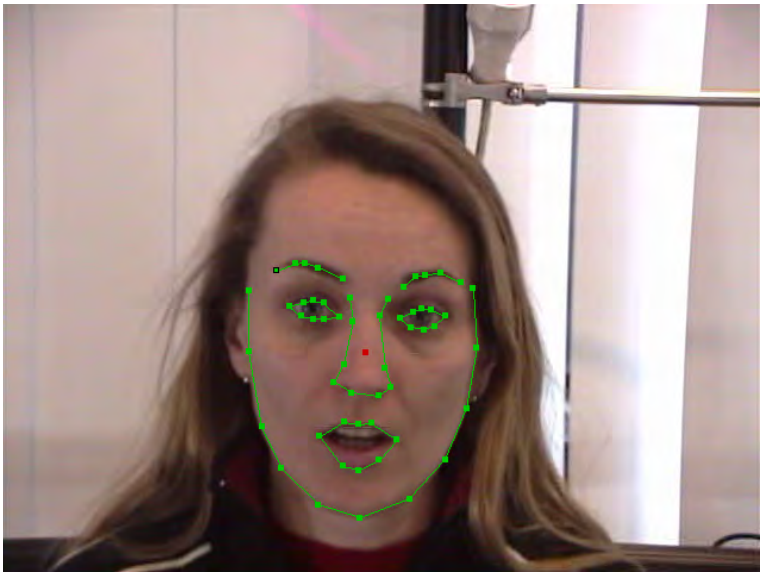


Figure 4: Mask tracked by AAM along a video sequence

using an active appearance model (AAM), see Cootes et al. (2002). It allows to extract facial distances and angles as well as facial texture information (such as levels of grey) 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. A total of 88 variables capturing distances (number of pixels) and angles (radians), as well as 100 elements of the vector $C$, are generated for each image in each video. This leads to 188 variables per image. We describe these explanatory variables in Appendix B. Note that the complete description of these variables can be found in Sorci et al. (2010a).

The video is discretized in groups of 25 images, each corresponding to one second clipping, *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 images are the features of the first image of the group. In the following, we use the word "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 (refer Equation (2)).

For a given frame $t$ and video $o$, three sets of variables are introduced: $\{x_{k,t,o}\}_{k=1,\ldots,188}$, $\{y_{k,t,o}\}_{k=1,\ldots,188}$, $\{z_{k,t,o}\}_{k=1,\ldots,188}$. $\{x_{k,t,o}\}_{k=1,\ldots,188}$ are the features extracted using the AAM (188 = 88 variables capturing distances + 100 elements of the C vector).

Frame dynamics is captured by variables $y_{k,t,o}$. For each $x_{k,t,o}$, $k = 1, \ldots, 188$,

$y_{k,t,o}$ is defined as

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

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.

Finally, another set of variables $z_{k,t,o}$, is introduced to capture the variation of $x_{k,t,o}$ within a frame. For each $x_{k,t,o}$, $k = 1, \ldots, 188$, $z_{k,t,o}$ 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

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

implying 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. The list of labels is composed of the seven basic expressions described by Keltner (2000): happiness (H), surprise (SU), fear (F), disgust (D), sadness (SA), anger (A), neutral (N). We have also added "Other" (O) and "I don't know" (DK), to avoid ambiguities in the survey. It is available at *http://transp-or2.epfl.ch/videosurvey/* since august 2008. A screen snapshot is shown at Figure 5.

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

## 3 Model specification

We consider a decision-maker who has to label a video sequence by choosing among the list of facial expressions described in Section 2 (happiness (H), surprise (SU), fear (F), disgust (D), sadness (SA), anger (A), neutral (N), other (O), not known (DK)). Five models based on different assumptions have been developed. We suppose that the perception of the respondent starts at the first frame of the

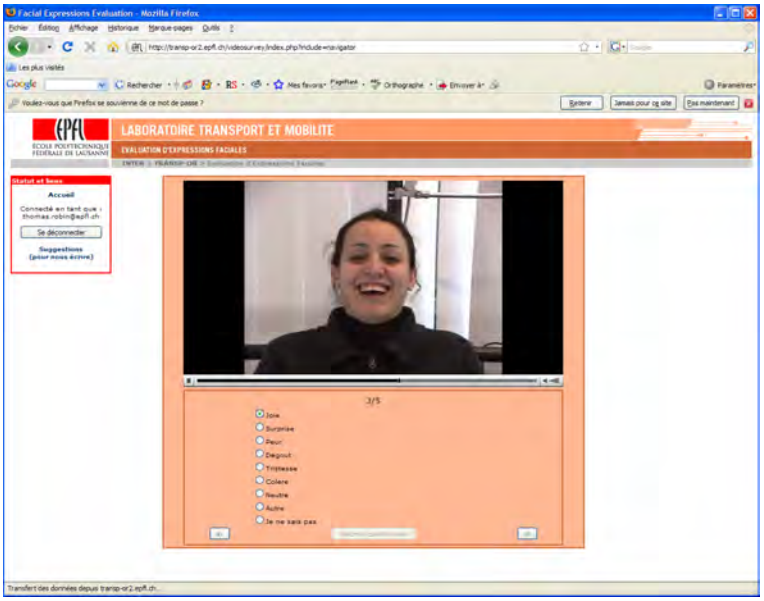



Figure 5: Snapshot of the internet survey screen (subject $N^o15$)

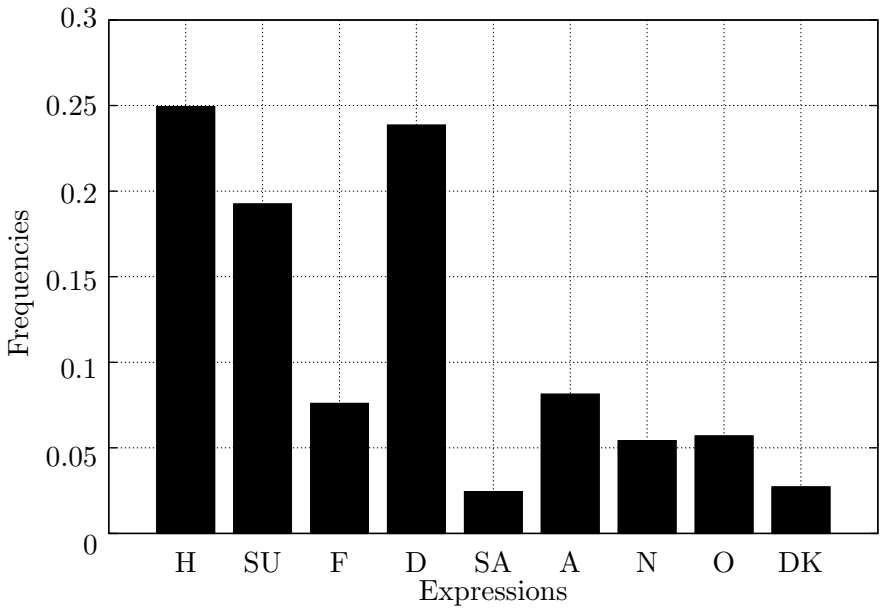



Figure 6: Distribution of the collected labels between expressions

video. Then, we assume that the respondent updates her perception every second, corresponding to every frame (see Section 2). In the first model we assume 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

features and it is considered as a reference point 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**. In the third model, we assume that it is the average perception of a group of consecutive frames which generates the choice of label. This is called **smoothed model**. Two supplementary models are proposed in order to account for the panel nature of the data, they are based on the first and second models and called **reduced model with panel effect** and **latent model with panel effect**. A **smoothed model with panel effect** is not considered here due to its estimation complexity.

The theoretical details and specification for each model are described in Sections 3.1, 3.2, 3.3, 3.4.1 and 3.4.2. They are all extensions to the model proposed by Sorci et al. (2010a), which is refered to as **static model**. Due to the small number of respondents, their socio-economic characteristics have not been included in the models.

### 3.1 The reduced model

We first assume that the perception of the last frame of a video is suggesting the choice of label. The filmed subject starts with a neutral face and evolves towards a certain expression which is triggered by the TV program that she is watching. The subject's face on the last frame should be expressive. The model is a direct application of the **static model** in the last frame. 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 $\theta_{M_1}$. 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).

$$
\begin{aligned}
V_{M_1}(H|T_o,o,\theta_{M_1}) &= ASC_{M_1,H} + \sum_{j=1}^{K_{M_1}} I_{M_1,H,j}\theta_{M_1,j} \sum_{k=1}^{188} I_{M_1,j,k} x_{k,T_o,o} \ , \\
V_{M_1}(SU|T_o,o,\theta_{M_1}) &= ASC_{M_1,SU} + \sum_{j=1}^{K_{M_1}} I_{M_1,SU,j}\theta_{M_1,j} \sum_{k=1}^{188} I_{M_1,j,k} x_{k,T_o,o} \ , \\
V_{M_1}(F|T_o,o,\theta_{M_1}) &= ASC_{M_1,F} + \sum_{j=1}^{K_{M_1}} I_{M_1,F,j}\theta_{M_1,j} \sum_{k=1}^{188} I_{M_1,j,k} x_{k,T_o,o} \ , \\
V_{M_1}(D|T_o,o,\theta_{M_1}) &= ASC_{M_1,D} + \sum_{j=1}^{K_{M_1}} I_{M_1,D,j}\theta_{M_1,j} \sum_{k=1}^{188} I_{M_1,j,k} x_{k,T_o,o} \ , \\
V_{M_1}(SA|T_o,o,\theta_{M_1}) &= ASC_{M_1,SA} + \sum_{j=1}^{K_{M_1}} I_{M_1,SA,j}\theta_{M_1,j} \sum_{k=1}^{188} I_{M_1,j,k} x_{k,T_o,o} \ ,
\end{aligned}
$$

$$
\begin{aligned}
V_{M_1}(A|T_o, o, \theta_{M_1}) &= ASC_{M_1,A} + \sum_{j=1}^{K_{M_1}} I_{M_1,A,j}\theta_{M_1,j} \sum_{k=1}^{188} I_{M_1,j,k} x_{k,T_o,o} \ , \\
V_{M_1}(N|T_o, o, \theta_{M_1}) &= 0 \ , \\
V_{M_1}(O|T_o, o, \theta_{M_1}) &= ASC_{M_1,O} + \sum_{j=1}^{K_{M_1}} I_{M_1,O,j}\theta_{M_1,j} \sum_{k=1}^{188} I_{M_1,j,k} x_{k,T_o,o} \ , \\
V_{M_1}(DK|T_o, o, \theta_{M_1}) &= ASC_{M_1,DK} \ , \qquad (4)
\end{aligned}
$$

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 associated to facial measurements $\{x_{k,t,o}\}$ in the **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

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

implying that a parameter $\theta_{M_1,j}$ is related to only one facial measurement $x_{k,T_o,o}$. Each utility contains an alternative specific constant $ASC_{M_1,i}$ 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 7 and 8. 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 $i$. This notations is used in all Tables in Appendix C. 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

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

Then the log-likelihood is

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

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 et al. (2010a) employed the database proposed by Kanade et al. (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 available data. We use 369 observations of labels against 38110 for the work of Sorci et al. (2010a). 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 the **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 $P_{M_2}(i|t, o, \theta_{M_2,1}, \alpha_{M_2})$. 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 $P_{M_2}(i|o, \theta_{M_2}, \alpha_{M_2})$, 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 $\theta_{M_2}$. $\alpha_{M_2}$ is a vector of parameters capturing the memory effects, which will be introduced in Equation (11), and has to be estimated ($\alpha_{M_2} = \{\alpha_{M_2,i}\}_{i=H,SU,F,D,SA,A,O}$). We obtain

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

For specifying the model $P_{M_2}(i|t, o, \theta_{M_2,1}, \alpha_{M_2})$, 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 $t-1$. $V_{M_2}(i|t, o, \theta_{M_2,1}, \alpha_{M_2,i})$ 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

