

# HOLISTIC AND FEATURE-BASED INFORMATION TOWARDS DYNAMIC MULTI-EXPRESSIONS RECOGNITION

Zakia Hammal

*International Laboratory for Brain Music and Sound Research (BRAMS), Université de Montréal  
Pavillon 1420 boul. Mont Royal, Montréal, Canada*

Corentin Massot

*Department of Physiology, McGill University, 3655 Prom. Sir William Osler, Montréal, Canada*

**Keywords:** Facial Expressions, Multiscale Spatial Filtering, Holistic Processing, Features Processing, Classification, TBM.

**Abstract:** Holistic and feature-based processing have both been shown to be involved differently in the analysis of facial expression by human observer. The current paper proposes a novel method based on the combination of both approaches for the segmentation of “emotional segments” and the dynamic recognition of the corresponding facial expressions. The proposed model is a new advancement of a previously proposed feature-based model for static facial expression recognition (Hammal *et al.*, 2007). First, a new spatial filtering method is introduced for the holistic processing of the face towards the automatic segmentation of “emotional segments”. Secondly, the new filtering-based method is applied as a feature-based processing for the automatic and precise segmentation of the transient facial features and estimation of their orientation. Third, a dynamic and progressive fusion process of the permanent and transient facial feature deformations is made inside each “emotional segment” for a temporal recognition of the corresponding facial expression. Experimental results show the robustness of the holistic and feature-based analysis, notably for the analysis of multi-expression sequences. Moreover compared to the static facial expression classification, the obtained performances increase by 12% and compare favorably to human observers’ performances.

## 1 INTRODUCTION

Significant efforts have been made during the past two decades to improve the automatic recognition of facial expressions in order to understand and appropriately respond to the users intentions. Applied in every day life situations (for example monitoring facial expression of Pain), such a system must be sensitive to the temporal behavior of the human face and able to analyze consecutive facial expressions without interruption. Yet, few efforts have been made so far for the dynamic recognition of multiple facial expressions in video sequences. Indeed, most of the past work on facial expressions recognition focused on static classification or at best assume that there is only one expression in the studied sequences. Recent studies have investigated the temporal information for the recognition of facial expressions (Pantic *et al.*, 2009). For example Pantic *et al.*, 2006; Valstar *et al.*, 2007; Koelstra *et al.*,

2008 introduced the temporal information for the recognition of Action Units (AUs) activation into 4 temporal segments (e.g. neutral, onset, apex, offset) in a predefined number of frames, while Tong *et al.*, 2007, introduced the temporal correlation between different AUs for their recognition. However, in our point in view these systems bypass the problem of facial expression recognition (which requires an additional processing step after detecting the AUs) and they do not allow to explicitly recognize more than one facial expression in a video sequence. Compared to these models, Zhang *et al.*, 2005; Gralewski *et al.*, 2006; Littlewort *et al.*, 2006, introduced the temporal information for facial expression recognition. However, the temporal information was mainly introduced in order to improve the systems’ performances. None of the proposed methods take explicitly into account the temporal dynamic of the facial features and their asynchronous deformation from the beginning to the

end of the facial expressions. Moreover, all the proposed methods are either holistic (analysis of the whole texture of the face, Littlewort *et al.*, 2006; Tong *et al.*, 2007) or feature-based (analysis of facial features information such as eyes, eyebrows and mouth, Pantic *et al.*, 2006; Valstar *et al.*, 2007; Koelstra *et al.*, 2008), or at best combine the permanent and transient facial features (i.e. wrinkles in a set of selected areas, Zhang *et al.*, 2005) for the automatic recognition of facial expression. However, it has been established in psychology that holistic and feature-based processing are both engaged in facial expressions recognition (Kaiser *et al.*, 2006). Compared to these methods, the current contribution proposed a new video based method for facial expressions recognition, which exploits both holistic and feature-based processing. The proposed holistic processing is employed for the automatic segmentation of consecutive “emotional segments” (i.e. a set of consecutive frames corresponding to a facial muscles activation compared to a Neutral state), and consists in the estimation of the global energy of the face by a multiscale spatial-filtering using log-Normal filters. The feature-based processing consists in the dynamic and progressive analysis of permanent and transient facial feature behavior inside each emotional segment for the recognition of the corresponding facial expression. The dynamic and progressive fusion process allows dealing with asynchronous facial feature deformations. The permanent facial features information is measured by a set of characteristic points around the eyes, the eyebrows and the mouth based on the work of Hammal *et al.*, (Hammal *et al.*, 2006). A new filtering-based method is proposed for transient facial features segmentation. Compared to the commonly proposed canny based methods for wrinkles detection (Tian *et al.*, 2001; Zhang *et al.*, 2005), the proposed spatial filtering method provides a precise detection of the transient features and an estimation of their orientation in a single pass. The fusion of all the facial features information is based on the Transferable Belief Model (TBM) (Smets *et al.* 1994). The TBM has already proved its suitability for facial expression classification (Hammal *et al.*, 2007) and to explicitly model the doubt between expressions in the case of blends, combinations or uncertainty between two or several facial expressions. Given the critical factor of the temporal dynamics of facial features for facial expressions recognition, a dynamic and progressive fusion process of the permanent and of the transient facial features information (dealing with asynchronous behaviour) is made inside each emotional segment from the beginning to the end

based on the temporal modelling of the TBM.

## 2 HOLISTIC AND FEATURE BASED PROCESSING

Facial expression results from the contraction of the permanent facial feature (such as eyes, eyebrows and mouth) and the skin texture deformations leading to the appearance of transient features (such as nasolabial furrows and nasal root wrinkles) (Tian *et al.*, 2005). Based on these considerations, a holistic (whole face analysis) and feature based (individual analysis of each facial feature) processing are proposed for measuring expressive deformation, transient feature segmentation and facial expression recognition.

