Spatiotemporal Facial Super-Pixels for Pain Detection

Lundtoft, Dennis Holm; Nasrollahi, Kamal; Moeslund, Thomas B.; Guerrero, Sergio Escalera

Published in:
IX Conference on Articulated Motion and Deformable Objects

DOI (link to publication from Publisher):
10.1007/978-3-319-41778-3_4

Publication date:
2016

Spatiotemporal Facial Super-Pixels for Pain Detection

Dennis H. Lundtoft, Kamal Nasrollahi, Thomas B. Moeslund, and Sergio Escalera

Abstract Pain detection using facial images is of critical importance in many Health applications. Since pain is a spatiotemporal process, recent works on this topic employ facial spatiotemporal features to detect pain. These systems extract such features from the entire area of the face. In this paper, we show that by employing super-pixels we can divide the face into three regions, in a way that only one of these regions (about one third of the face) contributes to the pain estimation and the other two regions can be discarded. The experimental results on the UNBC-McMaster database show that the proposed system using this single region outperforms state-of-the-art systems in detecting no-pain scenarios, while it reaches comparable results in detecting weak and severe pain scenarios.

Keywords: facial images, super-pixels, spatiotemporal filters, pain detection

1 Introduction

Remote monitoring of e.g. chronically ill patients is an increasing courtesy of physicians, as it improves quality of life of patients rather than staying at a hospital. However, remote extraction of data can be limited, as several measurements rely on direct contact, one such example is pain measurement using facial images.

Pain, which is a sensation of the body expressing itself to be damaged or in danger, is rather important for doctors to monitor. In long-durations it can heavily impact the quality of the life. A popular technique for measuring pain is patient self-report, however, for babies and for people in some illnesses, such as dementia, there are cases where the patient is unable to express their pain. To deal with such scenarios, automatic detection of pain using imaging techniques is of growing interest. The focus of this paper is therefore to develop a pain detection system using deformable facial images.

The rest of this paper is organized as follows: the related work in the literature on pain detection is reviewed in the next section in which the contributions of this work are also highlighted. Then, in section 3 the proposed system is explained. The obtained experimental results are reported then in section 4. Finally, the paper is concluded in section 5.

2 Related work

The current systems on pain detection can be divided into two groups: the first group only decides if a given image is of a painful case or not, while the second group besides clarifying the pain presence determines its level as well.

In the first group, the work of [3] and [14] use a Support Vector Machine (SVM) classifier that works with eigenfaces to see if there is any sign of pain in a given face of an infant or not. The work of [3] was then extended by [6], using a Relevance Vector Machine (RVM) instead of SVM as it introduces a degree of uncertainty to the estimated pain depending on the posterior probability score. The work of [11] utilized automatized facial expression analysis, using Gabor filters on eight different orientations and nine different spatial frequencies, to find facial Action Units (AU) to distinguish between faked and genuine pain. Then, these AUs were used to see if the pain was genuine of faked. AUs have also been used in [4] for defining a rule-based system for detecting the pain/no-pain case. Active Appearance Models (AAM) have been used in [2] to decouple shape and appearance parameters from digitized facial images, while in [15] Multiple Instance Learning (MIL) has been used to handle training data by putting it into bags, which are labeled as either positive, if the bag contains a positive instance, or negative, if no positive instances exist in the bag. Then, a Bag of Words (BoW) approach is used for determining whether a set of frames contains pain or not.

In the second group, where the focus is on determining the level of the pain, the work of [13], which is an extension of [2], uses facial expression analysis and 3D head pose to find the level of the pain. In [10], three feature sets of Facial landmarks (PTS), Discrete Cosine Transform coefficients (DCT) and Local Binary Patterns (LBP) are extracted from the facial images, and are then fed to a Relevance Vector Regression (RVR) to estimate the pain intensity. In [7] canonical appearance of the face using AAM are passed through a set of log-normal filters to get a discriminative energy-based representation of the facial expression which is then used to estimate the pain level. Inspired by this work, in [8] another energy-based system has been developed for pain estimation which uses spatiotemporal filters.

