Pain Recognition using Spatiotemporal Oriented Energy of Facial Muscles

Ramin Irani, Kamal Nasrollahi, and Thomas B. Moeslund
Visual Analysis of People (VAP) Laboratory
Rendsborggade 14, 9000 Aalborg, Denmark
{ri, kn, tbm}@create.aau.dk

Abstract

Pain is a critical sign in many medical situations and its automatic detection and recognition using computer vision techniques is of great importance. Utilizes this fact that pain is a spatiotemporal process, the proposed system in this paper employs steerable and separable filters to measure energies released by the facial muscles during the pain process. The proposed system not only detects the pain but recognizes its level. Experimental results on the publicly available pain database of UNBC show promising outcome for automatic pain detection and recognition.

1. Introduction

Pain is an unpleasant sensation that informs us about some (potential) damages or danger in the structure or the function of the body. It causes emotional effects like anger and depression and may even impact on the quality of life, social activities, relationships and our job. Yet pain is one of the most common reasons for seeking medical care, over 80% of patients complain about some sorts of pain [16]. So, for clinical trials and physicians, pain, similar to blood pressure, body temperature, heart-beat rate and respiration, is an important indicator of health. Therefore, reliable assessment of pain is essential for health related issues. That is why in 1995 Dr. James Campbell called the pain assessment as the fifth vital sign and suggested that quality care means that pain is measured and treated [17].

The most popular technique for pain assessment is Patient self-report. It is convenient and does not require special skills, but has some limitations. It includes inconsistent metrics, reactivity to suggestions, efforts at impression management and differences in conceptualizations of pain between clinicians and sufferers [15]. Moreover, self-reporting cannot be used, e.g., with children and those patients who cannot communicate properly due to neurological impairment or those who require breathing assistant. Craig et al. in [6] evidenced that changes in facial appearance can be a very useful cue for recognizing the pain.

In Atul Gawandes recent book [9], it has been shown that periodically monitoring of patients pain level by medical staff improves patients treatment. However, sustained monitoring of patients by this way is difficult, unreliable and stressful. To solve this issue, automatic recognition of pain using computer vision techniques, mostly from facial images, has received great attention over the past few years [3]-[19]. Brahman et al. [3] proposed a binary pain detection approach (pain versus no-pain) using Principal Component Analysis (PCA) and Support Vector Machines (SVM). Ashraf et al. [1] detected the pain using Appearance Active Model (AAM). Littlewort et al. [13] employed a two-layer SVM-based approach in order to detect real pain or posed pain. The above mentioned systems implement a binary classifier, meaning they recognize only two cases of pain versus no-pain, while based on the Prkachin and Solomon Pain Intensity metric [18], pain can be quantized into 16 discrete levels ranging from no-pain (0) to maximal pain (15).

To the Best of our knowledge, there are only few research articles that have estimated the pain level automatically, like those in [11-14]. In [14] a system has been developed which can detect three levels of pain intensity. It uses geometry-based and appearance-based features with a separate SVM classifier for each intensity level of pain. Kaltwang et al. [11] proposed an approach using a combination of appearance-based features, Local Binary Pattern (LBP), and Cosine Discrete Transform (DCT), for detecting intensity levels of pain. They applied a Relevance Vector Regression (RVR) model to predict the pain intensity from each feature set. The above mentioned systems use hand-crafted features like LBP and try different classifiers like PCA, SVM, and RVR to detected and recognize the pain. Though they produce interesting results, they do not consider the dynamics of the face. We have observed during our experiments that pain is exposed on the face through changes and motions of some of the facial muscles. These motions obviously release some energy. The level of the released energy is in direct relationship with the level of the pain. This is exactly the point that we want to exploit in this
paper: we develop a system for pain recognition that measures the level of the released energy of the facial muscles over the time. Changes (activation) of facial muscles during the pain have been previously used for pain recognition in Prkachin and Solomon [18]. However, they do not consider the released energy of the facial muscles, but detect the facial Action Units (AU)s and combine them to measure the pain.