### 2.1 Holistic Face Processing for Emotional Segment Detection

An emotional segment corresponds to all the frames between each pair of *beginning* and *end* of each facial expression. Facial muscle activation during facial expressions induces local changes in spatial frequencies and orientations of the face compared to the relaxation state (i.e. Neutral). These global changes can be measured by the energy response of a bank of filters at different frequencies and orientations. The current paper presents a holistic face processing technique based on a Log-Normal filtering process for dynamic detection of pairs of *beginning* and *end* of multiple emotional segments in video sequences.

**Log-Normal Filtering.** The studied face is first automatically detected in video streams using the method proposed by (Fasel *et al.*, 2005) and tracked in the remaining of the sequence (Hammal *et al.*, 2006). To cope with the problem of illumination variation, a preprocessing stage based on a model of the human retina (Beaudot, 1994) is applied to each detected face (see Figure 1.b). This processing enhances the contours and realizes a local correction of the illumination variation. To take away the frame border information and to only measure the facial deformations, a Hamming circular window is applied to the filtered face (Figure 1.b). The power spectra of the obtained face area is then passed through a bank of Log-Normal filters (15 orientations and 2 central frequencies), leading to a collection of features measuring the amount of energy displayed by the face at different frequency bands and across all orientations (Figure 1.c). The

Log-Normal filters are chosen because of their advantage of being easily tuned and separable in frequency and orientation (Massot *et al.*, 2008) which make them well suited for detecting features at different scales and orientations (see section 2.2.2). They are defined as follow:

$$|G_{i,j}(f,\theta)|^2 = |G_i(f)G_j(\theta)|^2 = A \cdot \frac{1}{f} \cdot \exp\left(-\frac{1}{2}\left(\frac{\ln(f/f_i)}{\sigma_f}\right)^2\right) \cdot \exp\left(-\frac{1}{2}\left(\frac{\theta-\theta_i}{\sigma_\theta}\right)^2\right) \quad (1)$$

Where  $G_{i,j}$  is the transfer function of the filter,  $G_i(f)$  and  $G_j(\theta)$ , respectively, represents the frequency and the orientation components of the filter;  $f_i$  is the central frequency,  $\theta_i$ , the central orientation,  $\sigma_f$ , the frequency bandwidth,  $\sigma_\theta$ , the orientation bandwidth and  $A$ , a normalization factor.

**Emotional Segments: Detection.** Facial muscle activity is measured by the energy of the obtained filters' responses. The amount of energy displayed by the face at two high frequencies ( $f_1=0.25$  and  $f_2=0.17$ ) and across all orientations are summed and called global energy as follow:

$$E_{global} = \sum_{i=1,2} \sum_{j=1..15} \|S_{frame}(f,\theta) * G_{i,j}(f,\theta)\|^2 \quad (2)$$

Where  $E_{global}$  is the global energy of the face and  $S_{frame}(f,\theta)$ , the power spectra of the current frame. The obtained results (Figure 1.d) show high-energy response (white areas) around the permanent facial features (such as eyes, eyebrows and mouth) and transient facial features (such as nasolabial furrows and nasal root wrinkles). These examples show that facial feature behaviors effectively induce a change of the measured global energy.

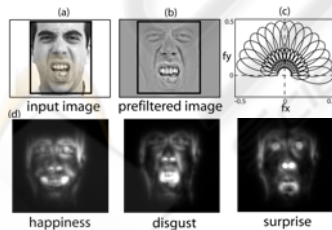


Figure 1: (a) input image, (b) after retinal filtering and with a hamming window, (c) bank of Log-Normal filters, (d) global energy response of Log-Normal filter during three facial expressions.

Figure 2 shows examples of the temporal evolution of the global energy of different subjects and for different facial expressions going from Neutral to the apex of the expression and coming back to Neutral. Similar evolutions can be observed for all the subjects independently of individual

morphological differences and facial expressions. The global energy is then used to detect each emotional segment as the set of frames between each pair of *beginning* and *end*. The *beginning* of each facial expression is characterized by the increase of the global energy of the face and the *end* as the coming-back of this energy to its value at the *beginning* taken as a reference value.

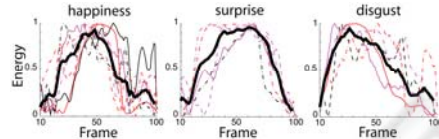


Figure 2: Time course of the global energy (normalized in amplitude and length) for 3 facial expressions and for 9 subjects from the Hammal-Caplier database. Black curves correspond to the mean curve of all the subjects.

The detection of the *beginning*  $F_b$  of each emotional segment is computed based on the derivative of the global energy signal  $\frac{d}{dt}(E_{global})$ . Indeed, a positive peak of the corresponding derivative function directly traduces an increase of the global energy. The temporal average  $M_t$  of the derivative function of the global energy and its standard deviation  $S_t$  from the *beginning* of the sequence (or from the *end* of a previous segment) are computed progressively. The *beginning*  $F_b$  corresponds to the first frame verifying:

$$\frac{d}{dt}(E_{global}(F_b)) > (M_t + S_t) \quad (3)$$

The detection of the *end* of each emotional segment  $F_e$  is made after each *beginning* frame. The *end* of each segment is considered as the coming back of the global energy to a reference value. To do so, the detection process begins 12 frames after the *beginning* of the segment (i.e. the minimum time necessary for a complete muscle activity (contraction and relaxation), see (Ekman *et al.*, 1978) and section 2.1.3). A temporal sliding window of 6 frames (time for muscle contraction, see section 2.1.3) is then used to measure the local average of the global energy signal. The first frame verifying equation 4 is considered as the *end* of the current emotional segment.

$$(M_t - S_t) \leq E_{global}(F_e) \leq (M_t + S_t) \quad (4)$$

It is important to notice that the proposed method allows the detection of each pair ( $F_b$ ,  $F_e$ ) on-line, without any post-processing step, and makes it independent of the absolute level of global energy that can be dependent of the expression intensity or face morphology. Figure 3 shows examples of

detection of expressive segments from the Hammal-Caplier and the MMI databases (the MMI-Facial Expression Database collected by M. Pantic and her group (www.mmifacedb.com), Pantic et al. 2005). The automatic segmentation appears very comparable to a manual segmentation and robust to variable duration of the expressive segments.