The proposed system in this paper is inspired by and based on the work of [8]. The current work treats pain as a spatiotemporal process and hence uses steerable filters to extract energy released from the face during the pain process. The main contributions of this paper can be summarized as below:

- As opposed to the work of [8] which uses facial landmarks to divide the face into three regions, we define our facial regions using super-pixels.
- It is shown in the experimental results that such a division of the face, results in three regions, similar to [8]. However, from these three regions, only one of them contributes properly to the pain estimation, as opposed to [8] which uses all the three regions. The proposed system therefore uses only this single facial region and the other two regions are discarded.
- Though the proposed system's performance in determining the level of the weak and severe pain is on average about 6% lower than the performance of

[8], its performance in determining the no-pain scenarios is about 15% better than the performance of [8]. These are further explained in the experimental results sections.

3 The Proposed System

The block diagram of the proposed system is shown in Figure 1. Having an input video sequence, for each frame, first, the face region will be detected and segmented from the background. Then, using Procrustes analysis, warping, and image registration the facial images found in the video sequence are aligned. Then, super-pixels are formed for each face image. These super-pixels are used to define the region of interest from which spatiotemporal released energies are found and used for detecting pain. These steps are explained in the following sub-sections.

Figure 1. The block diagram of the proposed system

3.1 Face Detection

Since the database employed in the experimental results, UNBC-MacMaster [12], already provides the positions of 66 facial landmarks in each frame, these landmark positions are used to detect the face. To do so, a Delauney triangulation is applied to the positions of the landmarks 2(a), which spans a facial mask as seen in Figure 2(b). This mask is used to segment the face from the rest of the image 2(c).

3.2 Facial Image Alignment

The segmented facial regions from the previous step, need to be aligned as they might have been displaced due to other sources of motion that are not directly related to pain, such as eye blinking and speaking motions. To do so, following [8], we utilize Procrustes analysis on the facial landmarks, followed by a piece-wise affine warping, and an inpainting step.

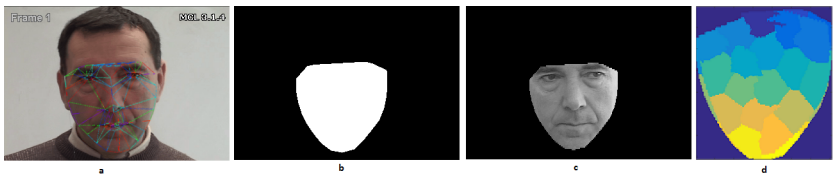

Figure 2. Segmenting the face area and obtaining super-pixels: a) Delauney triangulation of facial landmarks, b) the spanned mask by the triangulation, c) the segmented face, d) and its SLIC super-pixel labels.

3.3 Super-Pixel Regions

Having aligned the facial images, for each face, we form a set of super-pixels. The super-pixels are determined using the spatial proximity and the color similarity between pixels. We use SLIC super-pixel algorithm of [1] for this purpose. SLIC super-pixels are clustered in a five-dimensional $[labxy]$ space, where $[lab]$ is a color vector in CIELAB color space and $[xy]$ is the pixel position. The SLIC algorithm uses two parameters: the first, the desired number of approximately equally sized super-pixels in the image K, meaning the approximate size of each super-pixels in an image with N pixels is $\frac{N}{K}$ pixels, the second, cluster centers at every grid interval $S = \sqrt{\frac{N}{K}}$. The onset super-pixel centers $C_k = [l_k a_k b_k x_k y_x]^T$ is initially set at the regular grid intervals S. Since the spatial extent of a super-pixel is S^2, it is assumed that pixels associated with a cluster center lie within a $2S \times 2S$ area of the center on the xy plane, which is the spatial search area of pixels associated to the cluster center. We use the same distance measure of [1] to calculate the distance between the pixels and centers of the clusters.