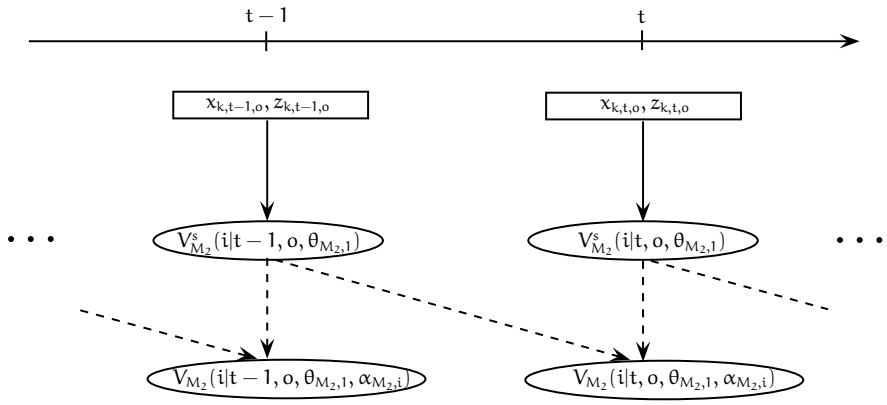



Figure 7: The dynamic process of the **latent model**

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 $V_{M_2}(i|t, o, \theta_{M_2,1}, \alpha_{M_2,i})$ 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)

$$
\begin{aligned}
V^s_{M_2}(H|t,o,\theta_{M_2,1}) &= ASC_{M_2,H} + \sum_{j=1}^{K_{M_2}} I_{M_2,1,H,j}\theta_{M_2,1,j} \sum_{k=1}^{188} I_{M_2,j,k} x_{k,t,o} \ , \\
V^s_{M_2}(SU|t,o,\theta_{M_2,1}) &= ASC_{M_2,SU} + \sum_{j=1}^{K_{M_2}} I_{M_2,1,SU,j}\theta_{M_2,1,j} \sum_{k=1}^{188} I_{M_2,j,k} x_{k,t,o} \\
&+ \sum_{j=1}^{K^z_{M_2}} I^z_{M_2,SU,j}\theta^z_{M_2,1,j} \sum_{k=1}^{188} I^z_{M_2,j,k} z_{k,t,o} \ , \\
V^s_{M_2}(F|t,o,\theta_{M_2,1}) &= ASC_{M_2,F} + \sum_{j=1}^{K_{M_2}} I_{M_2,F,j}\theta_{M_2,1,j} \sum_{k=1}^{188} I_{M_2,j,k} x_{k,t,o} \ , \\
V^s_{M_2}(D|t,o,\theta_{M_2,1}) &= ASC_{M_2,D} + \sum_{j=1}^{K_{M_2}} I_{M_2,D,j}\theta_{M_2,1,j} \sum_{k=1}^{188} I_{M_2,j,k} x_{k,t,o} \ ,
\end{aligned}
$$

$$
\begin{aligned}
V^s_{M_2}(SA|t,o,\theta_{M_2,1}) &= ASC_{M_2,SA} + \sum_{j=1}^{K_{M_2}} I_{M_2,SA,j}\theta_{M_2,1,j} \sum_{k=1}^{188} I_{M_2,j,k} x_{k,t,o} \ , \\
V^s_{M_2}(A|t,o,\theta_{M_2,1}) &= ASC_{M_2,A} + \sum_{j=1}^{K_{M_2}} I_{M_2,A,j}\theta_{M_2,1,j} \sum_{k=1}^{188} I_{M_2,j,k} x_{k,t,o} \ , \\
V^s_{M_2}(N|t,o,\theta_{M_2,1}) &= 0 \ , \\
V^s_{M_2}(O|t,o,\theta_{M_2,1}) &= ASC_{M_2,O} + \sum_{j=1}^{K_{M_2}} I_{M_2,O,j}\theta_{M_2,1,j} \sum_{k=1}^{188} I_{M_2,j,k} x_{k,t,o} \ , \\
V^s_{M_2}(DK|t,o,\theta_{M_2,1}) &= ASC_{M_2,DK} \ , \qquad (9)
\end{aligned}
$$

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

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

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, fast variations of facial characteristics could be perceived as surprise by respondents. $\{z_{kto}\}$ have been tested in the *reduced model*, but the associated parameters did not appear to be significant, certainly due to the simplistic assumption about the last frame triggering the expression choice. The detailed specification of $V^s_{M_2}(i|t,o,\theta_{M_2,1})\}$ is described in Tables 9 and 10. The reading of the tables is exactly the same as for Table 7 described in Section 3.1.

The utility function $V_{M_2}(i|t,o,\theta_{M_2,1},\alpha_{M_2,i})$ 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 $\alpha_{M_2,i}$, the parameter of memory effect. The specification of $V_{M_2}(i|t,o,\theta_{M_2,1},\alpha_{M_2,i})$ is defined in Equation (11).

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

Note that this is not anymore a linear-in-parameter specification for happiness, fear, sadness and anger, since $\{\alpha_i\}$ are estimated. Five memory effect parameters $\{\alpha_{M_2,i}\}_{i=SU,D,A,N,DK}$ have been fixed to 0 : 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 $ASC_{M_2,DK}$, 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 5 and Table 11). $\{\alpha_{M_2,i}\}_{i=H,F,SA,O}$ are supposed to be in the interval $[-1,1]$ because we hypothesize that the instantaneous perception of expression $i$ at time $t$ is more influenced by the instantaneous perception of expression $i$ at frame $t$ than at frame $t-1$. This dynamic specification has not been tested in the **reduced model**, as in this model we hypothesized that only the last frame of the video was triggering the expression (and not the two last frames). The model for $P_{M_2}(i|t,o,\theta_{M_2,1},\alpha_{M_2})$ is a logit model, that is

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

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 $\{y_{k,t,o}\}$ (see Equation (1)), and $\{z_{k,t,o}\}$ (see Equation (2)). We define $V_{M_2}(1|o,\theta_{M_2,2}) = 0$ and, for $t = 2,\ldots,T_o$,

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

and

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

$K^y_{M_2,2}$ and $K^z_{M_2,2}$ are 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

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

$I^z_{M_2,2,j,k}$ 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 12. Finally, the log-likelihood function is

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

### 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. In the first phase, the respondent is waiting for information, no perception of expressions is influencing the observed choice of label. 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 and 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 $\ell$ 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 $\theta_{M_3}$. $P_{M_3}(i|o,\theta_{M_3})$ is the average of $\{P_{M_3}(i|l,o,\theta_{M_3,1})\}_{l=t\dots T_o}$, weighted by $P_{M_3,n}(t|o,\theta_{M_3,2})$, sum up over $t=1\dots T_o$. We obtain

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

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).

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

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 13 and 14. Note that a dynamic formulation, as presented in Equation (11), has been tested in the expression utilities. It did not appear to be relevant, certainly due to the fact that the dynamics is already accounted for, by the consideration of the two phases. 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 $P_{M_3}(t|o,\theta_{M_3,2})$ captures the change 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. $\{y_{k,t,o}\}$ measure more drastic changes in the face compared to $\{z_{k,t,o}\}$ (see Section 5.3). The detailed specifications of the utilities are presented in Table 15. 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

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

### 3.4 Models with panel effect

The models presented in Sections 3.1, 3.2 and 3.3 do not account for the correlation between labels obtained through the internet survey. In this section, we assume that the labels are correlated through the filmed subject. Other panel structures have been tested (over respondents and videos) but this one appears to be the most relevant. Two models are developed based on the **reduced** and **latent** models.

#### 3.4.1 The reduced model with panel effect

This is a direct extension of the **reduced model** presented in Section 3.2. The utilities shown in equation 4 become

$$
\begin{aligned}
V_{M_4}(H|T_o,o,\theta_{M_4},\varepsilon_{M_4,s}) &= ASC_{M_4,H} + \sum_{j=1}^{K_{M_4}} I_{M_4,H,j}\theta_{M_4,j} \sum_{k=1}^{188} I_{M_4,j,k}x_{k,T_o,o} + \sum_{s=1}^{17} I_{o,s}\varepsilon_{M_4,s},\\
V_{M_4}(SU|T_o,o,\theta_{M_4,\varepsilon_{M_4,s}}) &= ASC_{M_4,SU} + \sum_{j=1}^{K_{M_4}} I_{M_4,SU,j}\theta_{M_4,j} \sum_{k=1}^{188} I_{M_4,j,k}x_{k,T_o,o} + \sum_{s=1}^{17} I_{o,s}\varepsilon_{M_4,s},\\
V_{M_4}(F|T_o,o,\theta_{M_4},\varepsilon_{M_4,s}) &= ASC_{M_4,F} + \sum_{j=1}^{K_{M_4}} I_{M_4,F,j}\theta_{M_4,j} \sum_{k=1}^{188} I_{M_4,j,k}x_{k,T_o,o} + \sum_{s=1}^{17} I_{o,s}\varepsilon_{M_4,s},\\
V_{M_4}(D|T_o,o,\theta_{M_4},\varepsilon_{M_4,s}) &= ASC_{M_4,D} + \sum_{j=1}^{K_{M_4}} I_{M_4,D,j}\theta_{M_4,j} \sum_{k=1}^{188} I_{M_4,j,k}x_{k,T_o,o} + \sum_{s=1}^{17} I_{o,s}\varepsilon_{M_4,s},\\
V_{M_4}(SA|T_o,o,\theta_{M_4},\varepsilon_{M_4,s}) &= ASC_{M_4,SA} + \sum_{j=1}^{K_{M_4}} I_{M_4,SA,j}\theta_{M_4,j} \sum_{k=1}^{188} I_{M_4,j,k}x_{k,T_o,o} + \sum_{s=1}^{17} I_{o,s}\varepsilon_{M_4,s},\\
V_{M_4}(A|T_o,o,\theta_{M_4},\varepsilon_{M_4,s}) &= ASC_{M_4,A} + \sum_{j=1}^{K_{M_4}} I_{M_4,A,j}\theta_{M_4,j} \sum_{k=1}^{188} I_{M_4,j,k}x_{k,T_o,o} + \sum_{s=1}^{17} I_{o,s}\varepsilon_{M_4,s},\\
V_{M_4}(N|T_o,o,\theta_{M_4},\varepsilon_{M_4,s}) &= 0\ ,\\
V_{M_4}(O|T_o,o,\theta_{M_4},\varepsilon_{M_4,s}) &= ASC_{M_4,O} + \sum_{j=1}^{K_{M_4}} I_{M_4,O,j}\theta_{M_4,j} \sum_{k=1}^{188} I_{M_4,j,k}x_{k,T_o,o} + \sum_{s=1}^{17} I_{o,s}\varepsilon_{M_4,s},\\
V_{M_4}(DK|T_o,o,\theta_{M_4},\varepsilon_{M_4,s}) &= ASC_{M_4,DK} + \sum_{s=1}^{17} I_{o,s}\varepsilon_{M_4,s},
\end{aligned}
\tag{24}
$$

where $\varepsilon_{M_4,s}$ is an error term capturing the correlation between observations associated to the filmed subject $s$. It is supposed normally distributed, $\varepsilon_{M_4,s} \sim N(0,\sigma_{M_4})$. $I_{o,s}$ is an indicator equal to 1 if the subject $s$ appears in video $o$, 0

otherwise. The probability of choosing the expression $i$ is

$$P_{M_4}(i|o,\theta_{M_4},\varepsilon_{M_4,s}) = \frac{e^{V_{M_4}(i|T_o,o,\theta_{M_4},\varepsilon_{M_4,s})}}{\sum_{j=1}^{9} e^{V_{M_1}(j|T_o,o,\theta_{M_4},\varepsilon_{M_4,s})}}. \tag{25}$$

Then, for the calculation of the log-likelihood, we have to integrate on $\varepsilon_{M_4,s}$

$$\mathcal{L}(\theta_{M_4},\sigma_{M_4}) = \sum_{s=1}^{17} log\left(\int (\prod_{o=1}^{O}\prod_{i=1}^{9} P_{M_4}(i|o,\theta_{M_4},\varepsilon_{M_4,s})^{w_{i,o}I_{o,s}}) f(\varepsilon_{M_4,s}) d\varepsilon_{M_4,s}\right), \tag{26}$$

where $f(\varepsilon_{M_4,s})$ is the probability density function (pdf) of $\varepsilon_{M_4,s}$.