There is not that many research work neither on exploiting the temporal axis nor on exploiting the released energy of the facial muscles for detecting and recognizing the pain. For example, [19] measures the pain over the temporal axis. However, it does not use the released energy of the muscles and is more focused on developing a classifier for pain recognition, which is based on Conditional Ordinal Random Fields (CORF). The only system that uses the released energy of facial muscles is the one developed by Hammal et al. [10]. This system uses a combination of AAM and an energy based filter, Log-normal filter, to estimate four intensity levels of pain. Though this system exploits the released energy of the facial muscles, it does that only on a frame by frame basis, in a spatial domain. The proposed system in this paper exploits the released energy of the facial muscles not only on the spatial domain, but also in the temporal one. To do that, we use a specific type of spatiotemporal filter which is shown to be very useful for extracting information in both spatial and temporal domains at the same time, for other applications, like region tracking in [7], [4].

The rest of this paper is organized as follows: the employed filter and the other details of the proposed system are given in the following section. Section 3 explains the performed experiments and discusses the results that are obtained on a public facial database. Finally, section 4 concludes the paper.

2. The Proposed System

The block diagram of the proposed system is shown in Fig. 1. Following the diagram, given an input video sequence, the faces are first detected, simply using the provided landmarks¹. Then, an Active Appearance Model (AAM) algorithm is used to align the detected faces in different frames of the video to a fixed framework using the provided landmarks. This registration to the fixed framework will cause losing some of the areas of the face, in some of the frames, which appear as holes or lines on the registered faces. To compensate for this, we use an inpainting algorithm. Then, the spatiotemporal filtering is performed in both x, y, and t dimensions to detect the energy released by the facial muscles motion of the aligned faces. Finally, the pain is detected and its level is recognized. These steps are explained in the following subsections.

2.1. Face Detection and Alignment

Detecting the face is an essential step in any facial analysis system, including, pain recognition. The employed database in this paper [18] provides the position of facial landmarks in all the frames of the dataset. We simply use these landmark positions to extract the facial regions in each frame. To do so, as it is shown in Fig. 2a, the facial landmarks are used as vertices of triangles which cover the entire face area, as it is done in [12]. This detected face needs to be segmented from the rest of the image. For this purpose, first, a binary mask (Fig. 2b) is generated such that:

\[
Mask = \bigcup_{k=1}^{K} I_k
\]  

where:

\[
I_k = \begin{cases} 
1 & P_{ij} \in T_k \\
0 & \text{Otherwise}
\end{cases}
\]  

where \(T_k\) is the \(k\)th triangle created by landmark points, \(P_{ij}\) is a pixel on the image located at \((i,j)\), \(I_k\) is a binary image corresponding to \(T_k\) and \(U\) is a union function. Finally, the face can be segmented from the rest of the image by applying the mask on the image (Fig. 2c).

As mentioned before, the proposed system measures the energy that is released due to the motion of the facial muscles. However, in a video sequence, such motions are not the only type of motion. For example, Fig. 3a shows the positions of 66 facial landmarks in a video sequence of 100 frames. If there was no motion in the video at all, one could only see 66 facial landmarks, but as it can be seen in Fig. 3a,
the position of each landmark is changing from one frame to another. This indicates the presence of other motions on the face, like motions resulting from the head pose. Such motions should be filter out. To do that, we employ the face alignment algorithm of [8]. The faces in this algorithm are aligned using the facial landmarks. The results of this alignment, applied to Fig. 3a, can be seen in Fig. 3b.

Having aligned the facial images of the input video sequence and generating an aligned facial video, using the above mentioned steps, the next step is to extract the spatiotemporal features. These features extract the direction and the level of the energies released by the facial muscles. These directions and levels are different for different facial expressions. For example, for a neutral face one should not expect too much energy to be released, while for a laughing face or a face suffering from pain, different levels of energy will be released by the facial muscles in different directions. Extracting of orientation and level of the released energy of
the facial muscles are explained in the following subsection.