In order to locating center of a super-pixel on the edge of the super-pixel, and to reduce the chance of choosing a noisy pixel as the center, a gradient descent algorithm in a 3×3 neighborhood is performed from the initial position of the centers. The gradient of the image is calculated using both the color and intensity information as:

$$G(x, y) = \|I(x+1, y) - I(x-1, y)\|^2 + \|I(x, y+1) - I(x, y-1)\|^2 \qquad (1)$$

where $I(x, y)$ is the lab vector at pixel position (x, y) and $\|, \|$ is the L_2 norm. The search radius around a cluster center is $2S \times 2S$. Pixels are initially associated with the nearest cluster center, and then a new cluster center is calculated using the average $labxy$ vector of all pixels belonging to the cluster. This is performed iteratively until a convergence is met. Lastly, connectivity is ensured by relabeling unconnected labels to largest neighboring cluster. The resulted super-pixels by the SLIC algorithm applied a detected face from 2(c) can be seen in Figure 2(d).

Having found the super-pixels, the next step is to use them to form some facial regions and use only those super-pixels/regions that are contributing to the pain detection and estimation. The state-of-the-art works of [8] and [9] use

facial landmarks positions to divide facial area into three regions: region 1) eyes and eyebrows, region 2) nose and the cheeks, and region 3) mouth and lower part of the face (Figure 3). The super-pixels are therefore grouped in a way that three such regions are formed.

Figure 3. The three different regions formed on the facial image, Left: Region1 consisting of eyes and upper face. Middle: Region2 consisting of nasal area and cheeks. Right: Region3 consisting of mouth and lower face.

Though three regions have been formed, as discussed in Section 4, when looking at the calculated spatiotemporal energy released from the different regions (discussed in the following subsection) it could be seen that region 2 was by far the most dominant and stable region when it came to pain detection, and that the other two regions often contributed a large amount of noise compared to the pain responses coming from them, hence it was decided to only use the second region, i.e., the nasal and cheek region, as a singular region of interest.

3.4 Spatiotemporal Features

Having detected the facial region of interest, we need to extract the features by which we detect the presence or absence of the pain and estimate its intensity. For this purpose, following [8,**?**] we use spatiotemporal energy released by the pixels. These are extracted by steerable filters. A steerable filter is an orientation-selective convolution kernel, which can be expressed by a linear combination of a set of rotated versions of itself. Such an oriented filter can be synthesized at any given angle, which is called steering. The steerable and separable filters are separated into basis filter banks, splitting them up into several sub-filters of lower complexity (i.e. separable). Once they are split up, the filters are multiplied by a set of gain maps, which adaptively control the orientation of the filters (i.e. steerable).

These filters have been proposed for extraction of spatiotemporal data in [5] in which the second derivative Gaussian filter G_2 and the Hilbert transform H_2 of the second derivative Gaussian are used. It is applied to a sequence of 2D images, utilizing the spatial domain $xandy$ as well as the temporal domain t. The formulas for a two dimensional Gaussian $G(x, y)$ and its second derivative with regard to x is as:

$$G(x,y) = e^{-(x^2+y^2)} G_2(x,y) = \frac{\partial^2 G}{\partial x^2} = (4x^2 - 2)e^{-(x^2+y^2)} \tag{2}$$

The Hilbert transform of the second derivative Gaussian is defined as:

$$H_2(x, y) = (-2.254x + x^3)e^{-(x^2+y^2)} \tag{3}$$

The second derivative Gaussian and Hilbert transform functions are then separated into basis functions, splitting the complexity of the functions up into fewer dimensions. For the second derivative Gaussian function six basis functions are needed for its separable set, as it has a 2nd order polynomial. The Hilbert transform requires 10 basis functions, as its polynomial is of 3rd order. The amount of basis functions required in the basis set is $M \geq \frac{(N+1)(N+2)}{2}$ where N is the order of the polynomial.

