Pain monitoring: A dynamic and context-sensitive system

Zakia Hammal a,b, Miriam Kunz a

a Brain, Music, and Signal Laboratory (BRAMS), CRAMAS, Suite 0 120, Pavillon 1-209, Boulevard Mont Royal, Université de Montréal, C.P. 6128 Station Centrale Ville, Montréal (Qué.), Canada HI6 2G7
b Centres de Recherche, Institut Universitaire de Gériatrie de Montréal, 4565, Chemin Queen-Mary, Montréal (Qué.), Canada H3H 1W5

ARTICLE INFO

Article history:
Received 29 February 2010
Received in revised form 18 August 2011
Accepted 24 September 2011
Available online 5 October 2011

Keywords:
Pain expression
Spontaneous facial expressions
Context based recognition
Transferable Relief Model
Classification

ABSTRACT

The current paper presents an automatic and context-sensitive system for the dynamic recognition of pain expression among the six basic facial expressions and neutral on acted and spontaneous sequences. A machine learning approach based on the Transferable Relief Model, successfully used previously to categorize the six basic facial expressions in static images [2,20] is extended in the current paper for the automatic and dynamic recognition of pain expression from video sequences in a clinical context application. The originality of the proposed method is the use of the dynamic information for the recognition of pain expression and the combination of different sensors, permanent facial features behavior, transient features behavior, and the context of the study, using the same fusion model. Experimental results, on 2-alternative forced choices and, for the first time, on 8-alternative forced choices (i.e. pain expression is classified among seven other facial expressions), show good classification rates even in the case of spontaneous pain sequences. The mean classification rates on acted and spontaneous data reach 81.2% and 84.5% for the 2-alternative and 8-alternative forced choices, respectively. Moreover, the system performance compares favorably to the human observer rates [76,80], and lead to the same doubt states in the case of blend expressions.

©2011 Elsevier Ltd. All rights reserved.

1. Introduction

To be more efficient, the next generation of computer systems needs to integrate the ability to detect and interpret human affective states, and be capable of adapting its responses accordingly. Several research areas could benefit from such systems: interactive games with the ability to adapt to the user's intentions; interactive teaching systems that permit teachers to be aware of student stress and attention; and medical tools for automatic diagnosis and monitoring. In this paper we will focus on the development of automatic pain assessment tools, which could be of considerable value in certain circumstances where patients are not able to communicate pain verbally (e.g. newborns, individuals with severe cognitive impairments like autism). Because of the potential of such systems in everyday applications, the past 20 years has witnessed an increasing number of efforts for automatic recognition of human facial expressions. However, until the beginning of the year 2000 most of the affect recognition methods attempted to recognize a small set of prototypic and acted expressions like happiness, surprise, fear, disgust, anger, and sadness [10,22,29]. Given the limited potential of such expressions in real applications, the last years have seen a shift within the affect recognition community towards non-prototypic (e.g. attentiveness [12], fatigue [13,14], frustration [84], and pain [83,51]) and spontaneous affect recognition [6–10].

1.1. Context of the study

In the current paper, we are interested in the automatic recognition of pain expression in a medical context application, where video taping adult subjects are not able to verbally express their own painful state. We focused on automatic recognition of pain facial expression due to its potential medical significance. For example, it can be used as a pain assessment tool for individuals who are not able to communicate pain verbally (e.g. newborns [57], individuals with severe cognitive impairments like autism [18,19]). Automatic recognition of pain from facial expression provides an alternative to common signs of pain, such as verbal reports, and provides information that may be more useful and valid in certain circumstances (for example schizophrenic patients [17]). For this, we propose a new contribution for the automatic recognition of pain based on facial information in a medical context application.

Significant efforts have been made in human behavior studies to identify reliable and valid facial indicators of pain [36,1]. The standard method requires manual labeling of facial Action Units...
faces. In order to recognize video images as a pain sequence or not, the output scores of individual frames are summed together to give a cumulative score for the entire sequence. A decision threshold is then used to determine if the current sequence corresponds to pain expression or not. A leave-one-subject-out validation leads to a hit rate of 81%.

The few proposed methods for the recognition of facial expression of pain have several limitations. First, these models are based on static classification or, at best, a combination of static classification results. None of them take explicitly into account the temporal dynamics of the facial features and their asynchronous deformation during a sequence of expression [7]. Second, automatic facial expression recognition implies face and facial features segmentation where noisy and partial data should be expected. Therefore, facial expression analysis should take into account noisy and partial data and should generate their conclusions with a confidence that reflects the certainty of face and facial features tracking [20]. None of the already proposed methods reported the uncertainty of the used features nor explicitly modeled it in the classification process. Third, they used that methods focused on a 2-alternative forced choice classification (pain vs. no pain or real vs. acted pain), making it difficult to evaluate whether the proposed models can dissociate pain from other basic facial expressions. In addition, these models skipped the problem of doubt between multiple expressions. Fourth, the existing models are context-independent. Few attempts have been made towards context-dependent interpretation of the observed facial expression [13,36-32]. However, facial expressions are accordingly displayed in a particular context, such as the location (outdoor, indoor), the situation (driving a car or being treated in a hospital), the undertaking task, the other people involved, and the identity and natural expressiveness of the individual [6]. Unless mistaken, none of the already vision-based models introduce one of the possible context variables for the recognition of facial expressions. The proposed model aims at overcoming these four limitations.

1.3. Proposed contribution

In the current paper we are interested in pain expression recognition from the six basic facial expressions (happiness, surprise, disgust, anger, fear, and sadness) and neutral. The classification is made in video sequences based on a dynamic fusion process of facial information in a medical context application (i.e., pain monitoring). To this end, we extended our previous model using a static approach (proposed for the recognition of the six basic facial expressions plus neutral [21]) and applied a new temporal and multi-cues modeling process in the lack of automatic and dynamic recognition of pain facial expression.

Facial expressions (such as pain) induce the contraction of facial muscles, which result in temporally deformed facial features such as eyes, eyebrows, lips, and skin texture, often revealed by wrinkles. Moreover, in a recent study, Roy et al. [25,26] have made a finer and less biased analysis of the importance of facial features for the discrimination of the basic facial expressions classification. Pain's timing showed that methods focused on one of the prominent facial features that serve the human observer for the recognition of pain expression. Therefore, in addition to the permanent facial features (like eyes, eyebrows, and mouth), nasal root wrinkles are introduced as a refinement sensor for the automatic recognition of pain expression. Our choice has been driven by the behavioral experiments presented in [25], which reported the used features among the most important features for the classification of the six basic facial expression and neutral, and reported by our own achievements [51] as the necessary features for the classification of the six basic facial expressions and neutral.
The temporal dynamics of facial behavior represent a critical factor for recognition of complex behaviors like pain, shame, and amusement [4]. A recent study by Cohn et al. showed as well that spontaneous smiles, in contrast to posed smiles, can have multiple apices (see Fig. 8) and are accompanied by other facial feature deformations that appear either simultaneously with the mouth corners rising or following it within 1 s [7]. Another study by Valstar et al. showed also that the order and timing of spontaneous and deliberately displayed eyebrow actions are different [10]. Based on these findings, facial expression can be described as the result of progressive deformations of a set of facial features appearing at different times (asynchronously) without any defined appearance order [46]. Using only aspects of temporal dynamics of facial expression such as the speed of a facial point displacement or the persistence of facial parameters over time [20] or including the information at time t−1 [65] is not sufficient. Compared to the already proposed models, and in addition to the information at time t−1, the asynchronous facial feature behavior is explicitly modeled in the current paper, by taking into account at each time (each frame) all the available information (facial feature states) from the beginning until the end of the sequence, to recognize the current facial expression.

In addition to the visual cues and their dynamic behavior, without context, even humans may misunderstand the observed facial expression. For example, as reported by [20] smiling in the context of downward head pitch communicates embarrassment rather than joy. Then, an important related issue that should be addressed in the recognition of all emotions is how to take into account the context information. Indeed, the context dependency remains an unexplored area for automatic facial expression recognition given the difficulty of modeling this variable and of its introduction in the classification process [20]. The current paper introduces the concept of context application as new information, which bias the classifier toward the most relevant expression (i.e. the context tell us which expressions are more likely to occur). A main advantage of the dynamic fusion process proposed in the current paper is the ability to introduce the context as a new refinement sensor in the Transferable Belief Model (TBM) based model.