### 3.4.2 The latent model with panel effect

This model generalizes the model proposed in Section 3.2. The utilities introduced in equation 11 are reformulated

$$\begin{aligned}
V_{M_5}(H|t,o,\theta_{M_5,1},\alpha_{M_5,H},\varepsilon_{M_5,s}) &= V^s_{M_5}(H|t,o,\theta_{M_5,1}) + \sum_{s=1}^{17} I_{o,s}\varepsilon_{M_5,s} \\
&+ \alpha_{M_5,H} V^s_{M_5}(H|t-1,o,\theta_{M_5,1}), \\
V_{M_5}(SU|t,o,\theta_{M_5,1},\alpha_{M_5,SU},\varepsilon_{M_5,s}) &= V^s_{M_5}(SU|t,o,\theta_{M_5,1}) + \sum_{s=1}^{17} I_{o,s}\varepsilon_{M_5,s}, \\
V_{M_5}(F|t,o,\theta_{M_5,1},\alpha_{M_5,F},\varepsilon_{M_5,s}) &= V^s_{M_5}(F|t,o,\theta_{M_5,1}) + \sum_{s=1}^{17} I_{o,s}\varepsilon_{M_5,s} \\
&+ \alpha_{M_5,F} V^s_{M_5}(F|t-1,o,\theta_{M_5,1}), \\
V_{M_5}(D|t,o,\theta_{M_5,1},\alpha_{M_5,D},\varepsilon_{M_5,s}) &= V^s_{M_5}(D|t,o,\theta_{M_5,1}) + \sum_{s=1}^{17} I_{o,s}\varepsilon_{M_5,s}, \\
V_{M_5}(SA|t,o,\theta_{M_5,1},\alpha_{M_5,SA},\varepsilon_{M_5,s}) &= V^s_{M_5}(SA|t,o,\theta_{M_5,1}) + \sum_{s=1}^{17} I_{o,s}\varepsilon_{M_5,s} \\
&+ \alpha_{M_5,SA} V^s_{M_5}(SA|t,o,\theta_{M_5,1}), \\
V_{M_5}(A|t,o,\theta_{M_5,1},\alpha_{M_5,A},\varepsilon_{M_5,s}) &= V^s_{M_5}(A|t,o,\theta_{M_5,1}) + \sum_{s=1}^{17} I_{o,s}\varepsilon_{M_5,s}, \\
V_{M_5}(N|t,o,\theta_{M_5,1},\alpha_{M_5,N},\varepsilon_{M_5,s}) &= V^s_{M_5}(N|t,o,\theta_{M_5,1}) = 0, \\
V_{M_5}(O|t,o,\theta_{M_5,1},\alpha_{M_5,O},\varepsilon_{M_5,s}) &= V^s_{M_5}(O|t,o,\theta_{M_5,1}) + \sum_{s=1}^{17} I_{o,s}\varepsilon_{M_5,s}
\end{aligned}$$

$$\begin{aligned} & + \quad \alpha_{M_5,O} V^s_{M_5}(O|t,o,\theta_{M_5,1}), \\ V_{M_5}(DK|t,o,\theta_{M_5,1},\alpha_{M_5,DK},\varepsilon_{M_5,s}) & = V^s_{M_5}(DK|t,o,\theta_{M_5,1}) + \sum_{s=1}^{17} I_{o,s}\varepsilon_{M_5,s}. \end{aligned} \tag{27}$$

where $\varepsilon_{M_5,s}$ is an error term capturing the correlation between observations implicating the same filmed subject $s$. $\varepsilon_{M_5,s}$ is supposed normally distributed, $\varepsilon_{M_5,s} \sim N(0,\sigma_{M_5})$. Note that $\{V^s_{M_5}(i|t,o,\theta_{M_5,1})\}$ are free of the error components, so there is no double counting of the error terms. The probability of choosing the expression $i$, within frame $t$ of video $o$ is

$$P_{M_5}(i|t,o,\theta_{M_5,1},\alpha_{M_5}) = \frac{e^{V_{M_5}(i|t,o,\theta_{M_5,1},\alpha_{M_5,i},\varepsilon_{M_5,s})}}{\sum_{j=1}^{9} e^{V_{M_5}(j|t,o,\theta_{M_5,1},\alpha_{M_5,j},\varepsilon_{M_5,s})}}, \tag{28}$$

and the probability of choosing the expression $i$ for video $o$ is

$$P_{M_5}(i|o,\theta_{M_5},\alpha_{M_5},\varepsilon_{M_5,s}) = \sum_{t=1}^{T_o} P_{M_5}(i|t,o,\theta_{M_5,1},\alpha_{M_5}) P_{M_5}(t|o,\theta_{M_5,2}), \tag{29}$$

where $P_{M_5}(t|o,\theta_{M_5,2})$ is the influence of the frame $t$ of video $o$ on the choice of label. It is the same than for the **latent model** (equation 14). The calculation of the log-likelihood function requires to integrate on $\varepsilon_{M_5,s}$

$$\mathcal{L}(\theta_{M_5},\alpha_{M_5},\sigma_{M_5}) = \sum_{s=1}^{17} log\left(\int (\prod_{o=1}^{O}\prod_{i=1}^{9} P_{M_5}(i|o,\theta_{M_5},\alpha_{M_5},\varepsilon_{M_5,s})^{w_{i,o}I_{o,s}}) f(\varepsilon_{M_5,s}) d\varepsilon_{M_5,s}\right), \tag{30}$$

where $f(\varepsilon_{M_5,s})$ is the pdf of the normal distribution $N(0,\sigma_{M_5})$.

# 4 Intermediary specification steps

The five models presented in this paper are the products of an extensive modeling process, where several intermediary models are generated. We started from the work of Sorci et al. (2010a) and first applied their model. The results were not satisfactory as the face positions and scales were not the same in the FEED and Cohn-Kanade databases as explained in Section 3.1. The model not only has been re-estimated, but also has been adapted due to the small number of observations available for this analysis. The obtained model is the **reduced model**, which has a strong basis for the development of the four other models. In the other

proposed models, the model managing the expression perception are extensions of this **reduced model**.

Next, we focus on the **latent model** (Section 3.1). In a first step, the model did not include any dynamics (Equation (11)). It came after the study of HMM and the dynamic formulation appeared to be meaningful in our work. Note that the accounting for previous probabilities instead of previous utilities was also tested, but it appeared heavy to manipulate. The incorporation of $\{z_{k,t,o}\}$ (Equation (2)) in the utility of surprise (Equation (9)) follows an analysis of the observed labels. Respondents have tendency to answer "surprise" when they perceive suddenness, and $\{z_{k,t,o}\}$ is well adapted to reflect it. Several specifications have been tested for the model managing the frame influence. We began with a model giving equal probabilities to all the frames, but the estimated results were not good. Regarding the frame utilities (Equation (13)), we started by incorporating only $\{x_{k,t,o}\}$ (Section 2), but parameters were not significant. We continued with models integrating only $\{y_{k,t,o}\}$ (Equation (1)), in order to account for the facial changes, which made a lot of sense. Finally we refined the model using both $\{y_{k,t,o}\}$ and $\{z_{k,t,o}\}$ in order to capture more details in the perception of the changes.

The study of the **latent model** predictions shows an instability of the model in case of several impressive frames (in the video) which are presenting different expressions. The improvement of this model passed by a smoothing of its behavior, by introducing the **smoothed model** (see Section 3.3). This new model is relevant for forecasting, because it is more robust to tiny fluctuations of facial descriptors, which can appear in noisy data. The dynamic formulation of the utilities has also been tested in the **smoothed model**, but it did not improve. It is certainly due to the fact that the dynamics is already accounted for with the assumption about the two behavioral phases (see Section 3.4.1). Several utility specifications of the model managing the phase changing have been tested (see Equation (20)). Only $\{z_{k,t,o}\}$ appeared to be significant, perhaps due to the fact that facial changes should be drastic for passing from one behavioral phase to the other.

We decided to refine the quality of the estimates for the **reduced** and **latent models** by accounting for the correlation between the observations in the data. It was not considered in the **smoothed** model due to practical difficulties in estimation, and modeling complexity. We obtained the **reduced** and **latent models with panel effect**. We sequentially considered the correlation per respondents, per videos and per filmed subjects. We retained this latter, the reasons are explained in Section 5.5. In this model, we additionally tried several specifications of the error term related to the panel effect. We tried to include *i.i.d.* and homoscedastic error terms in every alternatives (see Equation (24) and (27)). But the estimation results were similar. We retained the final specifications because it captured the panel effect as well as the correlation between all the alternatives, excepting the neutral. This mimics a nested structure, which makes sense as the neutral is the default expression.

# 5 Model estimation

The models are estimated by maximum likelihood (see Equations (7), (16), (23), (26) and (30)) using the biogeme software (Bierlaire, 2003; Bierlaire and Fetiarison, 2009). Except for the **reduced models** (with and without panel effect), these models are complex to estimate. The estimation results for the **latent** and **smoothed** models have been also obtained using codes based on biogeme. The estimation of models with panel effects required to perform numerical integration. A Monte-Carlo simulation with 1000 draws has been used. General estimation results are presented in Table 1.

## 5.1 The reduced model

The **Reduced model** is the simplest model as 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 7 and 8. Fourteen parameters are related to facial measurements characterizing AU (see Section 3.1). The parameter signs are consistent with the work of Sorci et al. (2010a), 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 $\geq 1.96$). 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 et al. (2010a).

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.

## 5.2 The latent model

For the **latent model**, the values and associated $t$-tests of the 34 parameters related to the model handling the expression perception are presented in Tables 9 and 10.

Signs and significance of parameters associated to AU, EDU and elements of the vector $C$ are correct and consistent with the estimated parameters obtained for the **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 the **reduced model**. This is due to the fact that the **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 indicating that the more the height of the mouth varies during the previous second, the more the surprise

will be favored, which is logical.

Four parameters of memory effect ($\alpha_{M_2,H}, \alpha_{M_2F}, \alpha_{M_2SA}, \alpha_{M_2O}$) appear to be significantly different from zero (see Table 11), and have the same magnitude. When unconstrained, their estimated values are in $[-1, 1]$ implying 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 12). 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 implies that the more a subject has the eye closed on a frame, the more the frame has influence on the observed choice of label.

### 5.3 The smoothed model

For the **smoothed model**, the model dealing with the perception of the expressions contains 36 parameters (see Tables 13 and 14).

Signs and significance of parameters associated to AU, EDU and C parameters are the same as that of the **reduced model**. This model contains 4 more parameters. $\theta_{M_3,1,4}$ and $\theta_{M_3,1,12}$ are respectively associated 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 the **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*"). They are included in the utility of surprise in order to capture the perception of suddenness. They are positive as expected, implying that the higher $z_{1,t,o}$ and $z_{3,t,o}$ are, 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 15). They are all linked with $\{y_{k,t,o}\}$. None of the parameters associated with $\{z_{k,t,o}\}$ appeared to be significant. The perception of the short time variations of facial characteristics is not relevant to activate the second phase of behavior, which seems logical. The change in the facial characteristics should be more drastic, which explains why $\{y_{k,t,o}\}$ are better adapted. As for the **latent model**, the magnitude of the parameters is larger than the parameters of the model handling the perception of the expressions. The interpretation remains the same as tht of the **latent model**.

### 5.4 Models with panel effect

For the models with panel effect, the parameters of the **reduced model with panel effect** are shown in Tables 16 and 17. The parameters of the **latent model with panel effect** are presented in Tables 18, 19, 20 and 21. In both cases, the parameter values are the same as that of **reduced** and **latent** models. Their interpretations remain unchanged. The standard errors $\sigma_{M_4}$ and $\sigma_{M_5}$ are significant, thereby verifying the hypothesis of correlation between labels associated to the same filmed subject.

### 5.5 Comparison of the five models

The final log-likelihood is improved between the **reduced** and **latent models**, and the **reduced** and **smoothed models**. The three first models cannot 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 for the models without panel effect, the **latent model** appears to be the best model, closely followed by the **smoothed model**. The improvement brought by the dynamic modeling is substantial. This is due to the nature of the videos (see Section 2). At the beginning of the videos, the facial expressions are neutral, and then they evolve towards other expressions, thereby making the faces distinclty expressive by the last frame of the video. This would explain why the **reduced models** tend to work well. Nevertheless, the proposed behavioral hypothesis makes sense. The assumption about one single frame triggering the choice seems to be the most relevant (**latent model**), closely followed by the assumption about the two behavioral phases (**smoothed model**). This order seems to be logical as the **latent model** focuses on a "pure" and "strong" perception, which is intuitively important, especially in short facial videos. Compared to this latter model, the **smoothed model** polishes the perceptions in the second behavioral phase. The main advantage of the **smoothed model** is its less sensitivity to data errors, compared to the **latent model**.

For the models with panel effect, the log-likelihood is improved between the **reduced model** and **reduced model with panel effect**, and the **latent model** and **latent model with panel effect**. Out of the five proposed models, the **latent model with panel effect** is the best in terms of fit. The accounting for the correlations between observations related to the same filmed subject, improves significantly the fit. A correlation by respondents has been tested but it did not appear to meaningful as the perception of the respondents seems to be homogeneous. This is logical as the respondents are also homogeneous in terms of socio-economic characteristics (they are mainly in Switzerland with an academic background). A correlation per videos has been checked and it showed similar results for the correlation per filmed subjects. It has not been kept because this model was very heavy to manipulate. Since the number of videos is higher than the number of filmed subjects, we increased the number of draws, resulting in a

Table 1: General estimation results

| | **Reduced** | **Latent** | **Smoothed** | **Reduced panel** | **Latent panel** |
|---|---|---|---|---|---|
| Nb obs. | 369 | 369 | 369 | 369 | 369 |
| Nb param. | 32 | 45 | 44 | 33 | 46 |
| Null $\mathcal{L}$ | -810.78 | -810.78 | -810.78 | -810.78 | -810.78 |
| Final $\mathcal{L}$ | -475.79 | -441.28 | -447.67 | -470.26 | -435.14 |
| $\bar{\rho}^2$ | 0.374 | 0.400 | 0.394 | 0.379 | 0.406 |

dramatic increase of the estimation time, making the cross-validation (see Section 6.2) impossible.