Next step is then to filter the image sequence $I(x, y, t)$ by G_2 and H_2 at the orientations $(\alpha, \beta, \gamma)_i$, which are found using the spherical coordinate orientation $(\theta, \phi = \frac{\pi}{2}, \rho = 1)$:

$$\alpha = cos(\theta)sin(\phi), \beta = sin(\theta)sin(\phi), \gamma = cos(\phi) \tag{4}$$

In this work θ takes the value of the four main directions $\theta = [0, 90, 180, 270]$. Filtering the image sequences at these orientations with the G_2 and H_2 filters provides a local energy measure as:

$$E_(x, y, t, \theta) = [G_2(\theta) * I(x, y, t)]^2 + [H_2(\theta) * I(x, y, t)]^2 \tag{5}$$

which is normalized by the sum of the consort response, by:

$$\hat{E}(x, y, t, \theta) = \frac{E(x, y, t, \theta)}{\sum_j E(x, y, t, \theta_j) + \epsilon} \tag{6}$$

where θ_j is all directions and ϵ is a bias constant to prevent numerical instability at small energy levels. Finally, the measured energy is filtered from too small values to remove likely noise, by:

$$\dot{E}(x, y, t, \theta) = \hat{E}(x, y, t, \theta) \cdot z(x, y, t, \theta) \tag{7}$$

where Z_θ is a constant to threshold low energy values and equals to one if $\hat{E}(x, y, t, \theta) > Z_\theta$, otherwise it is zero.

The measured energy is the spatiotemporal features of the images, which can be used to determine the pain index. The calculated pixel-based energy is then collected in oriented histograms over the region of interest, using:

$$H(t, \theta_j) = \sum E(x, y, t, \theta_j) \tag{8}$$

where H is the histogram of each respective direction θ_j which accumulates all the released energy.

Since muscles always move back to their resting position after exertion, the complimentary orientation histograms are combined. This means that histograms represent vertical or horizontal energy instead of a direction, by merging the histograms from complementary orientations together. We have observed that the the vertical motion is more active during pain expression while the horizontal

motion of muscles were fairly docile. This indicates that horizontal motion is weak at determining pain, thus this work will only utilizes the histogram representing vertical motions.

3.5 Pain Intensity Estimation

Since in this work we only utilize one facial region (compared to [8] which uses all the three facial regions), and since we consider only vertical muscle (compared to [8] which considers both vertical and horizontal motion), the pain index is calculated as:

$$PI = \sum_{t=1}^{n} UD_t \tag{9}$$

where n is the number of frames and UD_t is the histogram of the vertical muscle motion. Several post-processing steps are then used on the pain index, this includes smoothing it using a moving average filter and normalizing it to the ground truth. Furthermore, the estimated pain index using Eq.9 often has issues with negative values before and after a pain episode, resulting in lower values overall in the pain episode as it often starts from a negative value. In order to compensate for this discrepancy, we simply "lift" the pain episode by the most negative number before the pain episode, ensuring that it starts at 0 when the pain index starts ascending.

4 Experimental Results

In this section we first give the details of the employed database, then, we show why keeping only region 2 of Figure 3 is enough for detecting the pain. Finally, we represent the obtained results and compare them against state-of-the-art similar systems.

4.1 The Employed Database

The proposed system, which has been implemented in MATLAB 2014b, has been tested on the UNBC-MacMaster shoulder pain database [12]. This public benchmark database consists of 25 different subjects, with varying gender and age, having shoulder pain performing both active and passive movements while being filmed. It consists of in total 200 video sequences, each with ground truth pain intensity values and positions of facial landmarks for each frame of the video sequences. From the 200 video sequences 79 consisted of sequences containing pain according to the ground truth. Therefore, only these 79 pain sequences were used when testing the system. The ground truth pain intensity values were calculated using the Facial Action Coding System (FACS) metric, which considers the severity of movement of key facial action units, from which a pain intensity is calculated on a scale of 0-16.