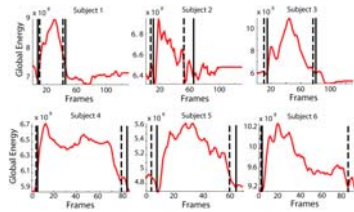


Figure 3: Example of detection of the *beginning* and the *end* on Hammal-Caplier (top) and MMI (bottom) databases. Dashed lined correspond to the automatic results and plain lines to expert manual segmentation.

The detection of the *beginning* and the *end* of facial expressions can also be applied several times during a multi-expression sequence. Figure 4 shows the evolution of the global energy during a sequence where the subject expressed 4 different facial expressions sequentially. Each *beginning* is detected (using equation 3) starting either at the first frame of the sequence or at the frame following immediately the last detected *end* (using equation 4).

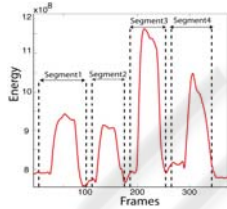


Figure 4: Example of automatic segmentation result of one sequence containing 4 expression segments. Dashed lines correspond to each detected pair of *beginning* and *end*.

The obtained result shows how the proposed method successfully detects the different emotional segments. At best of our knowledge, this is the first time where several facial expression segments are automatically detected in a video sequence. After the segmentation process each expressive segment is automatically and independently analyzed to recognize the corresponding facial expression based on a feature-based processing.

**Emotional Segments: Performances.** Intensive tests on dynamic facial expression sequences (single facial expression sequences such as Hammal-Caplier and MMI databases) and multi-expressions sequences (4 facial expression sequences acquired in

our laboratory) show the robustness of the proposed method to different acquisition conditions, individual differences and displayed facial expressions. Table 1 summarizes the mean frame differences between the automatic detection of the *beginning* and the *end* compared to a manual detection (which may also vary for different experts). Over all the used sequences (96 in total) the mean frame difference for the *beginning* and *end* detection is 8.1. This result can be related to findings that suggested that temporal changes in neuromuscular facial activity are from 0.25s to several minutes (Ekman *et al.*, 1978). The obtained error based on a minimum video frame rate of 24 frames/s is comparable to the shortest facial muscle activity duration (i.e. 6 frames).

Table 1: Detection errors of the *beginning* and the *end* of emotional segments and number of tested sequences.

	beginning	end errors	# seq.
Hammal_Caplier	9.24 frames	12.6 frames	63
MMI	3.42 frames	7.5 frames	29
Multi-expression	5.6 frames	10 frames	4

Considering that each result with an error less then 6 frames as a good detection, the performances of the emotional segments detection reach 89%. It is difficult to compare the obtained results to the few proposed works for the automatic recognition of multiple expressions because they either classify segments of AUs or did not report a quantitative evaluation of an expressive segment detection.

## 2.2 Feature-based Processing of Emotional Segments

Feature-based processing consists in the combination of the information resulting from the permanent and the transient facial feature deformations for the recognition of facial expressions during each emotional segment.

### 2.2.1 Permanent Facial Feature Information

The permanent facial feature behavior is measured based on the work of Hammal *et al.*, 2007. First, face and permanent facial features (eyes, eyebrows and mouth) are automatically segmented (see Hammal *et al.*, 2006 and Figure 5.a). Secondly, five characteristic distances  $D_i$   $1 \leq i \leq 5$  coding the displacement of a set of selected facial feature points according to the Neutral state are measured (Figure 5.b, see Hammal *et al.*, 2007 for detailed explanation of this choice). Facial expressions are then

characterized by the behavior of the measured characteristic distances.

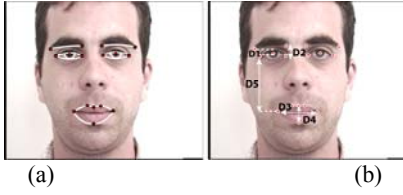


Figure 5: (a) Example of facial features segmentation (Hammal *et al.*, 2006); (b) associated characteristic distances (Hammal *et al.*, 2007).

## 2.2.2 Transient Facial Feature Information

In addition to the permanent facial features (Hammal *et al.*, 2007) and in order to provide additional information to support the recognition of facial expressions, the analysis of the transient facial features such as nasal root wrinkles (Figure 6 Areas 1) and the nasolabial furrows (Figure 6 Areas 2,3) (being part of the most important visual cues used by human observer for facial expression recognition (Smith *et al.*, 2005)) is introduced. Transient facial feature areas are first located based on the permanent facial features segmentation (Figure 6). The filtering based-method proposed in section 2.1 is then applied inside each selected area for the estimation their appearance and the corresponding orientation when necessary. Figure 7 shows the different processing steps.

After the selection process (Figure 7.b), a Hamming window is applied to each area (Figure 7.c).

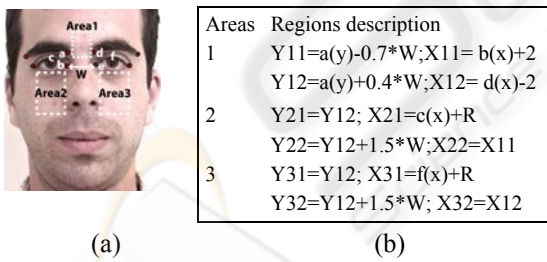


Figure 6: (a) detected wrinkles regions, (b) transient feature areas, R: eyes radius, W the distance between eyes corners.