1.4. Justification of the fusion based model

The proposed fusion process is based on the Transferable Belief Model (TBM) [41]. The TBM is a formalism of increasing interest for imprecise and uncertain data analysis (69,80,81), information fusion [76–78], and pattern recognition (77,78). It can be used in several applications such as image processing, neurosciences, medicine, robotics, and defense (82) and more recently in the analysis and the recognition of human facial and body behavior (2,61,79). The TBM based modeling has already proved its suitability for the static classification of the six basic facial expressions and neutral [2] even in the case of partially occluded facial parts [61]. The effectiveness of the TBM based approach for the purpose of facial expression classification has been also showed by its explicit composition with the Bayesian and HMMs formalisms [25]. The TBM based modeling for facial expression recognition is of great interest for two main reasons. Firstly it deals with imprecise and uncertain information, which could be the case with data measured resulting from video based segmentation algorithm. Indeed, the TBM based model allows the explicit modeling of the doubt on the sensor states (i.e. the facial feature states) and generates its conclusions so that the associated uncertainty varies with the certainty of the facial features detection. This is an important characteristic that has been reported as one of the most important limit of the existing approaches for facial expression recognition (see [60,20]). Secondly, the TBM based model allows an explicit modeling of blend expressions, which result from the combinations and uncertainty between two or more facial expressions [61]. Indeed, human facial expressions are not always binary, blend expressions often appear and lead to doubt between facial expressions even by human observers [50]. Compared to the already reported models, the TBM based model allows dealing with these considerations.

To summarize, the proposed work is a new development of a previously proposed model [2] significantly modified here for the automatic and dynamic recognition of pain expression in video sequences. The originality of the proposed method compared to the already proposed models for the recognition of pain expression lies in: (1) the automatic fusion of different sensors information: the permanent facial feature deformations (i.e. eyes, eyebrows, and mouth), the transient features information (i.e., nasal root wrinkles), and the context of the expression production (i.e. the place like a hospital); (2) dynamic and progressive fusion (i.e. taking into account asynchronous facial feature deformations) of the permanent facial feature behavior rather than combining the result of static classification for the recognition of pain expression in a video sequence; (3) modeling explicitly the imprecision and the uncertainty of the sensors detection in the fusion process based on the dynamic modeling of the TBM [33,34]; and (4) the recognition of pain expression among the other six basic facial expressions as well as neutral.

2. Automatic system

The current paper proposes a new development of a previously proposed model for static classification of the six basic facial expressions and neutral [2,61]. In the current paper, the previous model is generalized for the automatic and progressive recognition of pain expression in video sequences using a dynamic fusion of facial and contextual information. Fig. 1 shows an overview of the proposed model where the dashed square corresponds to the new development proposed in the current paper. First, in order to retrieve asynchronous information on the permanent facial feature states, a temporal refinement process is introduced (temporal evolution model and the asynchronous information); second, the detection and analysis of the transient facial features (e.g., nasal root wrinkles); and third, the introduction of the contextual information (e.g., place).

The first step of the proposed model is the automatic detection of the facial features (i.e., eyes, eyebrows, mouth, and wrinkles) and the context and then the representation of the corresponding behavior.

2.1. Features detection

2.1.1. Permanent facial features

The first step of the facial expression model [2] is the extraction of the contours of the permanent facial features (eyes, eyebrows, and mouth, a detailed description can be found in [37]). To cope with the illumination variation, a preprocessing stage based on a model of the human retina [37] is applied. The used retinal preprocessing enhances the contours and at the same time, realizes a local correction of the illumination variations (see Fig. 2).

A specific parametric model is then defined for each deformable feature (eyes, eyebrows, and mouth). Several characteristic points are extracted in the image to be processed to initialize each model (for example, eye corners, mouth corners, and eyebrow corners). In order to fit the model with the contours to be extracted, a gradient flow (of luminance and/or chrominance) is used to estimate the contours. The chosen models are flexible enough to produce realistic contours for the mouth,
the eyes, and the eyebrows (Fig. 3 shows an example of the obtained contours during happiness, surprise, and disgust facial expressions). Intensive tests (a comparison with a manual ground truth) on several databases under various face sizes (corresponding to a radius between 4 and 15 pixels) presenting a large range of varying conditions show the accuracy and the robustness of the method to specularities, luminance conditions, ethnicity, and facial expression deformations [37].

In the current paper, based on the segmentation results of the first frame, the set of characteristic points corresponding to the detected facial features (see Fig. 6 left) are selected and tracked in the remaining of the sequence. The algorithm we used to track these facial points is the Lucas-Kanade feature-tracking algorithm [38]. To be the most robust possible, the characteristic point positions are re-detected automatically at each eye blink [37].

2.1.2. Transient facial features

Based on the eye corner positions (Fig. 6 left) and a set of morphological constraints the transient facial feature area is first located (see Fig. 4). Inside each selected area the Canny edge detector is applied for the automatic detection of the wrinkles appearance. The presence or absence of wrinkles is made by comparing the number of edge points in the current expressive image with the number of edge points of a neutral facial image. If there are about twice more edge points in the current image than in the neutral image (neutral expression), wrinkles are considered to be present. The Canny edge threshold is set by expertise but is kept constant over all the databases and at a high value in order to minimize the risk of errors. In Section 3.2.2 we will describe how the proposed TBM based model allows the system to report doubt instead of taking the chance of making a wrong decision in the case of uncertainty on the appearance of wrinkles.

2.1.3. Context information

Two main issues need to be addressed during the introduction of the context of the application: (1) the context source detection and (2) the use of the context related information. It should be stressed that this work does not focus on how to extract context information, which can be performed by a complementary system from other research fields, but rather shows how such a system can be easily added as a new sensor to decide which expressions are more likely to occur (i.e. the context changes the a-priori probabilities). As a knowledge modeling process the proposed TBM based model can easily model several context variables and the corresponding information for facial expression classification (reported by [26] as the most important issue at introducing the context in the classification process). The purpose of the current model is the automatic recognition of pain expression of videotaped people in a medical environment. For this end, several context variables can be defined (such as the place, the task, the answer to a written questionnaire, etc.) allowing the elimination of a set of unexpected facial expressions. Considering a hospital context application (videotaped waiting room, aged people under camera monitoring, injured people), the aim of the proposed work is to automatically alert medical personnel – physician or nurse – that the patient in charge is suffering pain when this one is unable to report his pain by himself (see Fig. 5). The place variable is the most general and suitable context variable for our purpose. The place variable is defined with two values, medical or not medical (considering the expressor in the medical environment like hospital or not), and allows the support of the expected facial expression of pain (the modeling process of the context variable and the corresponding information are explained in Sections 2.2.2 and 3.2.2). A physician or a nurse using a simple pressure button, which changes the system parameters towards pain monitoring, can assess the medical context. Otherwise the whole set of possible facial expressions are expected without favoring any of them (see Fig. 5). The proposed modeling is only one among several possible modeling to introduce the context information. The proposed model can be easily generalized to incorporate several other context variables according to the studied facial expressions and their context dependency.

The strength of the proposed model is that the context information is easily added as a "reinforcement sensor" in addition to the permanent and transient facial feature behavior for the classification process. Once all required information is collected, it is incorporated into a fusion architecture based on the Transferable Belief Model.

2.2. Representation of features behavior

2.2.1. Facial feature states

Five important characteristic distances ($D_0$, $D_1$, $D_2$, $D_3$, $D_4$) [61] measure the permanent facial feature deformations occurring during each facial expression according to the neutral state (see Fig. 6). Each
distance is normalized with respect to the distance between the centers of both irises in the analyzed face. This makes the analysis independent of the variability of face dimensions and of the position of the face with respect to the camera. In the present work, we assume that facial expressions and the corresponding facial feature behaviors are symmetrical. Each characteristic distance value $D_i$ is then considered as the mean of its corresponding left and right side values. The relative importance of each of the five characteristic distances for the recognition of each one of the basic facial expressions and neutral has been recently examined (a detailed description of the specific contribution of each one of the characteristic distances can be found in [61]). The obtained results highlighted that the five characteristic distances are necessary for the classification of the six facial expressions and neutral, and then also for dissociating them from pain expression.