4.2 Why Only Region 2?

When looking at the calculated spatiotemporal energy released from the different regions, it could be seen that region 2 is by far the most dominant and stable region when it comes to pain detection, and that the other two regions often contributed a large amount of noise compared to the pain responses coming from them. Figure 4 shows an example where the first and third regions contained large amounts of noise along with the relevant frames, while the second region had less noise and mainly contained energy at the relevant frames indicated by the red lines. First frame is an example of where a frame results in noise due to non-pain related eye movement. Second frame is a "neutral" face and last showcased frame displays the subject's facial expression during pain. The drawback of only using region 2 is that people are vastly different, which also meant that the pain expression vary throughout the subjects. While some subjects mainly used cheeks/nasal area others expressed their pain by tightly closing their eyes or widely opening their mouth. This work chose to focus more on being autonomous, thus not requiring manual interaction to determine which region should be used which meant only the most stable region of interest is used, while this has been done manually in [8]. Hence, it was decided to only use the second region, i.e., the nasal and chin region, as a singular region of interest.

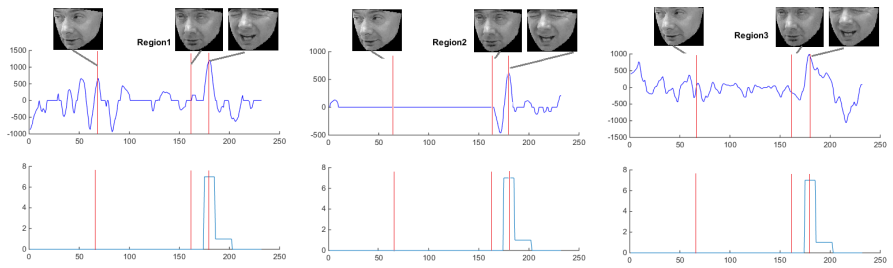

Figure 4. The results of the proposed system from the three different facial regions (top) against the ground truth (bottom) for a pain sequence. The x and y axises of the plots show the time and the pain index, respectively.

4.3 Results

Figure 5 shows the pain indexes obtained by the proposed system against the ground truth provided in the database for a simple (right) and a challenging (left) pain sequence. It can be seen from these figures that there are generally good agreement between the results of the proposed system and the ground truth.

The results of the proposed system are compared against two state-of-the-art pain detection systems of [7] and [8]. Following these two works, the obtained pain index of section 3.5, is classified into three different categories of no pain

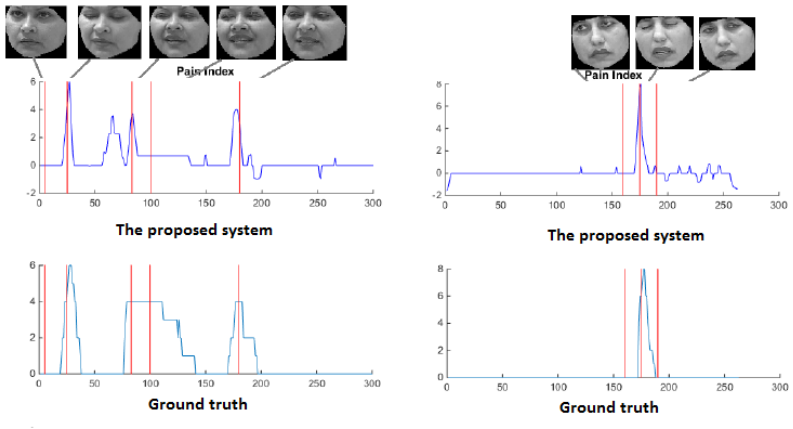

Figure 5. The results of the proposed system against the ground truth for a simple (right) and a challenging (left) pain sequence.

(if the pain index is zero), weak pain (if the pain index is either 1 or 2), severe pain (if the pain index is larger than or equal three). The results using the UNBC-McMaster database is shown in Table 1. It can be seen from this table, that our system is not as good as [8] (on average about 6% lower) in detecting the weak and severe pain, but it is much better this system in detecting no-pain cases (about 15% better on average).

System	No pain [%]	Weak pain [%]	Severe Pain [%]
The proposed system	**91.70**	55.75	63.4
[7]	65	36	70
[8]	77	**62**	**70**

Table 1. The results of the proposed system against two systems of [7] and [8] using the UNBC-McMaster database.