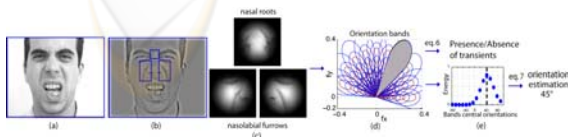


Figure 7: Transient features detection and orientation estimation.

The response of orientation bands  $B_{j,t}$  that corresponds to the sum of the responses of all filters sharing the same central orientation at different spatial frequencies (Figure 7.d grey) is measured as:

$$B_{j,t} = \sum_{i=1..7} \|S_i(f, \theta) G_{i,j}(f, \theta)\|^2 \quad (5)$$

This allows analyzing an oriented transient feature independently of its spatial frequency making the detection more robust to individual morphological differences.

Wrinkles detection: for each frame  $t$ , nasal root wrinkles and nasolabial furrows are detected based on the sum of total energy  $E_t$  over all the orientation bands  $j$  inside each selected area as:

$$E_t = \sum_{j=1..15} B_{j,t} \quad (6)$$

Wrinkles are “present” if  $E_t$  is higher than a predefined threshold and “absent” otherwise. Threshold values on the energy measure are obtained after learning process over three benchmark databases (Cohn-Kanade, Dailey-Cottrel and STOIC databases) and generalized on the Hammal-Caplier database. Table 2 shows the detection performances. The obtained results are more than sufficient to reinforce the permanent facial features information.

Indeed as explained in section 3.1.2, if they are present the corresponding information will be taken into account as a refinement of the classification process otherwise the doubt resulting from the permanent facial feature analysis is kept rather than making a wrong decision.

Table 2: Detection performances of the transient features; the threshold value has been chosen equal to 0.12 in order to minimize false alarms' rate.

	Recall %	Precision %	F-measure
Nasal Roots	73	71	72
Nasolabial Furrows	86	85	86

Nasolabial furrows orientation: once the nasolabial furrows detected, their orientation (the angle between their edge line and the horizontal plan of the corresponding area) is measured by linear combination of the orientation bands responses as:

$$\theta = \sum_{j=1..15} B_{j,t} \cdot \theta_j \quad (7)$$

Figure 8 shows examples of dynamic detection of nasolabial furrows and nasal roots wrinkles during sequences of Happiness and Disgust expressions.

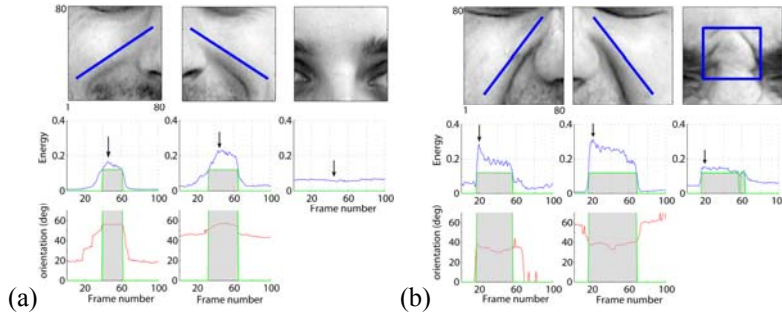


Figure 8: Example of nasolabial furrows and nasal roots detection during Happiness (a) and Disgust (b) sequences; gray temporal windows (second rows) indicate the temporal presence of the transient features based on the energy threshold (0.11, validated on three databases); third rows display the measured angles of the nasolabial furrows (around  $60^\circ$  for Happiness and  $45^\circ$  for Disgust).

One can see that nasal roots appear for Disgust but not for Happiness and that nasolabial orientations are different according of the expression. These examples show the usefulness of these wrinkles to characterize the corresponding facial expressions.

### 2.3 Numerical to Symbolic Conversion of Facial Features Behavior

A numerical to symbolic conversion translates the measured distances, transient features and the corresponding angles into symbolic states reflecting their behavior. First, the value of each characteristic distance  $D_i$  is coded with five symbolic states (based on the work of Hammal *et al.*, 2007) reflecting the magnitude of the corresponding deformations:  $S_i$  if  $D_i$  is roughly equal to its value in the Neutral expression,  $C_i^+$  (vs.  $C_i^-$ ) if  $D_i$  is significantly higher (vs. lower) than its value in the Neutral expression, and  $S_i \cup C_i^+$  (vs.  $S_i \cup C_i^-$ ) if the  $D_i$  is neither sufficiently higher (vs. lower) to be in  $C_i^+$  (vs.  $C_i^-$ ), nor sufficiently stable to be in  $S_i$ . Following this symbolic association, two states are introduced for Nasal root and nasolabial furrows behaviors: “present”  $P_j$  or “absent”  $A_j$   $1 \leq j \leq 2$  according to the corresponding energy measure as described in section 2.2.2. The explicit doubt of their state  $P_j \cup A_j$  ( $P_j$  or  $A_j$ ) is introduced and allows modeling the uncertainty of their detection (see section 3.1). Finally, two symbolic states are also introduced for nasolabial furrows angles: “opened”  $Op$  and “closed”  $Cl$ . If the angle is higher (resp. lower) than a predefined value the state  $Op$  (resp.  $Cl$ ) is chosen. As for the wrinkles detection a doubt state  $Op \cup Cl$  is also introduced to model

the uncertainty of the measured angles (see section 3.1).

Table 3 summarizes the characteristic distances, the transient feature and the nasolabial furrow angle states for each facial expression. However, a logic-based system is not sufficient to model the facial expressions. Indeed, an automatic facial expression system should explicitly model the doubt and uncertainty of the sensors (such as  $P_j \cup A_j$  states) generating its conclusion with confidence that reflects uncertainty of the sensors detection and tracking. For this reason, the Transferable Belief Model (TBM) is used.