A numerical to symbolic conversion is carried out using a fuzzy-like model for each characteristic distance $D_i$ (see Fig. 7). It allows the conversion of each numerical value to a belief in five symbolic states reflecting the magnitude of the deformation according to the neutral state $S_0$: the current distance is roughly equal to its corresponding value in the neutral expression, $C^+$ (vs. $C^-$); if the current distance is significantly higher (vs. lower) than its corresponding value in the neutral expression, then $S_0$ is $C^+$ (vs. $C^-$); if the current distance is neither sufficiently higher (vs. lower) to be in $C^+$ (vs. $C^-$), nor sufficiently stable to be in $S_0$. This conversion process has been validated on the three benchmark databases (the Cohn–Kanade database [39], the CAFÉ database [40], and the Hamanow-Capriol database [41]) for the recognition of the six basic facial expressions plus neutral [251]. Based on the same mapping, the analysis of the behavior of each characteristic distance (i.e., the permanent facial feature behavior) during pain expression sequences led to the definition of the rules (Table 1). This mapping has been obtained from the pain expression sequences validated by human observers on the STOIC database [36]. Different rules have been evaluated (i.e., with the corresponding inference model). From all these models, the most efficient rules (leading to the best performances) have been selected. The rules correspond to the facial feature deformations leading to the maximization of the correlation between the human and the system performances (see Table 1). In the current paper we reported the behavior of the characteristic distances for
Table 2
Transient feature and context states during pain expression.

<table>
<thead>
<tr>
<th>Facial Root Wrinkles</th>
<th>Pain</th>
<th>Anger</th>
<th>Disgust</th>
<th>Happiness</th>
<th>Surprise</th>
<th>Fear</th>
<th>Sadness</th>
<th>Neutral</th>
</tr>
</thead>
<tbody>
<tr>
<td>Context</td>
<td>P</td>
<td>P</td>
<td>P</td>
<td>P</td>
<td>P</td>
<td>P</td>
<td>P</td>
<td>P</td>
</tr>
<tr>
<td></td>
<td>MC</td>
<td>MC Or NMC</td>
<td>MC Or NMC</td>
<td>MC Or NMC</td>
<td>MC Or NMC</td>
<td>MC Or NMC</td>
<td>MC Or NMC</td>
<td>MC Or NMC</td>
</tr>
</tbody>
</table>

only pain expression. Their behavior (and then the corresponding characteristic distance states) during the six basic facial expressions has been clearly reported in previous publications (the interested reader can refer to [2,61] for the detailed description of the rules of the six basic facial expressions).

In addition to the permanent facial features, the analysis of the behavior of the facial root wrinkles (i.e., the number of pixels detected as wrinkles in each selected area) during the six basic facial expressions and pain compared to the neutral state is also made. Their behavior shows that they can have three characteristic states: P (meaning present) if the number of pixels detected as wrinkles is significantly higher than their number in the neutral state; A (meaning absent) if their number is roughly equal to its corresponding value in the neutral state; P or A if their number is neither sufficiently higher, to be in P, nor sufficiently stable to be in A. The facial root wrinkles during pain expression are reported in Table 2.

2.2.2. Context variable states

The context information is used to bias the classifier toward the most relevant expression. In the current application, the context is medical and pain expression is the more relevant expression. The context information is coded with two states: MC if the videoed sequence is in a medical context, for example a hospital, and NMC if not. In the current modeling, the context information allows refining the already obtained classification results (based on the facial sensors) where pain is already identified alone or where the system has doubt between it and another expression, discarding irrelevant facial expressions (see Section 4). In the context of pain monitoring (videoed waiting room, injured people), we deliberately choose to take a risk of having a false alarm in the case of doubt between pain and another expression (by biasing the classifier toward the most relevant expression, i.e., pain) rather than missing a painful state. In this case an alert is sent and a human observer (physician or nurse) can then confirm or not this information. The proposed modeling is the first attempt to introduce context information. This modeling can be however easily generalized to several other context variables (i.e., MC and NMC variables can be replaced by several other variables) as a selection sensors of relevant expressions (i.e., different expressions according to each context variable) according to the applications (e.g., context-based recognition of different kind of smiles) by a simple addition of the corresponding rules in Table 2.

At this point, pain expression is modeled by a set of rules according to the sensor states (see Tables 1 and 2). It should be stressed, however, that humans are not all or nothing, by it for the production or the recognition of emotional facial expressions [50]. Pain expression as all facial expressions may include a blend of expressions, which makes human observers often hesitating between several expressions (see Section 4.2.2.2). Moreover, automatic facial features segmentation can lead to measurement errors on the characteristic distance states as well as on the transient feature states. For these reasons, an all-or-nothing system based only on logical rules is not sufficient to reliably recognize facial expressions. These issues can be directly tackled by the TBM [33] modeling process of all the knowledge resulting from the characteristic distance states, the facial root wrinkles, and the context information.

3. Knowledge representation based on the Transferrable Belief Model

In information fusion, it is usually important to take into account the reliability of the different sources in the evidence aggregation process [70]. Indeed, in a realistic interaction environment, a facial expression analyzer should be able to deal with noisy and partial data and to generate its conclusions confidently with respect to the data. Compared to the already reported model for facial expression recognition the TBM is a model for reasoning that is well suited to represent imprecise and uncertain information using belief functions [43]. The TBM can be considered as an extension of Bayes's theorem in which probability measures are replaced by belief functions, and where no prior knowledge is assumed [75]. The TBM deals with imprecise and uncertain information by the explicit definition of doubt associated to the sensors information [2]. It provides a theoretical framework for the combination of various independent sensors or sources of information to obtain a more reliable decision [33,34]. In the following, we will introduce the main concepts of this theory and some essential notions that will be used in the proposed approach.

3.1. General concepts

The TBM modeling requires the definition of the basic belief assignment (BBA) associated to each independent source of information.

3.1.1. Basic belief assignment

Let $x$ be a variable taking values in a finite set $\Omega=\{H_1,...,H_n\}$ of $N$ exclusive and exhaustive hypotheses called frame of discernment and characterizing some situations. This means that the solution is unique and is one of the hypotheses of $\Omega$. In our case, the hypotheses correspond to one of the facial expressions and the space of discernment is $\Omega=\{\text{happiness}(F_1), \text{surprise}(F_2), \text{disgust}(F_3), \text{fear}(F_4), \text{anger}(F_5), \text{sadness}(F_6), \text{pain}(F_7), \text{neutral}(F_8)\}$.

The belief related to the variable $x$ is modeled by a basic belief assignment function (BBA) that assigns a piece of evidence (the equivalent of the probability in the Bayesian formalism) to the proposition $A$ of the power set: $2^\Omega=\{A\subseteq\Omega\}=\{\emptyset,\{H_1\},\{H_2\},...,\{H_n\}\}$. Such as

$$m^B(A):=\sum_{A\subseteq\Omega} m^B(A) = 1$$

where $m^B(A)$ is the part of belief on the proposition $A$ without favoring any of the propositions of $A$. This is the main difference with the Bayesian model, which implies equiprobability of the propositions of $A$. The subsets (or instances $\{H_i\}$) correspond to the logical propositions (for instance $H_1$ is true). In the following, the logical propositions are used instead of the subset, which means that the notation $\{H_i\}$ will be simplified to $H_i$ and $\{H_i,H_j\}$ to $H_i\cup H_j$ (i.e., $\text{happiness}(F_1)$ is the local element of $m^B(A)$ whenever $m^B(A)>0$. The local ignorance is represented by the BBA $m^B(\Omega)=1$. Moreover,
with TBM, the BAs are not required to be normalized, i.e., we may have $m^H(\phi) \propto \Omega(\phi)$ being the conflict between the sensors).

3.1.2. Dempster's rule of combination

The main feature of TBM is the set of powerful combination methods to fuse information from different sensors. The usual way of combining the beliefs resulting from each source of information (in the current case the visual cue and context states) is the Dempster's rule of combination (conjunctive sum operation), see Eq. (10) (33,34).

$$m_{\Omega} = \bigvee m_{\Omega}^H$$

(2)

In the current application CircLe (P, TF, CT), 1 ≤ i ≤ 5.