5 Conclusion

This paper proposed a spatiotemporal approach for detecting pain from facial images using steerable filters. To discard parts of the face which contribute negatively to the pain estimation process, we divided the face into three regions using super-pixels. Then, only of the region that contributes properly to the pain estimation has been kept and used. The experimental results on public benchmark database of UNBC-McMaster show that the proposed system outperforms state-of-the-art similar systems in detecting no-pain scenarios, while it produces comparable results in detecting weak and severe pains.

References

1. Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Susstrunk, S.: Slic superpixels compared to state-of-the-art superpixel methods. Pattern Analysis and Machine Intelligence, IEEE Transactions on 34(11), 2274–2282 (2012)
2. Ashraf, A.B., Lucey, S., Cohn, J.F., Chen, T., Ambadar, Z., Prkachin, K.M., Solomon, P.E.: The painful face âĂŞ pain expression recognition using active appearance models. Image and Vision Computing 27(12), 1788 – 1796 (2009), visual and multimodal analysis of human spontaneous behaviour:
3. Brahnam, S., Chuang, C.F., Shih, F.Y., Slack, M.R.: Machine recognition and representation of neonatal facial displays of acute pain. Artificial Intelligence in Medicine 36(3), 211–222 (2006)
4. Chen, Z., Ansari, R., Wilkie, D.J.: Automated detection of pain from facial expressions: a rule-based approach using aam. In: SPIE Medical Imaging. pp. 83143O–83143O. International Society for Optics and Photonics (2012)
5. Derpanis, K., Gryn, J.: Three-dimensional nth derivative of gaussian separable steerable filters. In: Image Processing, 2005. ICIP 2005. IEEE International Conference on. vol. 3, pp. III–553–6 (Sept 2005)
6. Gholami, B., Haddad, W.M., Tannenbaum, A.R.: Agitation and pain assessment using digital imaging. In: Engineering in Medicine and Biology Society, 2009. EMBC 2009. Annual International Conference of the IEEE. pp. 2176–2179. IEEE (2009)
7. Hammal, Z., Cohn, J.F.: Automatic detection of pain intensity. In: Proceedings of the 14th ACM international conference on Multimodal interaction. pp. 47–52. ACM (2012)
8. Irani, R., Nasrollahi, K., Moeslund, T.B.: Pain recognition using spatiotemporal oriented energy of facial muscles. In: Computer Vision and Pattern Recognition Workshop, 2015 IEEE Conference on. pp. 679–692 (2015)
9. Irani, R., Nasrollahi, K., Simon, M.O., Corneanu, C.A., Escalera, S., Bahnsen, C., Lundtoft, D.H., Moeslund, T.B., Pedersen, T.L., Klitgaard, M.L., et al.: Spatiotemporal analysis of rgb-dt facial images for multimodal pain level recognition (2015)
10. Kaltwang, S., Rudovic, O., Pantic, M.: Continuous pain intensity estimation from facial expressions. In: Advances in Visual Computing, pp. 368–377. Springer (2012)
11. Littlewort, G.C., Bartlett, M.S., Lee, K.: Automatic coding of facial expressions displayed during posed and genuine pain. Image and Vision Computing 27(12), 1797–1803 (2009)
12. Lucey, P., Cohn, J.F., Prkachin, K.M., Solomon, P.E., Chew, S., Matthews, I.: The UNBC-McMaster Shoulder Pain Expression Archive Database (2011), link to UNBC-MacMaster Shoulder Pain Database
13. Lucey, P., Cohn, J.F., Prkachin, K.M., Solomon, P.E., Chew, S., Matthews, I.: Painful monitoring: Automatic pain monitoring using the unbc-mcmaster shoulder pain expression archive database. Image and Vision Computing 30(3), 197–205 (2012)
14. Monwar, M., Rezaei, S.: Appearance-based pain recognition from video sequences. In: Neural Networks, 2006. IJCNN '06. International Joint Conference on. pp. 2429–2434 (2006)
15. Sikka, K., Dhall, A., Bartlett, M.: Weakly supervised pain localization using multiple instance learning. In: Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on. pp. 1–8. IEEE (2013)