## 3 TBM BELIEF MODELING

The TBM (Smets *et al.*, 1994) considers the definition of the frame of discernment of  $N$  exclusive and exhaustive hypotheses characterizing the six basic facial expressions and Neutral  $\Omega = \{Happiness (E_1), Surprise (E_2), Disgust (E_3), Fear (E_4), Anger (E_5), Sadness (E_6), Neutral (E_7)\}$ . The TBM requires the definition of the Basic Belief Assignment (BBA) associated to each independent source of information.

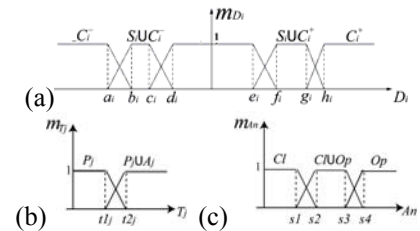



Figure 9: (a): model of BBAs for the characteristic distances (Hammal *et al.*, 2007); (b): model of BBAs for the transient features detection; (c): model of BBAs of the Nasolabial furrow angles.

Table 3: Rules table defining the visual cues states corresponding to each facial expression.



	$D_1$	$D_2$	$D_3$	$D_4$	$D_5$	$TF_1$	$TF_2$	$An$
Happiness	$C_1^-$	$S_2 \cup C_2^-$	$C_3^+$	$C_4^+$	$C_5^-$	$A_1$	$P_2$	$Op$
Surprise	$C_1^+$	$C_2^+$	$C_3^-$	$C_4^+$	$C_5^+$	$A_1$	$A_2$	-
Disgust	$C_1^-$	$C_2^-$	$S_3 \cup C_3^+$	$C_4^+$	$S_5$	$P_1$	$P_2$	$Cl$
Anger	$C_1^-$	$C_2^-$	$S_3$	$S_4 \cup C_4^-$	$S_5$	$P_1$	$P_2$	$Op \cup Cl$
Sadness	$C_1^-$	$C_2^+$	$S_3$	$C_4^+$	$S_5$	$A_1$	$A_2$	-
Fear	$C_1^+$	$S_2 \cup C_2^+$	$S_3 \cup C_3^-$	$S_4 \cup C_4^+$	$S_5 \cup C_5^+$	$A_1$	$A_2$	-
Neutral	$S_1$	$S_2$	$S_3$	$S_4$	$S_5$	$A_1$	$A_2$	-

### 3.1 Belief Modeling

The belief definition means the definition of the BBAs of each visual cue and is equivalent to the probabilities definition in the Bayesian model.

#### 3.1.1 Beliefs of the Permanent Facial Features

The BBA of the permanent facial features (characteristic distances) is based on the work of Hammal *et al.*, 2007. The BBA  $m_{D_i}^{\Omega_{D_i}}$  of each characteristic distance state  $D_i$  is defined as:

$$m_{D_i}^{\Omega_{D_i}} 2^{\Omega_{D_i}} \rightarrow [0, 1]$$

$$A^{\Omega_{D_i}} \rightarrow m_{D_i}^{\Omega_{D_i}}(A), \sum_{A \in 2^{\Omega_{D_i}}} m_{D_i}^{\Omega_{D_i}} = 1 \quad 1 \leq i \leq 5 \quad (8)$$

Where  $\Omega_{D_i} = \{C_i^+, C_i^-, S_i\}$  is the power set,  $2^{\Omega_{D_i}} = \{\{C_i^+\}, \{C_i^-\}, \{S_i\}, \{S_i, C_i^+\}, \{S_i, C_i^-\}, \{S_i, C_i^+, C_i^-\}\}$  the frame of discernment,  $\{S_i, C_i^+\}$  (vs.  $\{S_i, C_i^-\}$ ) the doubt state between  $C_i^+$  (vs.  $C_i^-$ ) and  $S_i$ ,  $m_{D_i}^{\Omega_{D_i}}(A)$ , the belief in the proposition  $A \in 2^{\Omega_{D_i}}$  without favoring any proposition of  $A$  in case of doubt proposition. This is the main difference with the Bayesian model, which implies equiprobability of the propositions of  $A$ . The piece of evidence  $m_{D_i}^{\Omega_{D_i}}$  associated with each symbolic state given the value of the characteristic distance  $D_i$  is defined by the model depicted in Figure 9.a.

#### 3.1.2 Beliefs of the Transient Facial Features

**Presence:** The BBA  $m_{TF_j}^{\Omega_{TF_j}}$  of the states of each transient feature  $TF_j$  is defined as:

$$m_{TF_j}^{\Omega_{TF_j}} 2^{\Omega_{TF_j}} \rightarrow [0, 1]$$

$$B^{\Omega_{TF_j}} \rightarrow m_{TF_j}^{\Omega_{TF_j}}(B), \sum_{B \in 2^{\Omega_{TF_j}}} m_{TF_j}^{\Omega_{TF_j}} = 1 \quad 1 \leq j \leq 2 \quad (9)$$