For example, if we consider two BAs $m_{\Omega}^1$ and $m_{\Omega}^2$ of the same frame of discernment $\Omega$. The joint belief $m_{\Omega}^3$ is given using the conjunctive combination (orthogonal sum) as

$$m_{\Omega}^3(H) = m_{\Omega}^1(H) \oplus m_{\Omega}^2(C) = \sum_{H \cap C} m_{\Omega}^1(H)m_{\Omega}^2(C)$$

(3)

where $A$, $B$, and $C$ correspond to hypotheses or subset of hypotheses.

3.1.3. Decision process

The decision is the ultimate step and consists in making a choice between various hypotheses (i.e., $\Omega$) and their possible combinations. However, making a choice means taking a risk, except if the result of the combination is perfectly reliable (i.e., $m_{\Omega}^H = 1$). As this is not always the case, several classical criteria can be used: the plausibility, which favors the single hypotheses in the case of mixture of expressions, and the belief, which favors the mixture of hypotheses (33). However, in some applications it is necessary to make a decision and choose the most reliable single hypothesis. To do so, Stieger proposed the use of the Pignistic transformation defined as

$$Bel(P) : \Omega \rightarrow [0,1]$$

$$I\cdot Bel(I) = \sum_{H \in \Omega} m^H(\phi)\cdot Car(I)(H) \quad \forall I \in \Omega$$

(4)

where $Bel(I)$ corresponds to the Pignistic probability of each of the hypothesis $I$ of $H$, $\phi$ corresponds to the conflict between the sensors, and $Car(I)(H)$ corresponds to the number of elements (hypotheses) of $H$.

3.2. TBM based modeling

The first step of the TBM modeling process is the definition of the BAs of each sensor (characteristic distance states, transient feature states, and context state) and then the combination of these BAs leading to the definition of the BAs of the facial expressions.

3.2.1. Basic belief assignment of the transient feature states

In this case the BAs of the facial root wrinkle states is defined as $m_{\Omega}^{\text{wr}}$ (see Eq. (1)) where $\Omega_{\text{wr}} = \{P,A\}$. From the power set $\mathcal{P}(\Omega)$ only the state $\Phi$ (we are sure that the wrinkles are present) and the state $P \cup A$ ("we don't know," corresponding to the doubt on their state) are considered leading to the subset $\mathcal{P}(\Omega_{\text{wr}}) = \{P,A,\Phi\}$. The wrinkles detection threshold has been derived by statistical analysis on the Hammam-Capdeville and the STOG databases (41,36). It is defined as the mean of the ratio of the number of the wrinkle pixels when the facial root wrinkles are present (ground truth) compared to their number in the case of neutral state. Thus, if the number of wrinkle pixels is higher than the threshold, the system is sure that the facial root wrinkles are present ($m_{\Omega}^{\text{wr}}(P) = 1$); otherwise the system keeps the doubt instead of taking the chance of making a wrong decision ($m_{\Omega}^{\text{wr}}(P \cup A) = 1$). Indeed, due to the difficulty of defining a reliable threshold for the intermediate values between the two classes $P$ and $\Phi$, binary beliefs are chosen with a high and precise threshold for the detection of the transient features. This choice can lead to some miss detections of the transient features. These miss detections will eventually lead to keeping the doubt between several expected facial expressions instead of taking the risk of making a wrong decision by eliminating one expected expression because of a false detection of the transient feature. The nasal root wrinkles are used for pain identification as reported in Table 2. Their presence allows a refinement of the classification eliminating the expressions happiness, surprise, fear, sadness, and neutral, reducing the number of the possible expressions to 3 rather than 5 as follows:

- if the nasal roots are present the current expression is pain or anger or disgust (without favoring any of them, see Table 2) and the corresponding piece of evidence is computed as

$$m_{\Omega}^{\text{wr}}(P) = m_{\Omega}^{\text{wr}}(\text{pain} \cup \text{anger} \cup \text{disgust}) = 1$$

- if they are absent the current expression is one of the 8 expressions and the corresponding piece of evidence is computed as

$$m_{\Omega}^{\text{wr}}(P \cup A) = m_{\Omega}^{\text{wr}}(\text{pain} \cup \text{anger} \cup \text{disgust} \cup \text{happy} \cup \text{surprise} \cup \text{fear} \cup \text{sadness} \cup \text{neutral}) = 1$$

3.2.2. Basic belief assignment of the context variable state

In this case the BAs of the contextual variable is defined as $m_{\Omega}^{\text{cont}}$ (see Eq. (1)) where $\Omega_{\text{cont}} = \{MC, NMC\}$. The contextual variable means medical context (the expresser is in a medical context and it is more likely that the expected expression corresponds to pain) and $NMC$ means not medical context. $MC$ or $NMC$ means that the context of the expresser is unknown. From the power set $\mathcal{P}(\Omega)$ only the state $MC$ (the current context is medical) and the state $NMC$ (context unknown) are taken into account, leading to the subset $\mathcal{P}(\Omega_{\text{cont}}) = \{MC, NMC\}$. In the current modeling the context variable is discrete, then, if the context is medical, the piece of evidence of the state $MC$ is equal to 1 ($m_{\Omega}^{\text{cont}}(MC) = 1$); otherwise the piece of evidence of the state $NMC$ is equal to 1 ($m_{\Omega}^{\text{cont}}(NMC) = 1$). The context information is used for pain selection as reported in Table 2. The piece of evidence of the corresponding expressions is computed as follows:

1. if the context of the application is medical, and then the aim is to know if the current videoed face is in pain or not, the piece of evidence of the corresponding expression is

$$m_{\Omega}^{\text{cont}}(MC) = m_{\Omega}^{\text{cont}}(\text{pain}) = 1$$

2. otherwise the piece of evidence of the corresponding expressions is

$$m_{\Omega}^{\text{cont}}(MC \cup NMC) = m_{\Omega}^{\text{cont}}(\text{pain} \cup \text{anger} \cup \text{disgust} \cup \text{happy} \cup \text{surprise} \cup \text{fear} \cup \text{sadness} \cup \text{neutral}) = 1$$

3.2.3. Basic belief assignment of the characteristic distance states

The BAs of each characteristic distance states $D_i$ is defined as $m_{\Omega}^{D_i}$ (see Eq. (1)), where $\Omega_{\text{cont}} = \{C_i, \bar{C}_i, S_i\}$. $\bar{C}_i - C_i \cup S_i$, $C_i \cup S_i$ is the set of possible focal elements. The states $C_i \cup \bar{C}_i$ and $S_i \cup C_i$ are not allowed, leading to the subset $\mathcal{P}(\Omega_{\text{cont}}) = \{C_i, \bar{C}_i, S_i\}$. The piece of evidence $m_{\Omega}^{D_i}$ associated with each symbolic state, given the
value of the characteristic distance \(D_0\) is obtained by the function depicted in Fig. 7. The normalized threshold values \((\alpha_0, \beta_0, \gamma_0, \delta_0, \epsilon_0, \zeta_0, \eta_0)\) have been derived by statistical analysis on the Hammal-Capler database \([41]\) for each characteristic distance. More precisely, the obtained thresholds have been computed based on the normalized average of minimum, maximum, and neutral values for each characteristic distance \(D_i\) for all the facial expressions and all the subjects. The generalization of these thresholds has already been proved on several benchmark databases such as the Cohn-Kanade database \([39]\) and the CAPFER database \([40]\) for the recognition of the six basic facial expressions and neutral. The interested reader can find a detailed description of their computing in \([2,61]\). The same thresholds, without re-setting, are used here.

As reported in the introduction, temporal dynamics of human facial behavior is a critical factor for the interpretation of the facial expressions \([7,44]\) and is moreover essential for the categorization of complex psychological states, like various types of pain and mood \([45]\). In the following, we take into account the dynamic behavior of the permanent facial features for the classification process introducing a dynamic FSM based model.

3.3. Temporal refinement

The temporal information is introduced at two levels: first, by taking into account at each time \(t\) the previous information (facial feature states) at time \(t-1\) to prevent some miss classifications that can appear due to wrong distance states estimation (temporal evolution in Fig. 1); second, by combining at each time all the previous information from the beginning until the end of the sequence to make a decision (asynchronous information in Fig. 1) taking into account the dynamic and asynchronous behavior of the facial features during the whole sequence. This is the main difference with the previous works, which take into account only the latest information at time \((t-1)\) or combine the information of a predefined number of frames \([8,20,65]\).