2.2. Spatiotemporal Feature Extraction

The extraction of the orientation and the level of the energies released by the facial muscles are done through steerable and separable filters of [4]. These filters compose of a second derivative Gaussian \( G_2(\theta, \gamma) \) followed by a Hilbert transform \( H_2(\theta, \gamma) \), in different directions of \( \theta \), and scales of \( \gamma \). We do not use a multiscale method, because the level of the energy is not that much visible in coarse scales, hence \( \gamma = 1 \). During the pain, however, the facial muscles can move in any directions, but such motions can be decomposed into four main directions. Therefore, we measure the released energies in four main directions corresponding to \( \theta = 0, 90, 180, \) and 270 degrees. The released energy from every pixel is then calculated by:

\[
E(x, y, t, \theta, \gamma) = [G_2(\theta, \gamma) * I(x, y, t)]^2 + [H_2(\theta, \gamma) * I(x, y, t)]^2
\]  

(3)

where * stands for a convolution operator, \((x, y, t)\) shows the pixel value located at the position of \( x \) and \( y \) of the \( t \)th frame (temporal domain) of the aligned video sequence of \( I \), and \( E(x, y, t, \theta, \gamma) \) shows the energy released by this pixel at the direction of \( \theta \) and the scale of \( \gamma \). To make the above obtained energy measure comparable in different facial expressions, we normalize it using:

\[
\hat{E}(x, y, t, \theta, \gamma) = \frac{E(x, y, t, \theta, \gamma)}{\sum E(x, y, t, \theta_i, \gamma) + \epsilon} \]  

(4)

where \( \theta_i \) considers all the directions and \( \epsilon \) is a small bias used for preventing numerical instability when the overall estimated energy is too small. Finally, to improve the localization, we weight the above normalized energy using [4]:

\[
\hat{E}(x, y, t, \theta, \gamma) = \hat{E}(x, y, t, \theta, \gamma).z(x, y, t, \theta)
\]  

(5)

where
\[
z(x, y, t, \theta) = \begin{cases} 
1 & \sum_{\gamma_i} \hat{E}(x, y, t, \theta, \gamma_i) > Z_\theta \\
0 & \text{Otherwise}
\end{cases}
\] (6)

in which \(Z_\theta\) is a threshold for keeping energies at the direction \(\theta\), as too small energies are likely to be noise. The weighted normalized energy obtained in Eq. 5 assigns a number to each pixel (corresponding to the level of the released energy by that pixel) in each of the four chosen directions of \(\theta = 0, 90, 180,\) and 270. Fig. 5 shows these pixel-based energies for a facial image computed at the four different orientations.

The above obtained pixel-based energies can be converted into a more understandable form, if we study the regional changes/motions of the facial muscles of different parts of the face. Based on our observations, different regions of the face contribute differently to the level and direction of the energy in different facial statuses. We have observed that facial muscles that are actively participating to the facial motions during the pain are coming from the three regions that are highlighted in Fig. 6. Besides this, the facial muscles on the left side and the right side of each of these three regions are participating differently in motions during the pain. Because of this, inside each region we have used different colors to distinguish between the left and the right sides.

The above obtained energies can inform us only about some muscles activities (motions), but we need a specific interpretation to see if these motions are due to the pain or not. To do that, we need to study the effect of the pain on the motions of the muscles in the temporal domain. To consider the time domain, we simply obtain the histograms of the directions for each facial region in the aligned input video (Fig. 8(left)). However, as mentioned, since the muscles will move back to their original locations at the end of the pain, instead of the measured directions, we simply consider the changes of the released energies of the muscles in two main orientations: up-down (UD) and left-right (LR). For UD we use \(UD_{R_i} = R_{R_i}(t, 0, \gamma) - R_{R_i}(t, 180, \gamma)\), and for \(LR_{R_i} = R_{R_i}(t, 90, \gamma) - R_{R_i}(t, 270, \gamma)\). These will convert the histograms of directions in (Fig. 8(left)) into changes in histograms of orientations, as shown in Fig. 8(right).

\[
A_{R_i UD} = \sum_{t=1}^{n} UD_{R_i} 
\] (8)

and

\[
A_{R_i LR} = \sum_{t=1}^{n} LR_{R_i} 
\] (9)

where \(A_{R_i UD}\) and \(A_{R_i LR}\) are the integrals of UD and LR for the \(i\)th region \((i = 1, 2, 3)\), respectively, and \(n\) is the number of the frames in the aligned video. Finally, the pain intensity, \(PI\), is obtained by calculating the above two integrals for each of the three regions.
\[ PI = \sum_{i=1}^{3} w_{R_{UD}} A_{R_{UD}} + \sum_{i=1}^{3} w_{R_{LR}} A_{R_{LR}} \]
obtained weights of the corresponding regional histograms of orientations. The $PI$ gives us an indication of the presence of the pain in each frame of the video. Depending on the value of $PI$ we find some experimentally achievable thresholds to classify the pain into three class of no-pain, weak, and strong. The experimental results are given in the next section.