Where  $TF_1$  means the *nasal root wrinkles*,  $TF_2$ , the *nasolabial furrow*,  $\Omega_{TF_j} = \{P_j, A_j\}$ ,  $2^{\Omega_{TF_j}} = \{\{P_j\}, \{A_j\}, \{P_j, A_j\}\}$ . From the frame of discernment  $2^{\Omega_{TF_j}}$  only the states  $P_j$  (the wrinkles are present without any doubt) and the state  $\{P_j, A_j\}$  (there is a doubt in their detection and noted  $P_j \cup A_j$ ) are considered. Then if the wrinkles are detected as present (the energy threshold is higher than the defined value) the corresponding state is  $P_j$  if not, the corresponding state is  $P_j \cup A_j$ . The piece of evidence  $m_{TF_j}^{\Omega_{TF_j}}$  of each state is derived according to the model depicted in Figure 9.b. The nasal root wrinkles are used as a refinement process and are associated to Disgust and Anger expressions (without favoring any of them). If they are present the current expression is Disgust or Anger with the piece of evidence:  $m_{TF_1}^{\Omega_{TF_1}}(P) = m_{TF_1}^{\Omega_{TF_1}}(E_3 \cup E_5) = 1$ . If they are not present, the current expression can be one of the 7 studied expressions with the piece of evidence:  $m_{TF_1}^{\Omega_{TF_1}}(P \cup A_1) = m_{TF_1}^{\Omega_{TF_1}}(E_1 \cup E_2 \cup E_3 \cup E_4 \cup E_5 \cup E_6 \cup E_7) = 1$ . If present, the nasolabial furrows are associated to Happiness, Disgust and Anger expressions with the piece of evidence:  $m_{TF_2}^{\Omega_{TF_2}}(P_2) = m_{TF_2}^{\Omega_{TF_2}}(E_1 \cup E_3 \cup E_5) = 1$ . If they are absent: the current expression is one of the 7 expressions with the piece of evidence:  $m_{TF_2}^{\Omega_{TF_2}}(P_2 \cup A_2) = m_{TF_2}^{\Omega_{TF_2}}(E_1 \cup E_2 \cup E_3 \cup E_4 \cup E_5 \cup E_6 \cup E_7) = 1$ .

**Orientation:** The BBAs of the nasolabial furrow angle states are defined as:

$$m_{An}^{\Omega_{An}} 2^{\Omega_{An}} \rightarrow [0, 1]$$

$$C^{\Omega_{An}} \rightarrow m_{An}^{\Omega_{An}}(C), \sum_{C \in 2^{\Omega_{An}}} m_{An}^{\Omega_{An}} = 1 \quad (10)$$

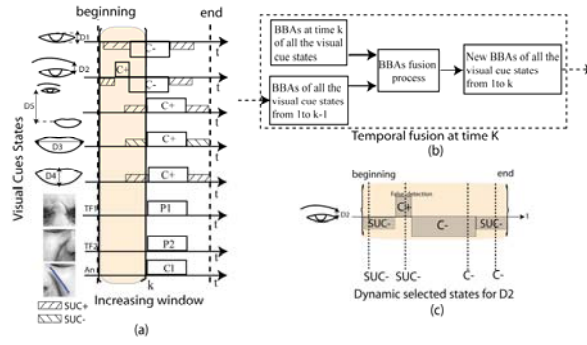


Figure 10: (a) Example of the increasing temporal window during a sequence of Disgust expression; (b) BBAs selection refinement process at time  $k$ ; (c) Example of the selection of the characteristic distance states inside the temporal window.

Where  $An$  is the angle,  $\Omega_{An} = \{Op, Cl\}$ ,  $2^{\Omega_{An}} = \{\{Op\}, \{Cl\}, \{Op, Cl\}\}$ ,  $Op$  and  $Cl$  mean opened and closed angles (see section 2.2.2)  $\{Op, Cl\}$  means  $Op$  or  $Cl$  and corresponds to the doubt between  $Op$  and  $Cl$  (noted  $Op \cup Cl$ ). The pieces of evidence associated to the states of the computed detected angles are defined using the model proposed in Figure 9.c. Based on the BBAs of the nasolabial furrow angle states, the piece of evidence associated to each one of the 3 expressions Happiness ( $E_1$ ), Anger ( $E_3$ ) and Disgust ( $E_5$ ) is

based on the fuzzy-like model of Figure 9.d as:

$$\begin{aligned} m_{An}^{\Omega_{An}}(An \leq s1) &= m_{An}^{\Omega_{An}}(Cl) = m_{An}^{\Omega_{An}}(E_5) = 1 \\ m_{An}^{\Omega_{An}}(An \geq s4) &= m_{An}^{\Omega_{An}}(Op) = m_{An}^{\Omega_{An}}(E_1) = 1 \\ m_{An}^{\Omega_{An}}(s2 \leq An \leq s3) &= m_{An}^{\Omega_{An}}(Op \cup Cl) = m_{An}^{\Omega_{An}}(E_1 \cup E_3 \cup E_5) = 1 \end{aligned}$$

In the other cases the piece of evidence of the expression or subset of expressions is equal to the projection of the angle value on the proposed model.

## 4 TEMPORAL INFORMATION

The dynamic and asynchronous behavior of the facial features is introduced by combining at each time  $t$  their previous deformations from the *beginning* until the *end* of each emotional segment (see Section 2) to take a decision. The analysis of the facial feature states is made inside an increasing temporal window  $\Delta t$  (Figure 10. a). The size of the window  $\Delta t$  increases progressively at each time from the detection of the *beginning* until the detection of the *end* of the expression. Then, at each time  $t$  inside the window  $\Delta t$ , the current state of each facial feature (i.e. characteristic distances and transient features) is selected based on the combination of their current state at time  $t$  and of the whole set of their past states since the *beginning* which then takes into account their dynamic and asynchronous facial

feature deformations (Figure 10.b). The dynamic fusion of the BBAs is made according to the number of appearance of each symbolic state noted  $Nb_{\Delta t}(state)$  and their integral (sum) of plausibility noted  $Pl_{\Delta t}(state)$  computed progressively inside the temporal window  $\Delta t$ . For instance, for a characteristic distance  $D_i$  and for the  $state = C^-$ :

$$K_t(C^-) = \begin{cases} 1 & \text{if } m_{D_i}(C^-) \neq 0 \\ 0 & \text{otherwise} \end{cases} \quad 1 \leq t \leq \Delta t \quad Nb_{\Delta t}(C^-) = \sum_{t=1}^{\Delta t} K_t(C^-) \quad (11)$$

$$Pl_{\Delta t}(C^-) = \sum_{t=1}^{\Delta t} (m_{D_i}(C^-) + m_{D_i}(S \cup C^-)) \quad (12)$$