The magnitude of the parameter values and signs are the same for the five models. For example, $\theta_{M_1,4}$, $\theta_{M_2,1,4}$, $\theta_{M_3,1,5}$, $\theta_{M_4,4}$ and $\theta_{M_5,1,4}$ are associated to the mouth opening ("*RAP_mouth*"), defined as the fraction between the height of the mouth ("*mouth_h*") and the width of the mouth ("*mouth_w*"). These paameters are in the utilities of surprise and fear. The associated parameters are all positive, indicating the stability of the models. Their positive sign is logical because when a person opens the mouth, the perceived facial expression is likely to be fear or surprise.

The specifications of the model related to the detection of the most impressive frame in the **latent models** (with and without panel effect) and to the detection of the first frame of the relevant group of frames in the **smoothed model** are quite similar. For the **latent models**, it contains parameters associated with both $\{y_{k,t,o}\}$ and $\{z_{k,t,o}\}$ and for the **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 8 displays the variation of this feature among frames of a video which are displayed at the top. The sign of the parameters associated with $y_{2,t,o}$ ($\theta_{M_2,2,6}$, $\theta_{M_5,2,6}$ 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.

In conclusion, the parameters of the models are significant and interpretable. Moreover, the addition of a dynamic feature in the models significantly improves the fit. The accounting of the panel effect is successful as the latent model with panel effect has the best fit.

# 6 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 for estimation (see Section 5). We proceed in three steps: the first one consists of comparing the percentages of badly predicted observations for the proposed models. In the second

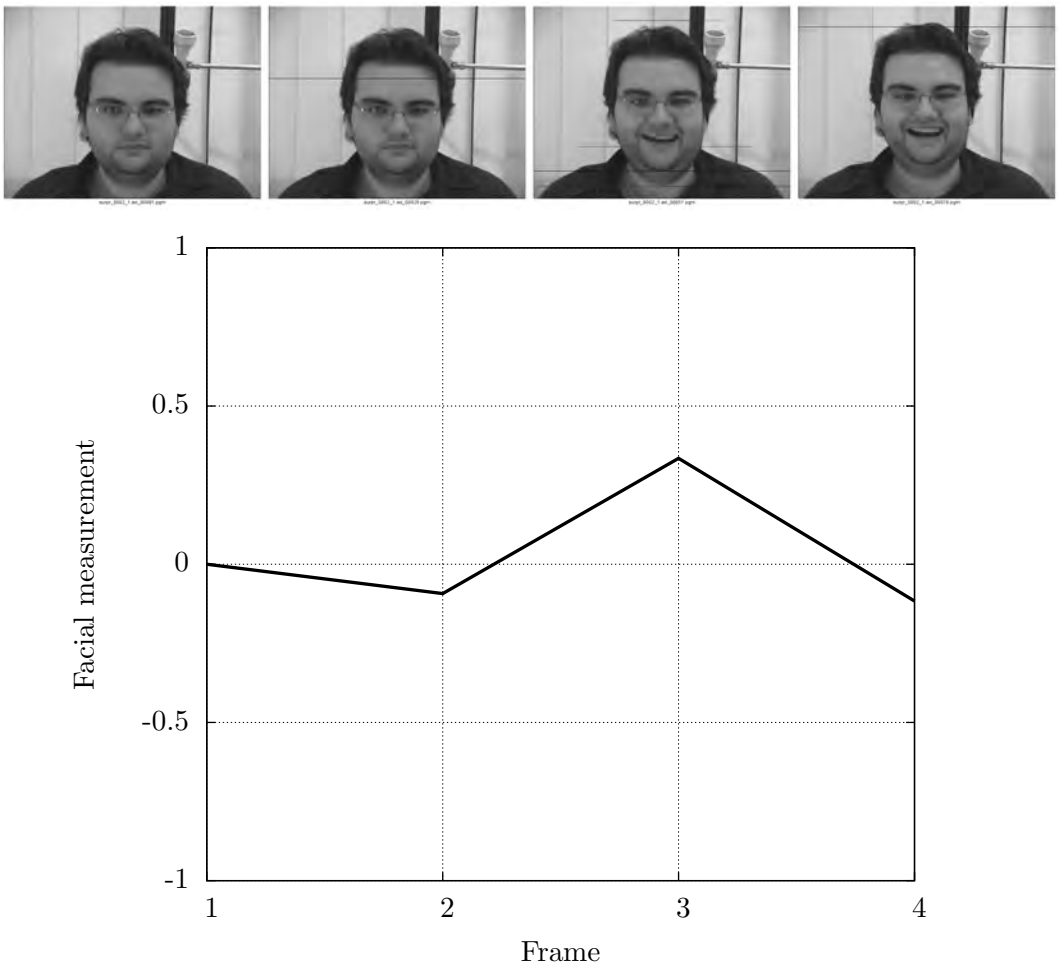



Figure 8: Examples of the variations of $y_{2,t,o}$, associated to the height of the mouth ("*mouth_h*"), for a video

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 analyzing the predictions of the models in detail.

## 6.1 Aggregate prediction

An observation is considered to be poorly 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 percentage of poorly predicted observations per model. The percentages are consistent with the fitting results presented in Section 5, which is good sign. The percentage of poorly predicted observations is already low for the **reduced model**. The

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

| **Reduced** | **Latent** | **Smoothed** | **Reduced panel** | **Latent panel** |
|---|---|---|---|---|
| 17.89 | 17.34 | 15.45 | 18.43 | 14.45 |

improvement brought by the **latent** and **smoothed models** compared to the **reduced model** is minor in terms of prediction. This can be explained by the structure of the considered facial videos. Since the "peak" emotion is often observed at the end of the video, there are fewer observations where the dynamic models could out perform. However the **latent model with panel effect** is the best.

The cumulative distributions of the choice probabilities predicted by the models are displayed in Figure 9. If the models were perfect, the curves would have been flat with a peack for choice probabilities equal to one. This would mean that the models replicate the observed choices of labels exactly. Of course this is not the case. The five curves are close in the "poorly predicted" interval (choice probabilities less than $\frac{1}{9} = 0.11$). This is consistent with the results shown in Table 2. In the interval $[0, 0.78]$ the **latent model with panel effect** performs the best. In the last interval, it is the **latent model** that predicts the highest probabilities (its curve is the last to reach the level of one). The **smoothed model**, is better than the **reduced model**, except on $[0.68, 1]$. Moreover the **latent model with panel effect** is always better than the **reduced models** (with and without panel effect), which demonstrates the added value of the dynamic modeling.

## 6.2 Cross-validation

The study of the poorly predicted observations, described in Section 6.1, 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 model quality should be tested on some new data, but we do not have access to such data. In this situation, the cross-validation allows us to validate the models. The methodology is inspired from the work of Robin et al. (2009) that successfully cross-validates a model for pedestrian behavior. The dataset is split into an estimation and a validation subset. The dataset is randomly split across the videos into five equal subsets of 13 videos out of 65. Four subsets are combined into the estimation dataset. After estimation, the model is applied on the remaining subset. This operation is repeated five times. The percentages of poorly predicted observations, calculated over the validation subsets are presented in Table 3.

For the models without panel effect, the two dynamic models (the **latent** and **smoothed models**) are always better than the **reduced model**. In addition, the percentages of badly predicted observations are close to those obtained on the entire estimation data (see Table 2) for the **latent** and **smoothed models**, but

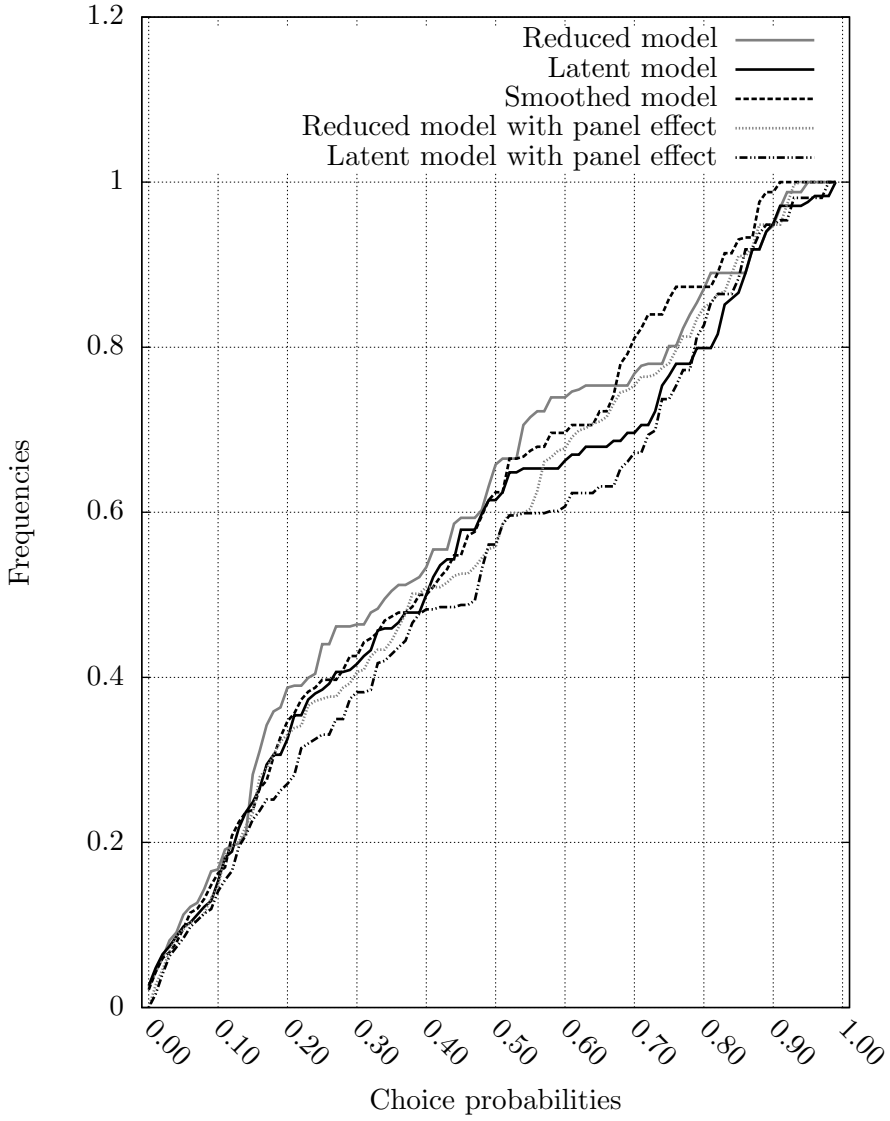



Figure 9: Cumulative distributions of the choice probabilities predicted by the five proposed models, on the estimation data

not the **reduced model**. The dynamic models appear to be much more robust than the **reduced model**. This justifies the approach and the validity of the dynamic models.

For the models with panel effect, results for the **reduced model with panel effect** are worse than the **reduced model**. This is also the case for the **latent model with panel effect** compared to the **latent model**. Note that for experience 4, the estimation of the **latent model with panel effect** did not converge

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

| Validation subsets | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| **Reduced** | 28.74 | 26.15 | 21.31 | 21.87 | 28.26 |
| **Latent** | 24.14 | 13.85 | 11.48 | 17.19 | 21.74 |
| **Smoothed** | 20.69 | 16.92 | 18.03 | 15.63 | 10.87 |
| **Reduced panel** | 28.73 | 26.15 | 22.95 | 23.43 | 28.26 |
| **Latent panel** | 28.70 | 15.38 | 21.29 | $*$ | 35.87 |

(this model is very difficult to estimate due to its complexity). We conclude that the two models with panel effect tend to over-fit the data.

### 6.3 Disaggregate prediction

We looked at the power of prediction over the estimation dataset, at the aggregate level. The study of a particular video allows us to precisely describe the predictions of the five models. The video is the same as the one considered in Figure 8. The detailed predictions of the models are shown in Figure 10 for the **reduced model**, Figure 11 for the **latent model**, Figure 12 for the **smoothed model**, Figure 13 for the **reduced model with panel effect** and Figure 14 for the **latent model with panel effect**. On these figures, each column is associated to a frame, except the one of the extreme right. The top row displays the considered frames. As mentioned in Section 2, each frame is the first of a group of images corresponding to one second of time in a video. The second row reflects the predictions of the model associated to the perception of the expressions. For each frame, the probability distribution of the expressions is shown. The third row shows the influence of the frames. The contributions of the frames sum up to one. For the **reduced models** (with and without panel effect), only the last frame is considered relevant. Therefore the peak is justified on this last frame. For the **latent models** (with and without panel effect), it shows the influence of each frame on the final expression choice. For the **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, we can 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 Figure 8), 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 retained 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. The **reduced model** predicts 65% for happiness, 35% for surprise, and 0% for anger. The prediction seems logical regarding only the facial characteristics in the last frame.

The **latent model** predicts 24% for happiness, 58% for surprise, 18% for disgust and 0% for anger. This is further away from the distribution of the collected labels, compared to the **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 8 for the height of the mouth). For this frame, the predicted probability of surprise is high. This is logical, because the utility of surprise contains $\{z_{k,t,o}\}$ (see Equation (9)), which account for the perception of suddenness. 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.