3.3.1. Temporal evolution model

The temporal evolution model applies a temporal constraint on the evolution of the BBAs. The proposed model allows predicting the BBAs of the permanent facial feature states at each time \(t\) according to their BBAs at time \(t-1\). In other words, the model predicts the BBAs \(m_{0,t}^{n}\) of each characteristic distance at time \(t\) according to the corresponding BBAs \(m_{0,t-1}^{n}\) at time \(t-1\). It is assumed that the two BBAs are close because the information between two consecutive frames is strongly related. Moreover, in the case of everyday life applications, it has been shown that spontaneous facial expressions are slower in time than posed facial expressions \([10,44]\).

The temporal evolution model needs first the definition of the conditional pieces of evidence, gathered in a "transmission matrix" for each characteristic distance. The predicted BBA is computed at each time \(t\) by the combination of a transmission matrix \(M(D_t)\) and the computed BBAs at time \(t-1\) in the following way:

\[
\begin{align*}
\tilde{m}_{0,t}^{n} &= M(D_t)m_{0,t-1}^{n} \\
\tilde{m}_{0,t} &= \begin{pmatrix}
m_{0,t}^{n}(S) \\
m_{0,t}^{n}(C) \\
m_{0,t}^{n}(S\cup C)
\end{pmatrix} \\
m_{0,t-1} &= \begin{pmatrix}
m_{0,t-1}^{n}(S) \\
m_{0,t-1}^{n}(C) \\
m_{0,t-1}^{n}(S\cup C)
\end{pmatrix}
\end{align*}
\]

The transition matrix \(M(D_t)\) is composed of a distribution of conditional pieces of evidence \([47]\). These pieces of evidence are associated to the transitions from each proposition \(A\) of the frame of discernment at time \(t-1\) (previous frame) to each one of the possible propositions \(A\) at time \(t\) (current frame), and is denoted as \(m_{0,t}^{n}(A|B)\) \((A\text{ and }B\in\{S,C,S\cup C\})\). For example, for a considered distance \(D_0\), \(m_{0,t}^{n}(S|C)\) corresponds to the piece of evidence (the belief) \(m_{0,t}^{n}(C)\) associated with the state \(C\) at time \(t\) (for the distance \(D_0\) given the fact that \(D_0\) was in the state \(S\) at time \(t-1\). The conditional piece of evidence for each characteristic distance \(D_0\) \((1\leq i\leq 5)\) is gathered in a specific transition matrix \(M(D_t)\) as:

\[
M(D_t) = \begin{pmatrix}
m_{0,t}^{n}(S|S) & m_{0,t}^{n}(S|C) & m_{0,t}^{n}(S|S\cup C) \\
m_{0,t}^{n}(C|S) & m_{0,t}^{n}(C|C) & m_{0,t}^{n}(C|S\cup C) \\
m_{0,t}^{n}(S\cup C|S) & m_{0,t}^{n}(S\cup C|C) & m_{0,t}^{n}(S\cup C|S\cup C)
\end{pmatrix}
\]

where the sum of all the conditional pieces of evidence belonging to the same column is equal to 1 and the matrix dimension is \([5\times5]\).

The piece of evidence of all the transitions are learned using the Hammal-Capler database and have already been validated for the three expressions: happiness, surprise, and disgust \([46]\). More precisely, for each expression, for each subject, and each sequence the conditional basic belief assignments are defined between two consecutive frames \((t-1, t)\) from the beginning to the end of the sequence as \((m_{0,t-1}^{n}(S) \rightarrow m_{0,t}^{n}(S)), (m_{0,t-1}^{n}(C) \rightarrow m_{0,t}^{n}(C)), \ldots, (m_{0,t-1}^{n}(S\cup C) \rightarrow m_{0,t}^{n}(S\cup C)))\), \(2 \leq t \leq N\), where \(N\) corresponds to the total number of frames per sequence.

A conditional matrix \(M_t(D_t)\) is thus defined for each distance \(D_0\) for each subject \(j\) and for each expression \(e\) by:

\[
m_{0,t}^{n} = M_t(D_t)m_{0,t-1}^{n}.
\]

Eq. (6) is defined for each transition between \((t-1, t)\). To obtain the transitions on the whole sequence, the \(m_{0,t}^{n}\) and \(m_{0,t-1}^{n}\) are concatenated as:

\[
\begin{pmatrix}
m_{0,t}^{n}(S) & \ldots & m_{0,t}^{n}(S|S) & m_{0,t}^{n}(S|C) & m_{0,t}^{n}(S|S\cup C) \\
m_{0,t}^{n}(C) & \ldots & m_{0,t}^{n}(C|S) & m_{0,t}^{n}(C|C) & m_{0,t}^{n}(C|S\cup C) \\
m_{0,t}^{n}(S\cup C) & \ldots & m_{0,t}^{n}(S\cup C|S) & m_{0,t}^{n}(S\cup C|C) & m_{0,t}^{n}(S\cup C|S\cup C)
\end{pmatrix}
\]
the predicted and computed BBAs is based on the conjunctive combination (see Eq. (2)) as
\[ m_{\text{pred}} = m_{\text{comp}} \odot m_{\text{evidence}}. \]

At each time \( t \) the BBAs of each characteristic distance \( D_i \) is then re-initialized by the resulting BBA as
\[ m_{\text{pred}} = m_{\text{evidence}}. \]

The combination leads sometimes to a conflict between the predicted and the computed pieces of evidence. This is mainly due to segmentation errors, so in this case, the BBA obtained by prediction \( m_{\text{pred}} \) is chosen to form the BBA associated to the corresponding characteristic distances at time \( t \).

1.3.2. Asynchronous behaviour modeling

Facial expression is the result of progressive deformations of a set of facial features appearing at different times (asynchronously) without any defined appearance order [48]. The followed modeling deals with these considerations taking into account, in addition to the information at time \( t-1 \) (reputed in the previous section), at each time (each frame) all the available information (BBAs of facial feature states) from the beginning until the end of the sequence to make a decision.

Spontaneous facial expressions are characterized by a beginning, one or more apaxes, and an end [7] (see Fig. 8). The beginning corresponds to the first frame where one or more facial features begin to move. The end corresponds to the frame where all the facial features come back to their neutral state (stable). In each expression sequence, the proposed model detects the beginning as the first frame where at least one of the permanent facial features (and then the corresponding characteristic distances) is no longer to the stable state \( S \), and the end as the first frame (after the beginning) where all the permanent facial features (and then the corresponding characteristic distance states) have come back to the stable state \( S \). However, there is no way in the proposed framework, and this is still an open issue in the community, to detect in real time the apexes of one expression sequence. The proposed method deals with this consideration taking into account all the available information (previously facial features deformation) between each pair of beginning and end frames to make a decision. The BBAs of the characteristic distance states are then redefined on the whole sequence and combined based on the rules table (see Table 1) to define the expression corresponding to all these deformations.

Once the beginning of the expression has been detected, the analysis of the distance states is made inside an increasing temporal window \( \Delta t \). The size of the window \( \Delta t \) increases progressively at each time from the beginning until the end of the expression (see Fig. 9). At each time \( t \) (each frame) inside the

![Fig. 8. Example of temporal evolution of spontaneous happiness expression.](image)

![Fig. 9. (a)-(d) show examples of the increasing temporal window during a sequence of pain expression: each row displays the temporal evolution of each distance state. It can be observed that with the temporal evolution each feature evolves asynchronously but the increasing window (gray window) includes all its behavior from the beginning until the end of the sequence.](image)
temporal window, the whole set of the previous information (the past states of the characteristic distances (computed as reported in Section 3.3.1) and the corresponding facial feature features) is taken into account to classify the current expression sequence. This processing allows explicitly dealing with the dynamic of the facial expressions and more importantly with asynchronous aspect of the facial feature deformations [7]. Fig. 9 gives an example of asynchronous facial feature deformations during pain expression sequence. In this case, combining the classification results coming from the obtained information at each time separately (frame by frame) can lead to a wrong decision. Thus, in the proposed model, once the beginning of the expression is detected, the dynamic classification consists in computing at each time $t$ ($k$ in Fig. 9) the BBAs of the characteristic distance states defined on $(C^+, C, S, U, C^-, S, U, C^-, U)$ according to their behavior from the beginning until the current frame. To do that, at each time $t$ ($k$ in Fig. 9), inside the current increasing window $\Delta t$ and for each characteristic distance, the whole of its behavior from the beginning is used to select the associated state. The selection is made according to the "integral sum of plausibility" of each state divided by its number of appearances. The number of appearances of each symbolic state is noted as $N_{AP}(state)$, state $\in (C^+, C, S, U, C^-, S, U, C^-, U)$ (see Eqs. (11) and (12) for an example) and their integral sum of plausibility is noted $PI_{AP}(state)$ and is computed inside the temporal window $\Delta t$ (see Eq. (12) for an example). The plausibility is one of the many decision criteria, which could be used in the TRM modeling framework [48,49]. It is used to choose from a set of hypothesis, or a subset of hypothesis, the most plausible one. In the current paper, the plausibility is used to reinforce the first condition in Eq. (11) (in the case where two states appear the same number of times for a given characteristic distance (see below)).