3. Experimental Results

The proposed system has been implemented in Matlab 2014b. We have used the publicly available UNBC-MacMaster Shoulder Pain Expression Archive Database [18] for evaluating the proposed system. This database is composed of 25 participants who suffer from pain in their shoulders. They have been filmed during series of movements in two different scenarios (active and passive). In the active scenario participants move their arms themselves, but in the passive scenario a physiotherapist is responsible for this. Videos were captured at a resolution of $320 \times 240$. The total number of the recorded frames is 48398. In this database, the ground truth pain information has been provided using AUs (see Fig. 9), by:

$$\text{Pain} = AU4 + \max (AU6, AU7) + \max (AU9, AU10) + AU43$$

Figure 9. Active AUs of Pain, the image is from the UNBC database of [18].

For each frame in the database, the AU intensities were coded on a 6 level scale except the AU number 43 which was coded on two levels [6].

To evaluate our system we selected randomly 50 sequences from 12 participants, containing 4926 frames. Table 1 shows the results of the proposed system against the results of the system developed in [10]. The last columns of the table show the percent by which the system have been able to recognize the pain in that specific level. It can be seen from Fig. 10 and this table, that our system not only detects the pain but also recognizes three different levels of the pain. These three levels are no-pain where $PI \leq 0$, weak pain where $1 \leq PI \leq 2$, and strong pain where $PI \geq 3$. The proposed system actually outperforms the system of [10] in terms of the accuracy of recognizing the level of the pain. It should be mentioned that system of [10] is the only energy based system in the literature for calculating the pain, but it is working in the spatial domain. Outperforming this system by our proposed system means that including the temporal information in an energy-based pain recognition system results in better outcomes.

We should take into account that the pain intensity in [10] has been classified into four levels. Two pain levels of Trace and Weak in [10] correspond to the level Weak in our system. Therefore, for comparison purposes mean of the Trace and Weak levels has been reported as Weak in Table 1.

Finally, Fig. 10 shows the $PI$ values obtained by the proposed system (using Eq. 10) for two different video sequences each containing 100 facial images, along with their ground truth data taken from the database employed. It can be seen from this figure, that there is a good overlap between the peaks of the estimated pain intensity curve and the ground truth. It should be noted that the negative values in the estimated values will be considered as zero (no-pain).

4. Conclusion

The proposed system in this paper uses separable steerable filters for automatic detection and recognition of pain. To do that, it applies these filters in both spatial ($x$, and $y$ axises) and temporal (time axis) domains and measures the energies released by the facial muscles that are active during the pain process. The proposed system has produced promising experimental results on the publicly available dataset of UNBC [18]. The results can be improved more by employing a better warping algorithm to compensate for the variations of head pose, for which we plan to use 3D information of facial landmarks in our future works.

References

<table>
<thead>
<tr>
<th>Semantic Ground Truth</th>
<th>Pain Index Ground Truth</th>
<th>Number of Frames</th>
<th>System of [10] (in %)</th>
<th>Proposed System (in %)</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Pain</td>
<td>0</td>
<td>4230</td>
<td>65</td>
<td>77</td>
</tr>
<tr>
<td>Weak</td>
<td>1, 2</td>
<td>387</td>
<td>36</td>
<td>62</td>
</tr>
<tr>
<td>Strong</td>
<td>⩾3</td>
<td>309</td>
<td>70</td>
<td>70</td>
</tr>
</tbody>
</table>

Table 1. Comparing the results of the proposed system (PS) against the system of [10] for different levels of pain from the images of the UNBC database [18].

Figure 10. Comparing the pain levels obtained by the proposed system (top row) against the Ground truth (bottom row), for two different pain scenarios.