The **smoothed model** predicts 58% for happiness, 38% for surprise, 4% for disgust and 0% for 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 the **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 high for this frame (see Figure 8). The model handling 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 consistent. 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 results are rather the same for the **reduced model with panel effect** and the **reduced model**. Regarding the **latent model with panel effect**, it predicts 35% of happiness, 52% of surprise, 13% of disgust and 0% of anger. The model has selected the frame 3 as being the most influential. Even if the results are quite similar compared to those obtained with the **latent model**, they are better because the difference between the predicted probabilities of happiness and surprise are smaller.

The predictions of the five models are explainable. The **smoothed model** seems to be the most interpretable. The **smoothed model** and **latent model with panel effect** predict the closest distributions of probability across the expressions to the collected labels. The **smoothed model** over-predicts happiness and under-predicts surprise, contrary to the **latent model with panel effect**.

# 7 Conclusions and Perspectives

In this paper, we propose a new approach of the dynamic facial expression recognition. The estimation of the models is based on labels collected through respondents of an internet survey. The developed models capture causal effects between facial characteristics and expressions. Statistical tests and model predictions have proved the model performance, and the added value of the dynamic formulation (the **latent models** and the **smoothed model** compared to the **reduced models**). In terms of fit, the **latent model with panel effect** is the best. The five models have been cross-validated on the estimation data. The **latent model** and the **smoothed model** appear to be more robust than the **reduced model**. The models with panel effect over fit the data. Consequently, they cannot be used for forecasting. Finally, some qualitative analysis of the model predictions allow us to confirm the modeler's intuition about the facial video. Regarding all the analysis, the **smoothed model** seems to be the most robust.

The proposed work overcomes the limitations of the standard approaches in the dynamic facial expression recognition. Standard approaches consist of associating any two examples with the same facial descriptors to the same expression. One of the main assumption is that facial expression labels, which are in the data, stand for the true expressions (Cohen et al., 2003; Bartlett et al., 2003). But this assumption does not hold in reality, as people can perceive the same expression differently. Facial expressions are characterized by the ambiguity. In our work, this ambiguity is directly taken into account as we have adopted a probabilistic approach. Another limitation of the previous approaches is the inability for interpreting the knowledge acquired by the systems. They are often black-boxes, where the interpretation of the links between the inputs (facial descriptors) and the output (expression) are not possible. Due to this black-box nature, it is also impossible to put knowledge in the model to improve it further. In our proposed work, psychological concepts are translated into mathematical equations, and the maximum likelihood estimation allows to confirm (or negate) the quality of the model. In addition, this allows us to learn the behavioral patterns contained in the data. In particular, we have quantified the concepts introduced by Ekman and Friesen (1978).

We generalize the work of Sorci et al. (2010a), as we worked with facial videos and not with images. More generally, this work is the first attempt for analyzing videos using discrete choice models.

Regarding discrete choice modeling, we have developed models inspired from recent works (Ben-Akiva, 2010). Original formulations have been introduced to capture the dynamics, which can be reused for other analysis. The proposed models are based on different assumptions about video perceptions. The estimation, validation and comparison of the models underline the relevance of these assumptions. We learned that respondents have tendency to make their expression choices when watching specific frames and we can select these frames in different ways. Our approach explains and quantifies these psychological concepts.

As such, these models can be used directly for applications. The major difficulty concerns the computation of the variables. The quality of the considered videos should be quite high, in terms of definition and size of the face. 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) but remain quite general. Some case studies have to be conducted in order to completely prove the model applicability to transportation (Denis, 2009).

In the context of "Aware" vehicles, we think of a system that would be able to manage automatically the interior features of the car, based on the driver's characteristics, including the facial expression. In case of dedicating the proposed model for this application, a data collection in two stages should be performed. In the first stage, we can conduct a survey in a car simulator by placing the respondents in controlled real driving situations and recording their faces. Then, the respondents would be asked to perform actions using the interior car features. In the second stage, the collected facial videos would be labeled using a similar survey to the proposed internet survey. A model handling the choice of action using the driver's characteristics and expression as inputs, can be developed. Then, in a real context, the face of a driver can be monitored with a camera and the proposed model applied.

The proposed models may be also used to analyze travelers satisfaction with public transportation (Friman and Garling, 2001). The facial expression could be used as a measure of satisfaction when conducting transportation surveys. For on-site measures, it is not worth as the facial expressions are most of the time generated by stimuli not related to transportation. The experimental design of the survey should be carefully set in order to use adapted stimuli.

More generally, for the estimation of hybrid choice models, some indicators of the latent variables are needed. Bolduc and Alvarez-Daziano (2010) propose an hybrid choice model handling the vehicle choice. In that case, the facial expression of the survey respondent could be used as an indicator of the two latent variables: "Environmental concern" and "Appreciation of new car features". In addition to the rational behavior, the latent variables capture the emotional states. The facial expression results from a short emotion and it could be used as a proxy of this emotion, or combined with other emotion indicators (questionnaires, for example) to reveal it. Practically, in addition to the questionnaires, some well-chosen stimuli should be shown to the survey respondents, such as short and impacting environmental documentaries, or advertisements of cars having new features, while their faces are recorded. Then, a DFER model is needed to determine the facial expressions. For this application, the facial expression is not an input of the prediction process, but it helps to reinforce the quality of the estimated model.

Finally in the marketing context, MacInnis et al. (1991) studied the ability of individuals to process the brand information from advertisements. The facial expression could be used as inputs for the model, in addition to eye-tracking data.

Even if this new modeling framework is meaningful, there is some scope for

improvements. The model has been estimated on a small dataset. More observations would help. The number and type of videos is also a critical aspect. Feature variabilities are quite low and should be increased. This would allow us to have a more complete specification of the utilities. We could use the specification proposed by Hensher (2010) for the processing of explanatory variables, which is highly relevant due to large amount of information provided on a face. In addition, more complex structures could be tested for the choice models. In the **latent** and **smoothed models**, the model handling the detection of the most attractive frame and the first frame of the relevant group of frames can be modified for considering the correlation between frames. A cross-nested logit seems to be well adapted to the frame choice, using two nests: "attractive" and "not attractive". Each frame could belong to the two nests. The membership degrees of the frame to each nest should be defined as a function of their attractiveness. They cannot be generic, as the videos are varying from one observation to the other and the associated frame set. This stands as a research topic on its own. Finally, a comparison with the state of the art machine learning method, such as neural networks (NN) or hidden markov models (HMM), would be interesting.

## Acknowledgments

We are very grateful to Dr. Matteo Sorci who provided the necessary codes to extract facial features using AAM. We thank Dr. Prem Kumar Viswanathan who helped in improving the paper editing.