For instance, for a characteristic distance $D_2$, and for the state $C^+$ the selection parameter is defined as:

$$K_{PI}(C^+) = \begin{cases} 1 & \text{if } N_{AP}(C^+) > 0 \\ 0 & \text{otherwise} \end{cases} \quad 1 \leq t \leq \Delta t \quad (10)$$

$$\Pi_{AP}(C^+) = \frac{\sum_{t=1}^{\Delta t} K_{PI}(C^+)}{\Delta t} \quad (11)$$

$$\Pi_{PI}(C^+) = \frac{\Pi_{AP}(C^+)}{1} \quad \Pi_{PI}(C^-) = \frac{\Pi_{AP}(C^-)}{1} \quad 1 \leq t \leq \Delta t \quad (12)$$

where $K_{PI}$ indicates the occurring, or not, of a symbolic state at time $t$.

From the two parameters $N_{AP}(state)$ and $PI_{AP}(state)$, the distance states at each time $t$ inside the temporal window $\Delta t$ are selected as:

$$\text{states}(D_i) = \text{max}(\Pi_{AP}(state)/N_{AP}(state)) \quad 1 \leq i \leq 5 \quad \text{state } \in (C^+, C, S, U, C^-, S, U, C^-, U) \quad (13)$$

Fig. 10 shows three examples of the temporal evolution of the states of the characteristic distance $D_2$ inside the increasing window and the corresponding selected states according to Eq. (13).

The piece of evidence associated to each chosen state for each characteristic distance corresponds to its maximum piece of evidence inside the current temporal increasing window as:

$$\Pi_{max}(D_i) = \text{max}(\Pi_{AP}(state)) \quad 1 \leq i \leq 5 \quad (14)$$

Finally, at each time $t$ (see time $k$ in Fig. 9a-d) between the beginning and the end of the expression sequence, once the BBAs of all the characteristic distances are defined (the selected states and the corresponding pieces of evidence), the current expression or the subset of possible expressions is selected (according to the permanent facial features behavior) according to the rules table (see Table 1). The obtained results are then fused to the information resulting from the nasal root wrinkles and the context variable to give the best possible decision (see Fig. 1). The main feature of this classification is that the recognition of the current expression can be made at each time $t$ (each frame) of the sequence.

1. taking into account all the past BBAs of the characteristic distance states (and then the whole dynamic of the corresponding facial features) from the beginning until the current frame [46];

2. at the beginning of the sequence, all the expressions are in the set of possible expressions and, during the sequence, this set is progressively reduced by a dynamic refinement process;

3. when reaching the end (the current frame is then the last frame of the sequence), the decision depends on the whole set of past BBAs of the characteristic distance states and it gives the classification of the entire expression sequence.

3.4. Dynamic fusion process

The main feature of the TRM is the powerful combination operator [33,48] that integrates information from different sensors. In the current paper the sensors are the characteristic distance states, the nasal root wrinkles, and the context variable.

The fusion requires the definition of the fused information on the same frame of discernment (Fig. 11).

From the rule tables (see Tables 1 and 2), and in order to take into account all the available information, the facial expression classification is based on the fusion process of all the used feature states ($D_2$, $FP$, $CT$, see Fig. 11). The BBAs of the characteristic distance states $m_{D_2}^{\text{prev}}$, the transient feature states $m_{FP}^\text{trans}$, and the context states $m_{CT}^\text{cont}$ are defined on different frames of discernment. For the fusion process, it is necessary to redefine the BBAs on the same frame of discernment $\Delta t$, where:

$$\Omega = \text{happiness}(F_1), \text{surprise}(E_2), \text{disgust}(E_3), \text{fear}(E_4), \text{anger}(E_5), \text{sadness}(E_6), \text{pain}(E_7), \text{neutral}(E_8)$$

the set of expressions (pain needs to be identified from the set of $B$ possible facial expressions). From the rule tables (Tables 1 and 2) and the BBAs $m_{D_2}^{\text{prev}}$, $m_{FP}^\text{trans}$, and $m_{CT}^\text{cont}$ of all feature states, a set of BBAs on facial expressions $m_{D_2}$, $m_{FP}$, and $m_{CT}$ is first derived for each characteristic distance $D_2$, transient feature $FP$, and context.
variable \( CT \). In order to combine all this information, a fusion process of their BBAs is performed using the conjunctive combination rule \([33,34]\) described in Section 3.1.2 (see Eq. (2)) and it results in \( m^b\), the BBA of the corresponding expression or subject of expressions. For example, considering two BBAs \( m^b_{E_1}(E_1 \cup E_2) \) and \( m^b_{E_2}(E_2) \) of the same frame of discernment \( Z \), their combination leads to the definition of the expressions \( E_1 \) corresponding to \( E_1 \cup E_2 \cup E_2 \) with the piece of evidence \( m^b_{(E_1 \cup E_2)} = m^b_{E_1} \cup m^b_{E_2} \).

The decision is the ultimate step of the classification process and it allows choosing between various hypotheses \( E_i \) and their possible combinations. The Pignistic probability, which only deals with singleton expressions, is then used as decision criteria (see Section 3.1.3 Eq. (3)). The decision is made using the maximum of Pignistic probability BeLP \([49]\).

4. Experimental results

4.1. Facial expression data

Despite the great efforts in the computer vision community to build and make accessible a large number of databases of the six basic facial expressions, there is still a lack of accessible non-prototypic facial expressions databases including pain expression. In the current paper the performances of the proposed model are evaluated across two new acquired pain expression databases: an acted facial expression database (STOIC database) \([36]\) and a spontaneous pain expression database \([88]\). The databases were recently recorded and ground-truthed by two research groups from Université de Montréal and the Institut Universitaire de Gériatrie de Montréal, experts in facial expression perception \([36]\), especially on pain expression \([68]\). A short description of the two databases acquisition and validation is presented in the following sections.

4.1.1. Acted pain expression database

The STOIC database (see Fig. 12) developed and validated by Roy et al. from the Université de Montréal \([88]\) is used in the current paper. The STOIC database is one of the first dynamic facial expression databases validated by human observers (compared to the usual FACS validation process) independently of the underlying facial expressions.

4.1.1.1. Acted data acquisition and validation. Forty students from theatrical schools aged between 20 and 45 years old were recruited. In all 7000 videos were recorded using a camera located directly in front of the subjects, who were asked to perform the six basic facial expressions as well as pain and neutral expressions \([36]\). Videos were captured and then aligned on a template that fitted in a box centered on the face with a resolution of 256 × 256 pixels \([36]\). Each video begins with the neutral state and evolves until the apex of the underlying facial expression (see Fig. 12). Thirty-five participants (20 females and 15 males) from Montréal were recruited for the validation of the recorded expressions as one of the six basic facial expressions, pain, and neutral independently of the underlying facial expression \([36]\). Eighty-five videos were unambiguously recognized as the six basic facial expressions, pain, and neutral expression \([36]\). Fifteen videos (seven male and eight female actors) from the eighty-five videos were recognized as pain expression. Fig. 12 shows two examples of the obtained pain sequences. A detailed description of the database and a free access can be found in [http://www.mapageweb.umontreal.ca/rostek/STOIC.tar]. The obtained pain sequences were divided into two parts: the training set (10 of the pain videos) and the test set (10 the remaining five pain videos). The training set was used to define the rules for pain expression recognition. The rules correspond to the Facial Action Coding System (FACS) and were explained in Section 2.2.1, the corresponding characteristic distance states leading to the maximization of the correlation between the human and the system performances. This process allows defining a combination rule based on human observer validation (see Tables 1 and 2).