From the two parameters  $Nb_{\Delta t}(state)$  and  $Pl_{\Delta t}(state)$ , the selected states of each visual cues at each time  $t$  inside the temporal window  $\Delta t$  are chosen as:

$$State_{\Delta t}(D_i) = \max(Pl_{\Delta t}(state_{D_i}) / Nb_{\Delta t}(state_{D_i})) \quad (13)$$

$$state_{D_i} \in \{C_i^+, C_i^-, S_i \cup C_i^+, S_i \cup C_i^-\} \quad 1 \leq i \leq 5$$

$$State_{\Delta t}(TR_j) = \max(Pl_{\Delta t}(state_{TR_j}) / Nb_{\Delta t}(state_{TR_j})) \quad (14)$$

$$state_{TR_j} \in \{P_j, P_j \cup A_j\} \quad 1 \leq j \leq 2$$

$$State_{\Delta t}(An) = \max(Pl_{\Delta t}(state_{An}) / Nb_{\Delta t}(state_{An})) \quad (15)$$

$$state_{An} \in \{Op, Cl, Op \cup Cl\}$$

Figure 10.c shows an example of the temporal selection of the states of the characteristic distance  $D_2$ . One can see the correction of the false detection state ( $C^+$ ) by the temporal fusion process (equation 13). The piece of evidence associated to each chosen state corresponds to its maximum piece of evidence inside the temporal window as:



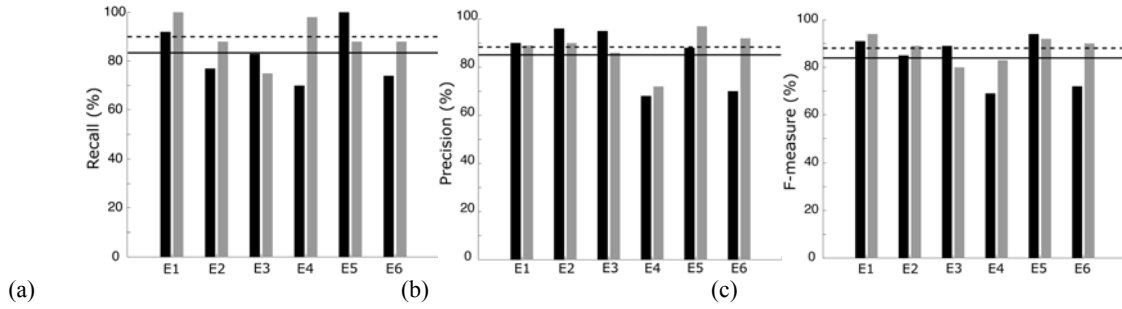


Figure 11: Recall, Precision and F measures (in %) of the model and human performances (*Happiness* ( $E_1$ ), *Surprise* ( $E_2$ ), *Disgust* ( $E_3$ ), *Fear* ( $E_4$ ), *Anger* ( $E_5$ ), *Sadness* ( $E_6$ ), *Neutral* ( $E_7$ )). Black bars: model performances; grey bars, human performances. Plain and dashed horizontal lines mean model and human performances over all the expressions respectively.

$$m_{\Delta}^{State(Cues),\Delta t} = \max(m_{Cues,1,\dots,\Delta t}) \quad (16)$$

$$Cues \in \{D_i, TF_j, An\} \quad 1 \leq i \leq 5, 1 \leq j \leq 2$$

Then at each time  $t$  from the *beginning* to the *end* of the expression sequence, once the basic belief assignments of all the visual cues are refined, the corresponding expression is selected according to the rules table 3.

## 5 BELIEFS FUSION

The fusion process of all the visual cue states is done at each time (Figure 10.a) using the conjunctive combination rule (Denoeux, 2008) and results in  $m^\Omega$  the BBA of the corresponding expression or subset of expressions:

$$m^\Omega = \oplus m_{Cues}^\Omega \quad (17)$$

From Table 3 and the BBAs of the sensor states: the characteristic distance states  $m_{D_i}^{\Omega_{Di}}$ , the transient features' states  $m_{TF_i}^{\Omega_{TF_i}}$  and the angles' states  $m_{An}^{\Omega_{An}}$ , a set of BBAs on facial expressions is derived for each sensor as:  $m_{D_i}^\Omega$ ,  $m_{TF_i}^\Omega$  and  $m_{An}^\Omega$ . The fusion process of the BBAs  $m_{D_i}^\Omega$ ,  $m_{TF_i}^\Omega$  and  $m_{An}^\Omega$  is performed successively using the conjunctive combination rule (equation. 17). For example, for two characteristic distances  $D_i$  and  $D_j$  the joint BBA  $m_{D_i, j}^\Omega$  using the conjunctive combination is:

$$m_{D_i, j}^\Omega(A) = (m_{D_i}^\Omega \oplus m_{D_j}^\Omega)(A) = \sum_{E \cap F = A} m_{D_i}^\Omega(E) * m_{D_j}^\Omega(F) \quad (18)$$

The obtained results are then combined to the BBAs of the transient features' states as:

$$m_{D_i, TF_j}^\Omega(G) = (m_{D_i}^\Omega \oplus m_{TF_j}^\Omega)(G) = \sum_{A \cap B = G} m_{D_i}^\Omega(A) * m_{TF_j}^\Omega(B) \quad (19)$$

The obtained results are finally combined to the BBAs of the angles' states as:

$$m_{D_i, TF_j, An}^\Omega(H) = (m_{D_i, TF_j}^\Omega \oplus m_{An}^\Omega)(H) = \sum_{G \cap C = H} m_{D_i, TF_j}^\Omega(G) * m_{An}^\Omega(C) \quad (20)$$

Where  $A, B, E, F, G, H, C$  denote propositions and  $E \cap F, A \cap B, G \cap C$  the conjunction (intersection) between the corresponding propositions. This leads to propositions with a lower number of elements and with more accurate pieces of evidence.