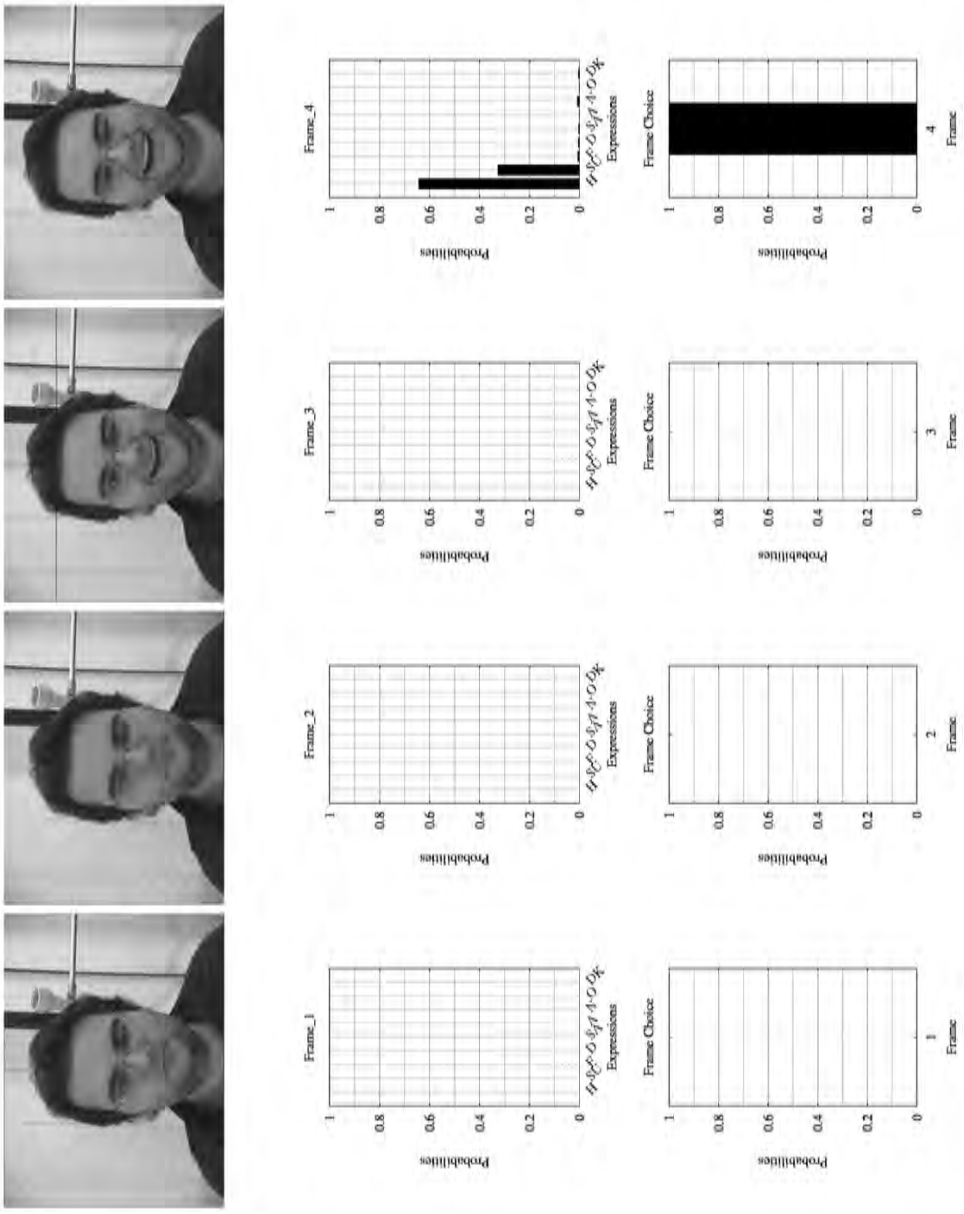



Figure 10: Example of a detailed prediction of the **reduced model**

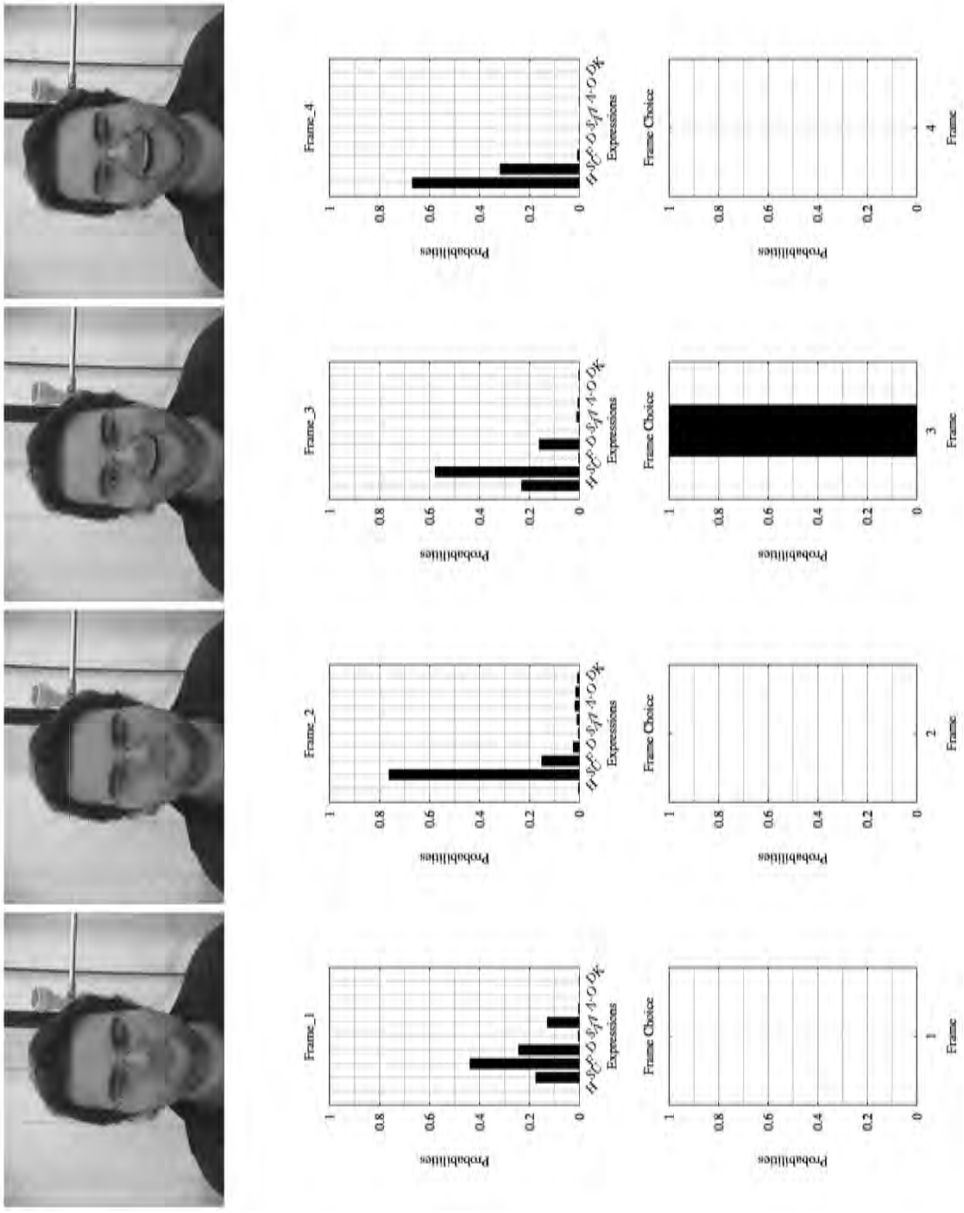



Figure 11: Example of detailed prediction of the **latent model**

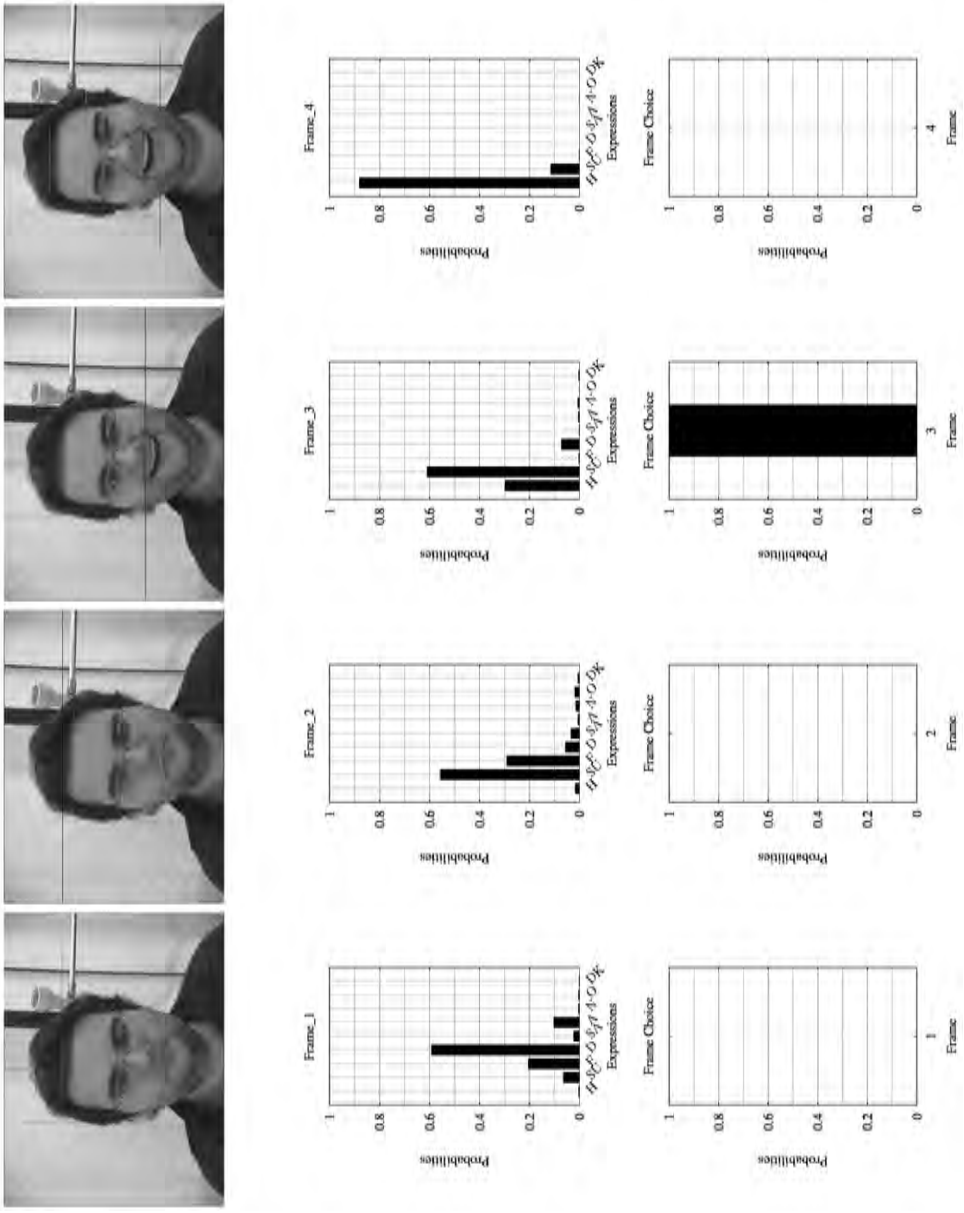



Figure 12: Example of detailed prediction of the **smoothed model**

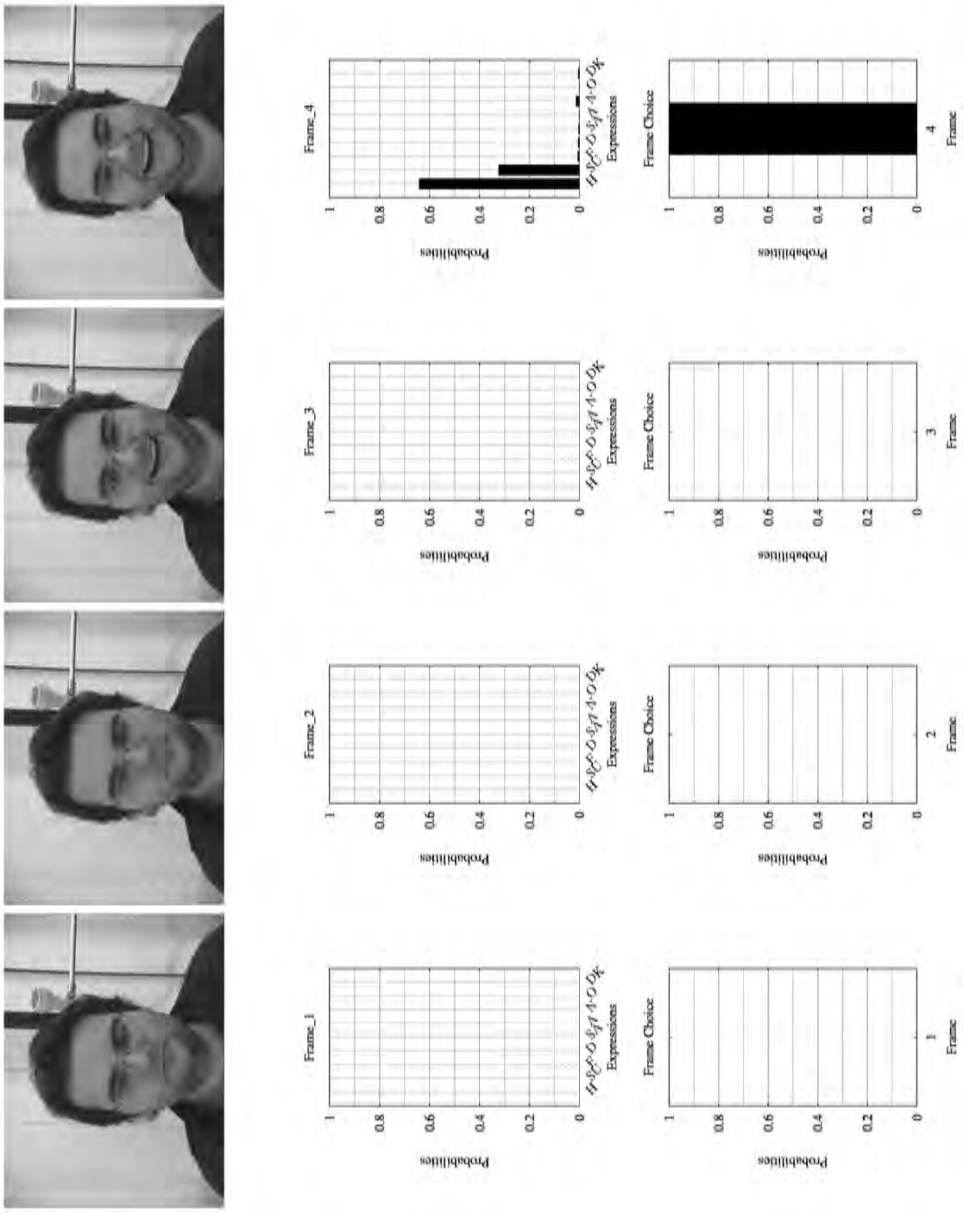



Figure 13: Example of detailed prediction of the **reduced model with panel effect**

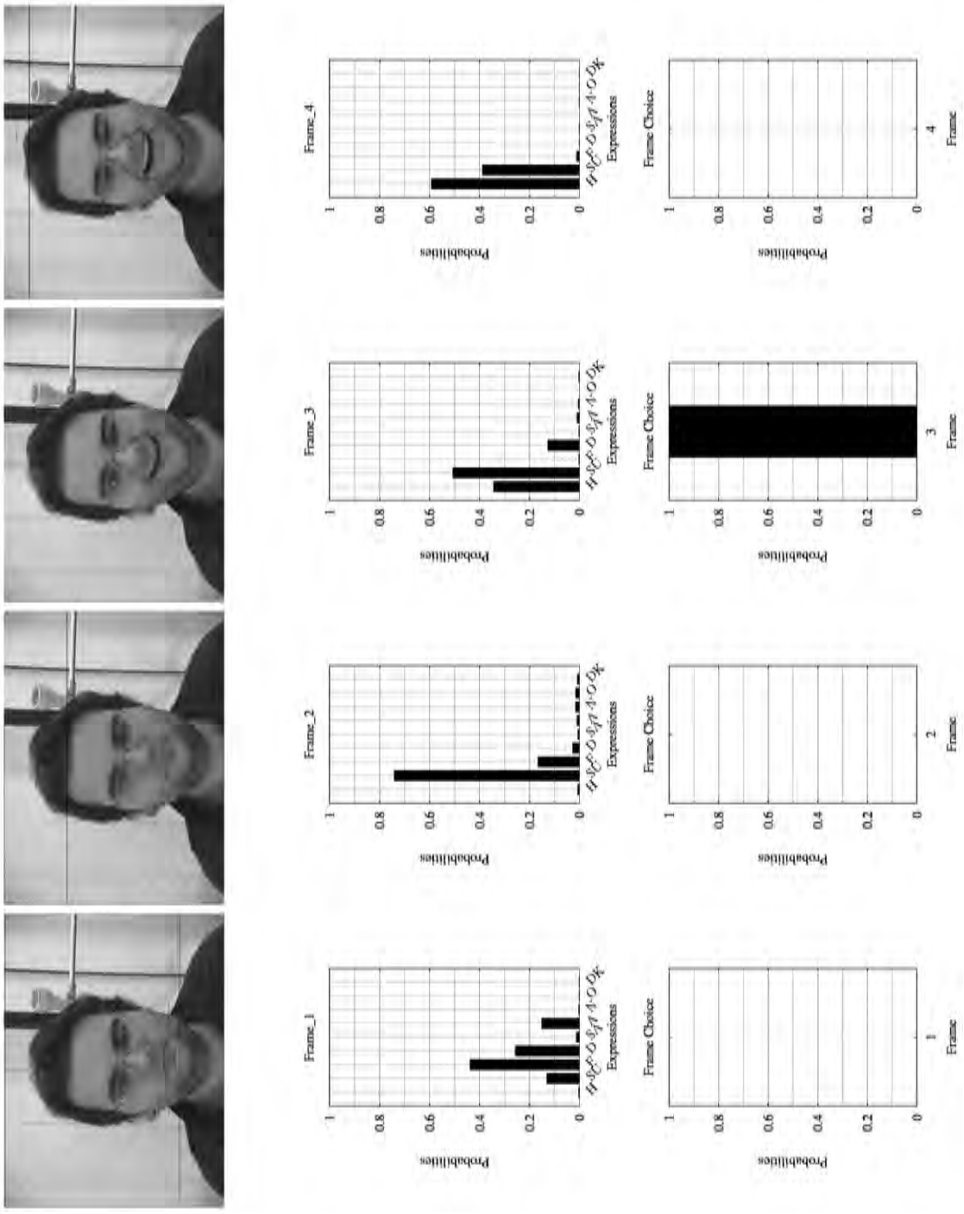



Figure 14: Example of detailed prediction of the **Latent model with panel effect**

## References

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

Antonini, G., Sorci, M., Bierlaire, M., Thiran, J., 2006. Discrete choice models for static facial expression recognition. In: Blanc-Talon, J., Philips, W., Popescu, D., Scheunders, P. (Eds.), 8th International Conference on Advanced Concepts for Intelligent Vision Systems. Vol. 4179 of Lecture Notes in Computer Science. Springer Berlin / Heidelberg, Berlin, pp. 710–721.