4.1.2. Spontaneous pain expression database

In addition to the acted database and in order to better test the model generalization, with less controlled pain sequences, a spontaneous database was also used. The database was acquired by a research group at the Institut Universitaire de Gériatrie de Montréal working on the relation between facial expressions and the neural networks of pain. The database was recorded during one of their studies and is not freely accessible. A detailed description of the database can be found in \([88]\).

4.1.2.1. Spontaneous data acquisition and validation. The subjects participated in a study on the relation between pain catastrophizing and pain responsiveness of pain in healthy, pain-free individuals \([68]\). Forty subjects (20 females and 20 males) were recorded in total. The subjects knew that they were being recorded and that they were taking part in a pain induced experiment. Pain was induced experimentally by means of a Peltier-based, computerized thermal stimulator (Medoc TST-2001; Medoc Ltd., Ramat Yishai, Israel) with a 3 × 3 cm² contact probe. The warmer probe was attached to the lower leg. Baseline temperature was always set to 35°C. Once uncomfortable (1°C below the individual pain threshold) and two painful thermal stimuli (2-3°C above the individual pain threshold) were applied in a random order. The temperature increased from baseline with a heating rate of 4°C/s to the pre-set temperatures, remained at a plateau for 5 s and returned to baseline with a rate of 4°C/s. ISI varied between 30 to 35 s \([68]\). The faces of subjects were videotaped during the experiments. The video camera was placed in front of the subject at a distance of approximately 4 m. For subject's confidentiality rights we cannot show the images of this database.
(the interested reader can directly contact Kunz et al. [68]). Before applying a stimulus, subjects were always instructed to focus on an emotion-neutral picture being positioned next to the camera in order to ensure a frontal view of the face. Subjects were also instructed not to talk during thermal stimulation. To mark the onset of stimulation on the videotape (for further analysis), the light signal was switched alternatively [68]. The light was visible to the camera but not to the subject. Videos were captured at a resolution of 720 x 576 pixels, out of which the face area spanned an average of approximately 120 x 470 pixels. Each video begins with the neutral state and evolves until the apex of the unfolding facial expression (in the case of pain expressions) or remains at neutral state in the case of neutral videos. Facial expressions displayed on each frame of the obtained videos have been validated as pain expression using the Facial Action Coding System (FACS) [69]. Special software designed for analysis of observational data (the Observer Video-Pro (Noldus Information Technology)) was used to segment the videos and to enter the FACS codes into a time-related database [68]. Ten 5s segments selected for scoring began immediately after the stimulus had reached its maximum. Twenty FACS validated subjects (and the corresponding sequences) were used for the validation process in the current paper. The data correspond to 40 sequences (20 subjects x 2 [neutral + pain] expressions). A large variability was noted between the acted and spontaneous pain expression samples. Compared to a human observer, the proposed model gives comparable classification rates on these data (see Section 4.2.2.2).

4.2. Simulation results

The simulation results were obtained on the STOC database (85 video sequences containing the 6 basic facial expressions, pain, and neutral) and the spontaneous database (40 video sequences containing 20 neutral and 20 pain sequences). A total of 1,25 videos in which 25 different subjects expressed pain: 15 video sequences (15 subjects) from the STOC database (validated by human observers, see Section 4.1.1) and the 20 spontaneous pain sequences (20 subjects, FACS validated, see Section 4.1.2). The number of sequences used for the validation process is in fact a higher number than in all the validation databases used by the already proposed models for pain expression recognition (see Section 1.3). The generalization performances of the proposed model for pain expression recognition were performed, first by establishing the rules only on one part of the subjects of the STOC database and the validation of the model on the whole database (new subjects) and, secondly, by evaluating the model performances on the spontaneous database (20 new subjects recorded in another experimental condition). The proposed results show, first, the performance of the proposed model for the recognition of pain expression in the case of 2-alternative forced choices (pain vs. neutral) and 8-alternative forced choices (the 5 basic facial expressions, pain, and neutral); second, they prove the robustness of the context information; third, they allow comparison between these performances and those obtained by a human observer in similar conditions, demonstrating the usefulness of the model in a real-life application. Two simulations are reported: first, a 2-alternative forced choice classification between pain and neutral to compare the obtained performances to the 2-alternative forced choice classification systems previously developed (see Section 1.3); second, an 8-alternative forced choice classification between pain and the 6 basic facial expressions plus neutral.

4.2.2. Model performance for pain expression recognition

1. Classification performances are reported in two conditions, 2-forced choice classification and 8-forced choice classification on the STOC and spontaneous databases. The validation of the proposed model on spontaneous data allows investigating how the proposed model performed on 20 new subjects expressing “spontaneous” pain expressions, which included head movement with both in-plane and out-of-plane rotations. Thus these spontaneous data allows evaluating the ability of the model to generalize to a new database presenting less controlled pain expression sequences. For example, compared to the acted data (STOC database) even if the subject had exaggerated their expression, which is very unlikely given that pain was physically induced, subjects displayed greater variability than the actors, which challenges even more the robustness of the automatic recognition system. In the following, the obtained classification performances are first reported on each database separately (i.e. STOC and spontaneous). The classification performances (C) (i.e., the percentage agreement with the ground truth), recall (R), precision (P), and F-measure (F), which combine recall and precision as = 2 × recall × precision/(recall + precision), are compared to evaluate the performances of the system in detail. Notably, precision gives an estimation of the amount of false alarms introduced when classifying either pain expression or one of the six basic facial expressions and neutral among 8-alternative choices. Figs. 13 and 14 show the results plots.

4.2.2.1. Performance of the proposed model on 2-alternative forced choice classification

In the first simulation the system performed a 2-alternative forced choice classification between neutral and pain facial expressions. The obtained results are C = 92.3%, P = 96%, and F = 94%. Fig. 13 summarizes the obtained results. For all the studied criteria, the obtained results on acted data are the highest and are better than the performances reported in the

4.2.2.2. Human behavior: experimental material and methods

Human visual system remains far ahead of everything that has been done in automatic facial expression recognition. Thus, in
expressions plus neutral. Interestingly, the performances obtained on spontaneous data compare favorably to those of human observers (see Section 4.2.2.1) on the same data (C: 72.6%, R: 73.5%, P: 80%, and F: 80%). In fact, human and model performances are not significantly different (two-way ANOVA, P ≤ 0.01). Then, even if the obtained performances on spontaneous data are lower than those obtained on acted data they are very encouraging given the fact that the system was not trained on samples of spontaneous pain expression while it is still able to generalize to them. In our point of view this comparison is a more appropriate evaluation of the system's performances considering humans as experts in facial expression recognition.

To summarize, the mean performances obtained on the 2-alternative forced choices are very encouraging (C: 81.2%, R: 78%, P: 84%, and F: 80.5%). They are much higher than the already proposed models for pain expression recognition on acted data and compare favorably to human observer performances in the case of spontaneous data. In daily life situations, in addition to pain expression other facial expressions can be expressed. This is the case for example of the hospital context we introduced previously (see Section 3.1.1). Pain expression needs to be dissociated from at least the six basic facial expressions and neutral to have a better estimation of the classification performances.