The decision is the ultimate step and consists in making a choice between various hypotheses  $E_e$  and their possible combinations. The decision is made using the credibility as:

$$Bel: 2^\Omega \rightarrow [0, 1] \quad (21)$$

$$I \rightarrow Bel(I) = \sum m^\Omega(B), \forall I \in \Omega$$

## 6 CLASSIFICATION RESULTS

In order to measure the introduction of the transient features and the temporal modeling compared to the model proposed by Hammal *et al.*, 2007, the classification results were performed on the six basic facial expressions from three benchmark databases, Cohn-Kanade, Hammal-Caplier and Stoic databases (a total of 182 videos). Recall (R), Precision (P) and F-measure (F), which combines evenly *Recall* and *Precision* as:  $F = 2 * Recall * Precision / (Recall + Precision)$  are used for the evaluation of the proposed method.

Figure 11 shows the performances obtained by the proposed model (black bars). The mean classification recall and precision reaches 83% and 85% respectively and the mean f-measure (f) 84% (horizontal plain lines). The best performances are obtained for Anger (f=94%) and Happiness (f=91%). The lowest performances are obtained for Fear (f=72%) and Sadness (f=69%) expressions.

These results can be explained by two doubt states that appear frequently: the doubt between Fear and Surprise and the doubt between Sadness and Anger expressions. Interestingly, these expressions are also notoriously difficult to discriminate for human observers (Roy *et al.*, 2007). Compared to the model of Hammal *et al.*, 2007 the introduction of the temporal modeling of all the facial features information leads to an average increase of 12% of the performances. To better evaluate the quality of the obtained results, the model performances are compared with those of human observers on the same data. 15 human observers were asked to discriminate between the six basic facial expressions on 80 videos randomly interleaved in 4 separate blocks. Figure 11 reports the human performances (grey bars). The human and model performances are not significantly different (two-way ANOVA,  $P > 0.33$ ).

## 7 CONCLUSIONS

The current paper proposes a model combining a holistic and a feature-based processing for the automatic recognition of facial expressions dealing with asynchronous facial feature deformations and multi-expression sequences. Compared to the static results, the introduction of the transient features and the temporal modeling of the facial features increase the performances by 12% and compare favorably to human observers. This opens promising perspectives for the development of the model. For example, preliminary results on spontaneous pain expression recognition proved its suitability to generalize to non-prototypic facial expressions. A future direction would be the synchronization of the facial and the vocal modalities inside each detected emotional segment and the definition of a fusion process towards a bimodal model for multi-expression recognition.

## REFERENCES

- Hammal Z., Couvreur L., Caplier A., Rombaut M., 2007. Facial expressions classification: A new approach based on transferable belief model. *International Journal of Approximate Reasoning*, 46(3), 542-567.
- Hammal Z., Eveno N., Caplier A., Coulon, P-Y., 2006. Parametric models for facial features segmentation, *Signal processing*, 86, 399-413.
- Smith M., Cottrell G., Gosselin F. Schyns P.G, 2005. Transmitting and decoding facial expressions of emotions, *Psychological Science*, 16, 184-189.
- Pantic M., Patras I., 2006. Dynamics of Facial Expression: Recognition of Facial Actions and Their Temporal Segments from Face Profile Image Sequences, *IEEE Trans. SMC- Part B*, 36(2), 433-449.
- Smets P., Kruse R., 1994. The transferable belief model, *Artificial Intelligence*, 66, 191-234.
- Tian Y., Kanade T., Cohn J.F., 2001. Recognizing Action Units for Facial Expression Analysis, *IEEE Trans. PAMI*, 23(2), 97-115.
- Massot C., Herault J., 2008. Model of Frequency Analysis in the Visual Cortex and the Shape from Texture Problem, *Int. Journal of Computer Vision*, 76(2).
- Denoeux T., 2008. Conjunctive and disjunctive combination of belief functions induced by non-distinct bodies of evidence, *Artificial Intelligence*, 172:234-264.
- William Beaudot, 1994. Le traitement neuronal de l'information dans la rétine des vertébrés : Un creuset d'idées pour la vision artificielle, Thèse de Doctorat INPG, Laboratoire TIRF, Grenoble (France).
- Tian Y. L., Kanade T., Cohn J.F., 2005. Facial expression analysis, In S" Z. Li & A.K. Jain (Eds), *Handbook of face recognition*, 247-276. NY: Springer.
- Littlewort G., Bartlett M. S., Fasel I., Susskind, J. & Movellan J. 2006 Dynamics of facial expression extracted automatically from video. *J. Image Vis. Comput.* 24, 615-625.
- M.F. Valstar and M. Pantic, 2007. Combined Support Vector Machines and Hidden Markov Models for Modeling Facial Action Temporal Dynamics, in *Proc. IEEE Workshop on Human Computer Interaction*, Rio de Janeiro, Brazil, 118-127.
- Zhang Y., & Qiang, J. 2005. Active and dynamic information fusion for facial expression understanding from image sequences. *IEEE Trans. PAMI*, 27(5), 699-714.
- Gralewski L., Campbell N. & Voak I. P. 2006 Using a tensor framework for the analysis of facial dynamics. *Proc. IEEE Int. Conf. FG*, 217-222.
- Tong Y., Liao W. & Ji Q. 2007 Facial action unit recognition by exploiting their dynamics and semantic relationships. *IEEE Trans. PAMI*. 29, 1683-1699.
- Pantic M., Valstar M.F., Rademaker R. & Maat L. 2005 Web-based database for facial expression analysis. *Proc. IEEE Int. Conf. ICME'05*, Amsterdam, The Netherlands, July.