Bartlett, M. S., Littlewort, G., Fasel, I., 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 5, 53.

Ben-Akiva, M., 2010. Planning and action in a model of choice. In: Hess, S., Daly, A. (Eds.), Choice modelling: the state-of-the-art and the state-of-practice. Emerald, pp. 19–34.

Ben-Akiva, M., Mcfadden, D., Train, K., Walker, J., Bhat, C., Bierlaire, M., Bolduc, D., Boersch-Supan, A., Brownstone, D., Bunch, D. S., Daly, A., De Palma, A., Gopinath, D., Karlstrom, A., Munizaga, M. A., 2002. Hybrid choice models: Progress and challenges. Marketing Letters 13, 163–175.

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

Bierlaire, M., Chen, J., Newman, J. P., 2010. Using location observations to observe routing for choice models. In: Proceedings of the 89th Transportation Research Board Annual Meeting. Washington D.C., US.

Bierlaire, M., Fetiarison, M., 2009. Estimation of discrete choice models: extending biogeme. In: Proceedings of the 9th Swiss Transport Research Conference. Ascona, Switzerland.

Bierlaire, M., Frejinger, E., 2008. Route choice modeling with network-free data. Transportation Research Part C 16 (2), 187–198.

Bolduc, D., Alvarez-Daziano, R., 2010. On estimation of hybrid choice models. In: Hess, S., Daly, A. (Eds.), Choice modelling: the state-of-the-art and the state-of-practice. Emerald, pp. 259–288.

Choudhury, C. F., 2007. Model driving decisions with latent plans. Ph.D. thesis, Massachusetts Institute of Technology.

Cohen, I., Sebe, N., Garg, A., Chen, L. S., 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., 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. Tech. rep., Transport and Mobility Laboratory (TRANSP-OR), EPFL, EPFL ENAC

INTER TRANSP-OR, Station 18, CH-1015 Lausanne, Switzerland.

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

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

Friman, M., Garling, T., 2001. Frequency of negative critical incidents and satisfaction with public transport services. Journal of Retailing and Consumer Services 8 (2), 105 – 114.

Hensher, D. A., 2010. Attributes processing, heuristics and preference construction in choice analysis. In: Hess, S., Daly, A. (Eds.), Choice modelling: the state-of-the-art and the state-of-practice. Emerald, pp. 35–70.

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

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

Lerner, J. S., Keltner, D., 2000. Beyond valence: Toward a model of emotion-specific influences on judgement and choice. Cognition and Emotion 14, 473 – 493.

MacInnis, D. J., Moorman, C., Jaworski, B. J., 1991. Enhancing and measuring consumers' motivation, opportunity, and ability to process brand information from ads. The Journal of Marketing 55 (4), 32–53.

Mellers, B. A., McGraw, A. P., 2001. Anticipated emotions as guides to choice. Current Directions in Psychological Science 10 (6), 210–214.

Miwa, H., Itoh, K., Matsumoto, M., Zecca, M., Takanobu, S., Rocella, S., Carrozza, P., Dario, A., A., T., 2004. Effective emotional expressions with emotion expression humanoid robot we-4rii - integration of humanoid robot hand rch-1. In: International Conference on Intelligent Robots and Systems. Vol. 3. pp. 2203–2208.

Pantic, M., 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.

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

Small, D., 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., Thiran, J.-P., 2010a. 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., Antonini, G., 2010b.

Capturing human perception of facial expressions by discrete choice modelling. In: Hess, S., Daly, A. (Eds.), Choice Modelling: The State-of-the-Art and the State-of-Practice. Emerald Group Publishing Limited, pp. 101–136.

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

Wallhoff, F., 2004. Fgnet-facial expression and emotion database. Tech. rep., Technische Universität München.

Weinberg, P., Gottwald, W., 1982. Impulsive consumer buying as a result of emotions. Journal of Business Research 10 (1), 43 – 57.

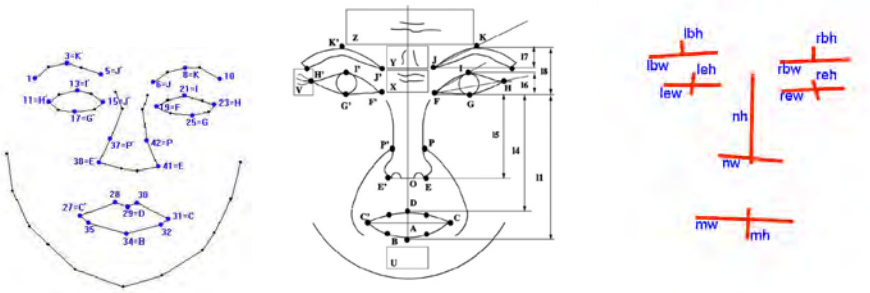
# A Notations

Table 4: Summary of the mentioned acronyms

| **Acronym** | Definition |
|---|---|
| A | Anger |
| AAM | Active appearance model |
| ASC | Alternative specific constant |
| AU | Action unit |
| D | Disgust |
| DCM | Discrete choice model |
| DFER | Dynamic facial expression recognition |
| DK | I don't know |
| EDU | Expression descriptive unit |
| EDU_6 | EDU defined as the ratio between the average of the eye width and the mouth width |
| EDU_8 | EDU defined as the ratio between the average of the eyes height and the average of the brow-eyes height |
| F | Fear |
| FER | Facial expression recognition |
| FACS | Facial action coding system |
| FEED | Facial expression and emotion database |
| HMM | Hidden markov model |
| H | Happiness |
| MEV | Multivariate extreme value |
| N | Neutral |
| NN | Neural networks |
| O | Other |
| PCA | Principal component analysis |
| RAP_brow | EDU defined as the ratio between the average of the brow-eyes height and the brow-eyes width |
| RAP_mouth | EDU defined as the ratio between the height and the width of the mouth |
| SA | Sadness |
| SFER | Static facial expression recognition |
| SU | Surprise |
| SVM | Support vector machine |

# B Detailed description of the explanatory variables

The details of the mask extracted using an AAM, are shown in Figure 15(a), as well as the geometrical relationship of the facial measure points (Figure 15(b)) and some facial descriptors (Figure 15(c)). The correspondences between the measures on the mask displayed in Figure 15(b) and the mask presented in Figure 15(c), are shown in Table 5.




(a) (b) (c)

Figure 15: a) Facial landmarks (55 points); b) the geometrical relationship of facial feature points; c) some facial descriptors;

Different explanatory variables based on the outputs of the AAM, are used to reflect the perception of facial expressions. They are coming from the facial action coding system (FACS); they are expression descriptive units (EDU), and also C parameters.

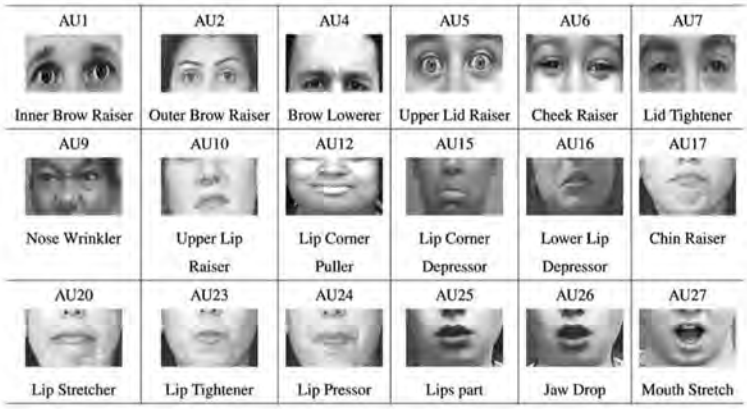



Figure 16: Sample of AU

The FACS associates tensenesses and relaxations of muscles to each expression. They call them action units (AU). A sample of AU is presented in Figure 16. For

Table 5: Correspondences between measures on masks 15(b) and 15(a)

| **FACS Measures** | **Measures on mask** 15(a) |
|---|---|
| $\overline{JJ'}$ | $Dist(P5, P6)$ |
| $\overline{JF}$ | $Dist(P6, P19)$ |
| $\overline{J'F'}$ | $Dist(P5, P15)$ |
| $\overline{KG} \equiv l8$ | $Dist(P8, P25)$ |
| $\overline{K'G'}$ | $Dist(P3, P17)$ |
| $\overline{GI} \equiv l6$ | $Dist(P25, P21)$ |
| $\overline{G'I'}$ | $Dist(P13, P17)$ |
| $\overline{PF}$ | $Dist(P19, P42)$ |
| $\overline{P'F'}$ | $Dist(P15, P37)$ |
| $\overline{FC}$ | $Dist(P19, P31)$ |
| $\overline{F'C'}$ | $Dist(P15, P27)$ |
| $\overline{FD} \equiv l4$ | $Dist(P25, P29)$ |
| $\overline{F'D}$ | $Dist(P17, P29)$ |
| $\overline{OD}$ | $Dist(\left(\frac{P39+P40}{2}\right), P29)$ |
| $\overline{OB}$ | $Dist(\left(\frac{39+40}{2}\right), 33)$ |
| $\overline{DB}$ | $Dist(P29, P33)$ |
| $\overline{C'C}$ | $Dist(P27, P31)$ |
| $\measuredangle FHJ$ | $\measuredangle P19P23P6$ |
| $\measuredangle F'H'J'$ | $\measuredangle P15P11P5$ |
| $\measuredangle HFI$ | $\measuredangle P23P19P21$ |
| $\measuredangle H'F'I'$ | $\measuredangle P11P15P13$ |
| $\measuredangle HGF$ | $\measuredangle P23P25P19$ |
| $\measuredangle H'G'F'$ | $\measuredangle P15P17P11$ |

example AU 6 is associated to happiness. The details of these associations are presented in Ekman and Friesen (1978). We translate the facial distances and angles extracted from the mask, into AU.

EDU are reported in Table 6 and introduced by in Antonini et al. (2006). Additionally to the FACS, they account for the interactions between facial descriptors. The first 5 EDU represent, respectively, the eccentricity of eyes, left and right eyebrows, mouth and nose. The EDU from 7 to 9 represent the eyes interactions with mouth and nose, while the $10^{th}$ EDU is the nose-mouth relational unit. The last 4 EDU relate the eyebrows to mouth and nose. The EDU can be intuitively interpreted. For example, in a face displaying a surprise expression, the eyes and the mouth are usually opened and this can be captured by EDU7 $(eye_{height}/mouth_{height})$.

Another vector $C$ of values capturing both the facial texture and shape is also

Figure 17: Examples of synthesized faces obtained varying the first $C$ parameter from the mean face ($\pm 3std$).

Table 6: Expressions Descriptive Units

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

generated by the AAM. FACS and EDU provide measures of local facial features but they do not provide a description of a face as a global entity. This information can be obtained considering the appearance vector $C$ matching the face in the processed image. Figure 17 shows the effect of varying the first appearance model parameter, showing changes in identity and expression.