4.2.2.2. Performance of the proposed model on 8-alternative forced choice classification. In order to evaluate the robustness of the proposed model for the recognition of pain expression and to discriminate it from the six basic facial expressions and neutral, a second simulation was carried out where the system performed an 8-alternative forced choice classification. This simulation was carried out on the STOIC database (i.e. on the 85 videos displaying the 6 basic facial expressions, neutral, and pain) and on the spontaneous database (i.e. on the 40 videos displaying either pain or neutral expressions). Fig. 14 summarizes the obtained results. The classification performances obtained in the case of 8-alternative forced choices on the STOIC database are C: 82%, R: 85%, P: 75%, and F: 80%. These results are slightly lower than those obtained on the same data for the 2-alternative forced choices (see Section 4.2.2.1). Notably, the precision drops from 80% to 75%. Given the task difficulty (8-alternative choices) the obtained performances remain satisfactory (F-measure ~ 80%). Moreover, similarly to the 2-alternative forced choices, the classification performances on spontaneous database (C: 77%, R: 77%, P: 72%, and F: 74%) show that the obtained precision (72%) is lower (but satisfactory) than the one obtained on the acted database (82%). Moreover, as pain is more relevant in a medical context (modulated in our case by the introduction of the context in the classification process), recall is more important than precision (i.e. in medical context having a false alarm is better than missing a painful state). Interestingly, the obtained performances reflect this characteristic, leading to a higher recall (R: 83%) compared to the precision (P: 75%, R: 72%) in the case of acted as well as spontaneous data respectively. The obtained results on acted and spontaneous data are very encouraging given the difficulty of the task due to the large variation between individuals and the subtle facial feature changes displayed during spontaneous expressions. It has to be emphasized as well that the model was not trained on spontaneous data, but only one part of the STOIC database, which shows its ability to generalize to other databases. Unfortunately, there are no available results in the state of the art classifying pain from the 8-alternative forced choices (neither on acted or on spontaneous data) to compare with. The only possible comparison was with the performances obtained by human observers. Based on the
Table 3
Misclassification results (%) of acted and spontaneous pain sequences in the case of an 8-alternative choice with the context variable.

<table>
<thead>
<tr>
<th></th>
<th>Happy</th>
<th>Disgust</th>
<th>Surprise</th>
<th>Fear</th>
<th>Sadness</th>
<th>Anger</th>
<th>Pain</th>
<th>Neutral</th>
<th>Sadness, anger, pain</th>
<th>Sadness, anger</th>
<th>Ignorance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acted pain</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>92</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Spontaneous</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>77</td>
<td>0</td>
<td>0</td>
<td>15</td>
<td>8</td>
</tr>
</tbody>
</table>

same data (STOIC database) the obtained results in the case of 8-alternative forced choices compare favorably to the performances of human observers (76%, 75%, 80%, 80%). Importantly, both humans and the model show comparable precision levels and their overall performances are not significantly different (two-way ANOVA, P > 0.01). However, compared to the performances of the proposed context-based model, recall is lower than precision in the case of human observer performances. We explain these results by our implementation of the dynamic fusion process, which uses the most important visual information for the human observer [25,26] and especially the refinement process due to information conveyed by the context variable (as described in Section 2.2.4). It is important to note that people usually do not display "pure" facial expressions [26], which introduces doubts between expressions. Indeed, the 8-alternative forced choice behavioral experiment on the STOIC database, reported that human observers misclassify pain expression as sadness, disgust, or happiness. In addition to the performances reported in Fig. 11, a refined analysis shows that before the introduction of the context information different kinds of confusions appear according to the used database. Compared to the human observers, the same doubt states appeared between the same expressions on STOIC database (i.e., doubt between sadness, disgust, and happiness). Notably the obtained results are supported by the results of Ray et al. [25,26], where both human observers were asked to classify pain expression with the 8-alternative forced choice paradigm (also on the STOIC database). The authors reported that humans also misclassify pain with sadness, disgust, and happiness expression. More interestingly, we obtain the same confusions when we asked naive human observers to recognize the same data. These results prove the ability of the proposed model to report doubt in the case of confused expressions, where even human observers cannot discriminate between them without context. The same analysis shows as well that in the spontaneous database, these doubts are no longer between the expected expressions: sadness, disgust, and happiness but between sadness and anger expressions. In these cases the system is sure that the current expression is one of these 2 (or 3) and never one of the 5 (or 6) other expressions (their corresponding rates are equal to zero). These results further attest that the model shows a striking similarity with the human observers and that the introduction of the context information allows solving these confusions. These results are of first interest as they demonstrate the ability of the proposed model to describe a given facial expression with a set of possible responses (in the case of blend expressions) corresponding to doubt that would also be reported by human observers. In some critical applications with high human validation it is requested this ability may appear to be very important and useful. For example, considering the application in the hospital context (waiting room, sick people under camera monitoring [see Fig. 11]), such information is more than sufficient to alert somebody in charge that the patient is suffering pain (displaying also the possible doubt with one or two other facial expressions rather than making a wrong decision). Medical personnel—physician or nurse—can then confirm or not confirm this information. The introduction of the context information (as described in Section 2.2.4) solves the doubt state in the case of blend expressions, increasing the performances by 15%. Indeed, in the case of pain monitoring application (videotaped waiting room, injured people), the current modeling the context information is used as additional information and allows the support of the more relevant facial expression (i.e., pain) in the case of doubt. Table 3 summarizes the obtained misclassification results with the context information (detailed misclassification of the six basic facial expression can be found in [26,26]).

To summarize, the proposed model performances on the 8-alternative forced choice classification prove its robustness to recognize pain expression on acted as well as on spontaneous data. It shows its robustness to new situations (i.e., applied to new sets of data not used for training). Comparison between the proposed model and the human behavior allows better quantification of the model performances. Despite the fact that humans are more precise than the automatic model, the time required for coding, which is accomplished manually by a human expert, makes the automatic model more suitable for real-time situations.

5. Conclusion

In the current paper we present a new dynamic modeling process for the automatic classification of pain expression. The proposed model is based on a dynamic fusion process of facial and context information for the automatic classification of pain expression from 8-alternative forced choices. The system proves its suitability in dealing with spontaneous sequences and the obtained performances compare favorably to those of human observers, leading to similar doubt between confused expressions. These results are encouraging given the context of the application. They prove that an automatic detection of pain expression in video sequences is a feasible task.

The obtained results compare favorably to the already proposed models for the automatic recognition of pain expression in the case of 2-alternative forced choice classification. Moreover, the proposed model shows its robustness and suitability in the case of a challenging 8-alternative forced choice classification. This opens promising perspectives for the development of the model and its application in everyday-life situations. For example, preliminary results on the recognition of spontaneous pain expression intensities [26] proved its potential to be applied in a hospital context application. Even though in the current paper we focus our work on recognizing pain expression on adult subjects it is of interest to test it on the recognition of newborns' pain expression. Finally, the dynamic and multi-layer model has been tested and validated for pain expression. However, the same approach can be easily applied to other facial expressions. For example, the proposed model is the first contribution where the context information is explicitly taken into account for the classification of facial expressions. This modeling can be easily generalized to take into account several other context variables according to the studied facial expression. Moreover, the original version of the proposed model (static recognition based only on the permanent facial features information) has been extensively validated on benchmark databases. The addition of new sensors and the dynamic information enhanced also the classification.
performances for the six basic facial expressions (this claim is based on the preliminary results obtained in [56,60]).

Acknowledgments

We would like to thank Pierre Rainville and colleagues of the Cetnre de Recherche de l'Universite de Geratique de Montreal and of the Otto-Friedrich University Bamberg, Germany, for allowing us to use part of their non-publically available spontaneous pain expression data. The proposed work is in a continuation of the work of Zakia Hammal on the automatic and dynamic analysis of facial expression using the TBM based modeling. The first version of the proposed work appeared in Hammal (2006) [48].

References


Zahia Havranek received her Ph.D. degree in cognitive science from the Université Joseph Fourier, Grenoble, France, in 2006. From 2006 to 2010, she was a postdoctoral research associate at the Université de Montréal, Montreal, Canada. She is currently at the RoboCust Institute of the Carnegie Mellon University, Pittsburgh, USA. Her main research topics include computer vision, machine learning and visual perception applied to human-machine interaction and emotion recognition. Havranek has published over 30 technical papers in the areas of affective computing, image processing, machine learning and visual perception in prestigious journals and conferences in these fields.

Mohamed R. Masson received his bachelor’s degree (2000) and master’s degree (2002) in Psychology from the University of Berlin (FU, Germany). His doctoral research focused on pain processing in patients with endomteria and he received his Ph.D. degree in Psychology from the University of Bamberg in 2006. During her post-doctoral stay (2007–2008) at the University of Montreal (Canada) he studied which brain activation patterns underlie the regulation of facial expressions. At the moment he is working as a researcher at the University of Bamberg (Germany).