# C Estimation results

Table 7: 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 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $ASC_{M_1,A}$ | | | | | | × | | | | 1 | 0.95 | 0.28 |
| $ASC_{M_1,D}$ | | | | × | | | | | | 1 | 25.38 | 7.88 |
| $ASC_{M_1,DK}$ | | | | | | | | | × | 1 | -0.69 | -1.79 |
| $ASC_{M_1,F}$ | | | × | | | | | | | 1 | 0.49 | 0.19 |
| $ASC_{M_1,H}$ | × | | | | | | | | | 1 | -3.14 | -0.79 |
| $ASC_{M_1,O}$ | | | | | | | | × | | 1 | 6.95 | 3.20 |
| $ASC_{M_1,SA}$ | | | | | × | | | | | 1 | 10.80 | 2.54 |
| $ASC_{M_1,SU}$ | | × | | | | | | | | 1 | -11.27 | -5.63 |

Table 8: Estimation results and description of the specification of **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 |
| $\theta_{M_1,23}$ | × | | | | | | | | | mouth_nose_dist | -236.37 | -5.65 |
| $\theta_{M_1,24}$ | × | | | | | | | | | mouth_w | 188.42 | 4.90 |

Table 9: Estimation results of the constants for the **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 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $ASC_{M_2,A}$ | | | | | | × | | | | 1 | -5.86 | -1.31 |
| $ASC_{M_2,D}$ | | | | × | | | | | | 1 | 22.73 | 4.48 |
| $ASC_{M_2,DK}$ | | | | | | | | | × | 1 | -0.71 | -1.83 |
| $ASC_{M_2,F}$ | | | × | | | | | | | 1 | -4.55 | -1.13 |
| $ASC_{M_2,H}$ | × | | | | | | | | | 1 | 3.02 | 0.22 |
| $ASC_{M_2,O}$ | | | | | | | | × | | 1 | 14.44 | 4.22 |
| $ASC_{M_2,SA}$ | | | | | × | | | | | 1 | 8.54 | 1.57 |
| $ASC_{M_2,SU}$ | | × | | | | | | | | 1 | -25.69 | -7.08 |

Table 10: Estimation results and description of the specification of the **latent 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 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\theta_{M_2,1,1}$ | | | | × | | | | | | 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 |
| $\theta^z_{M_2,1,1}$ | | | | | | | | | | mouth_h, $z_{1,t,o}$ | 50.21 | 3.04 |

Table 11: Estimation results of the **latent model**, associated to the memory effects parameters

| parameter | value | $t$-test 0 |
|---|---|---|
| $\alpha_{M_2,H}$ | -0.62 | -8.18 |
| $\alpha_{M_2,F}$ | -0.33 | -2.73 |
| $\alpha_{M_2,SA}$ | -0.46 | -2.04 |
| $\alpha_{M_2,O}$ | -0.70 | -2.68 |

Table 12: Estimation results and description of the specification of the **latent model**, associated to the model which detects the most meaningful frame

| 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 13: Estimation results of the constants for the **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 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $ASC_{M_3,A}$ | | | | | | × | | | | 1 | -7.53 | -1.63 |
| $ASC_{M_3,D}$ | | | | × | | | | | | 1 | 20.28 | 4.03 |
| $ASC_{M_3,DK}$ | | | | | | | | | × | 1 | -0.69 | -1.79 |
| $ASC_{M_3,F}$ | | | × | | | | | | | 1 | -0.35 | -0.09 |
| $ASC_{M_3,H}$ | × | | | | | | | | | 1 | -7.66 | -1.43 |
| $ASC_{M_3,O}$ | | | | | | | | × | | 1 | 12.95 | 4.38 |
| $ASC_{M_3,SA}$ | | | | | × | | | | | 1 | 4.17 | 1.04 |
| $ASC_{M_3,SU}$ | | × | | | | | | | | 1 | -29.15 | -7.07 |

Table 14: Estimation results and description of the specification of the **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 |
| $\theta_{M_3,1,16}$ | | | × | | × | | | | | 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 15: Estimation results and description of the specification of the **smoothed model**, associated to the model related to the detection of the first frame of the relevant group of frames

| 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 16: Estimation results of the constants for **reduced model with panel effect**

| parameter | H | SU | F | D | SA | A | N | O | DK | $x_{k,T_o,o}$ | value | $t$-test 0 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $ASC_{M_4,A}$ | | | | | | × | | | | 1 | 1.61 | 0.42 |
| $ASC_{M_4,D}$ | | | | × | | | | | | 1 | 25.40 | 5.80 |
| $ASC_{M_4,DK}$ | | | | | | | | | × | 1 | -0.067 | -0.10 |
| $ASC_{M_4,F}$ | | | × | | | | | | | 1 | 1.14 | 0.37 |
| $ASC_{M_4,H}$ | × | | | | | | | | | 1 | -3.69 | -0.94 |
| $ASC_{M_4,O}$ | | | | | | | | × | | 1 | 7.44 | 2.95 |
| $ASC_{M_4,SA}$ | | | | | × | | | | | 1 | 11.60 | 3.37 |
| $ASC_{M_4,SU}$ | | × | | | | | | | | 1 | -9.91 | -4.83 |

Table 17: Estimation results and description of the specification of **reduced model with panel effect**

| parameter | H | SU | F | D | SA | A | N | O | DK | $x_{k,T_o,o}$ | value | $t$-test 0 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\theta_{M_4,1}$ | | | | × | | | | | | EDU_6 | -6.68 | -3.27 |
| $\theta_{M_4,2}$ | | | | × | | | | | | EDU_8 | -4.57 | -3.68 |
| $\theta_{M_4,3}$ | | × | | | | × | | | | RAP_brow | 6.38 | 4.41 |
| $\theta_{M_4,4}$ | | × | × | | | | | | | RAP_mouth | 2.70 | 3.33 |
| $\theta_{M_4,5}$ | × | | | | | | | | | RAP_mouth | 9.66 | 5.50 |
| $\theta_{M_4,6}$ | × | | | | | | | | | C_1 | -16.70 | -2.33 |
| $\theta_{M_4,7}$ | | | | | | × | | | | C_2 | 22.76 | 2.80 |
| $\theta_{M_4,8}$ | | | | × | | | | | | C_2 | 25.20 | 4.01 |
| $\theta_{M_4,9}$ | × | | | | | | | | | C_3 | 15.84 | 2.47 |
| $\theta_{M_4,10}$ | | × | | | | | | | | C_3 | 15.92 | 6.03 |
| $\theta_{M_4,11}$ | | | | | × | | | | | broweye_l2 | 158.76 | 3.00 |
| $\theta_{M_4,12}$ | | × | | | | | | | | broweye_l3 | 82.23 | 5.75 |
| $\theta_{M_4,13}$ | | × | × | × | × | × | | | | broweye_r2 | -52.02 | -3.20 |
| $\theta_{M_4,14}$ | | | × | | × | | | | | eye_angle_l | 55.23 | 3.12 |
| $\theta_{M_4,15}$ | | | | | × | | | | | eye_brow_angle_l | -143.11 | -7.56 |
| $\theta_{M_4,16}$ | | | | × | | | | | | eye_mouth_dist_l2 | -66.87 | -2.49 |
| $\theta_{M_4,17}$ | × | | | | × | | | × | | eye_mouth_dist_l | -42.45 | -3.40 |
| $\theta_{M_4,18}$ | | | | | | × | | | | eye_nose_dist_l | 252.55 | 5.46 |
| $\theta_{M_4,19}$ | | | × | × | × | | | × | | eye_nose_dist_l | 153.93 | 3.38 |
| $\theta_{M_4,20}$ | | | × | × | × | × | | × | | eye_nose_dist_r | -214.88 | -3.93 |
| $\theta_{M_4,21}$ | | × | × | | | | | | | leye_h | 22.90 | 1.80 |
| $\theta_{M_4,22}$ | | | | | × | × | | | | mouth_nose_dist2 | -93.02 | -2.01 |
| $\theta_{M_4,23}$ | × | | | | | | | | | mouth_nose_dist | -235.84 | -3.82 |
| $\theta_{M_4,24}$ | × | | | | | | | | | mouth_w | 202.92 | 4.48 |
| $\sigma$ | | | | | | | | | | | 1.47 | 4.33 |

Table 18: Estimation results of the constants for the **latent model with panel effect**, associated the expression perception model

| parameter | H | SU | F | D | SA | A | N | O | DK | $x_{k,t,o}$ | value | $t$-test 0 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $ASC_{M_5,A}$ | | | | | | × | | | | 1 | -5.29 | -1.44 |
| $ASC_{M_5,D}$ | | | | × | | | | | | 1 | 20.90 | 4.44 |
| $ASC_{M_5,DK}$ | | | | | | | | | × | 1 | -0.180 | -0.25 |
| $ASC_{M_5,F}$ | | | × | | | | | | | 1 | -3.30 | -0.63 |
| $ASC_{M_5,H}$ | × | | | | | | | | | 1 | -11.08 | -0.96 |
| $ASC_{M_5,O}$ | | | | | | | | × | | 1 | 14.70 | 3.00 |
| $ASC_{M_5,SA}$ | | | | | × | | | | | 1 | 10.09 | 1.95 |
| $ASC_{M_5,SU}$ | | × | | | | | | | | 1 | -22.50 | -6.45 |

Table 19: Estimation results and description of the specification of the **latent model with panel effect**, associated to the expression perception model

| parameter | H | SU | F | D | SA | A | N | O | DK | $x_{k,t,o}$ | value | $t$-test 0 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\theta_{M_5,1,1}$ | | | | × | | | | | | EDU_6 | -6.10 | -4.01 |
| $\theta_{M_5,1,2}$ | | | | × | | | | | | EDU_8 | -3.85 | -3.65 |
| $\theta_{M_5,1,3}$ | | × | | | | × | | | | RAP_brow | 7.62 | 3.37 |
| $\theta_{M_5,1,4}$ | | × | × | | | | | | | RAP_mouth | 3.96 | 2.98 |
| $\theta_{M_5,1,5}$ | × | | | | | | | | | RAP_mouth | 17.70 | 3.27 |
| $\theta_{M_5,1,6}$ | × | | | | | | | | | C_1 | -30.40 | -4.20 |
| $\theta_{M_5,1,7}$ | | | | | | × | | | | C_2 | 43.40 | 5.52 |
| $\theta_{M_5,1,8}$ | | | | × | | | | | | C_2 | 46.10 | 5.68 |
| $\theta_{M_5,1,9}$ | × | | | | | | | | | C_3 | 21.60 | 3.21 |
| $\theta_{M_5,1,10}$ | | × | | | | | | | | C_3 | 25.30 | 3.99 |
| $\theta_{M_5,1,11}$ | | | | | × | | | | | broweye_l2 | 238.00 | 4.76 |
| $\theta_{M_5,1,12}$ | | × | | | | | | | | broweye_l3 | 87.70 | 4.30 |
| $\theta_{M_5,1,13}$ | | × | × | × | × | × | | | | broweye_r2 | -51.60 | -3.14 |
| $\theta_{M_5,1,14}$ | | | × | | × | | | | | eye_angle_l | 39.3 | 1.80 |
| $\theta_{M_5,1,15}$ | | | | | × | | | | | eye_brow_angle_l | -190.00 | -7.84 |
| $\theta_{M_5,1,16}$ | | | | × | | | | | | eye_mouth_dist_l2 | -67.8 | -1.82 |
| $\theta_{M_5,1,17}$ | × | | | | × | | | × | | eye_mouth_dist_l | -84.30 | -3.71 |
| $\theta_{M_5,1,18}$ | | | | | | × | | | | eye_nose_dist_l | 258.00 | 3.15 |
| $\theta_{M_5,1,19}$ | | | × | × | × | | | × | | eye_nose_dist_l | 106.00 | 1.59 |
| $\theta_{M_5,1,20}$ | | | × | × | × | × | | × | | eye_nose_dist_r | -223.00 | -3.01 |
| $\theta_{M_5,1,21}$ | | × | × | | | | | | | leye_h | 46.50 | 3.02 |
| $\theta_{M_5,1,22}$ | × | × | × | × | × | × | | | | mouth_h | 103.00 | 2.42 |
| $\theta_{M_5,1,23}$ | | | | | × | × | | | | mouth_nose_dist2 | -121.00 | -1.91 |
| $\theta_{M_5,1,24}$ | × | | | | | | | | | mouth_nose_dist | -327.00 | -3.00 |
| $\theta_{M_5,1,25}$ | × | | | | | | | | | mouth_w | 215.00 | 4.74 |
| $\theta^z_{M_5,1,1}$ | | | | | | | | | | mouth_h, $z_{1,t,o}$ | 55.20 | 3.06 |
| $\sigma_{M_5}$ | | | | | | | | | | | 1.20 | 2.44 |

Table 20: Estimation results of the **latent model with panel effect**, associated to the memory effects parameters

| parameter | value | $t$-test 0 |
|---|---|---|
| $\alpha_{M_5,H}$ | -0.557 | -4.29 |
| $\alpha_{M_5,F}$ | -0.314 | -2.14 |
| $\alpha_{M_5,SA}$ | -0.381 | -1.31 |
| $\alpha_{M_5,O}$ | -0.585 | -2.64 |

Table 21: Estimation results and description of the specification of the **latent model with panel effect**, associated to the model which detects the most meaningful frame

| parameter | $y_{k,t,o}$ | value | $t$-test 0 |
|---|---|---|---|
| $\theta^{y}_{M_5,2,1}$ | C_2 | -506.23 | -3.58 |
| $\theta^{y}_{M_5,2,2}$ | eye_brow_angle | 311.53 | 3.93 |
| $\theta^{y}_{M_5,2,3}$ | mouth_w | 438.40 | 3.69 |
| $\theta^{y}_{M_5,2,4}$ | C_4 | 441.12 | 3.85 |
| $\theta^{y}_{M_5,2,5}$ | eye_h | -634.03 | -3.63 |
| $\theta^{y}_{M_5,2,6}$ | mouth_h | 123.99 | 3.66 |
| $\theta^{z}_{M_5,2,1}$ | brow_dist, $z_{4,t,o}$ | 295.89 | 3.76 |