Published in final edited form as:

IEEE Int Conf Autom Face Gesture Recognit Workshops. 2015 May ; 2015: . doi:10.1109/fg.2015.7163107.

Three Dimensional Binary Edge Feature Representation for Pain Expression Analysis

Xing Zhang[1], **Lijun Yin**[1], **Jeffrey F. Cohn**[2]

[1]Binghamton University-SUNY

[2]University of Pittsburgh

Abstract

Automatic pain expression recognition is a challenging task for pain assessment and diagnosis. Conventional 2D-based approaches to automatic pain detection lack robustness to the moderate to large head pose variation and changes in illumination that are common in real-world settings and with few exceptions omit potentially informative temporal information. In this paper, we propose an innovative 3D binary edge feature (3D-BE) to represent high-resolution 3D dynamic facial expression. To exploit temporal information, we apply a latent-dynamic conditional random field approach with the 3D-BE. The resulting pain expression detection system proves that 3D-BE represents the pain facial features well, and illustrates the potential of noncontact pain detection from 3D facial expression data.

Keywords

Pain; Facial Expression; Latent-Dynamic Conditional Random Field (LDCRF); Emotion

I. Introduction

Traditionally, pain assessment relies on subjective rating scales. Because subjective ratings are prone to error or may not even be possible for some populations or applications (e.g., intubated patients), objective assessment is desirable. Within past decade, significant progress has been made in automatic pain detection from video recordings of the face.

Reliable indicators of pain in facial expression have been identified [6]. Several machine-learning based automatic pain detection methods have been developed. Littlewort et al. [8] reported an optimal facial action unit (AU) recognizer with Gabor wavelet features and classified by support vector machine (SVM) in 2D video. They used it to identify distributions of 20 AUs to differentiate the genuine and fake pain expressions. Lucey and colleagues [7][18] applied active appearance models (AAM) to automatically identify pain-related facial actions and pain in 2D video.

Previous work in automatic pain detection has several limitations. It is limited to 2D registration, which is unable to represent 3D shape or motion and is less robust to the moderate to large head rotation that may be common in clinical settings. 3D depth may be important to recognition of facial actions such as lip pucker and funneling that have pronounced changes in 3D position. Previous work also has neglected the timing of facial

actions. Although it has been shown that temporal information is important to facial expression understanding [4][5], no research has been reported to date of utilizing the temporal information for pain expression detection with 3D data. Moreover, the AU set to differentiate pain expression from the other expressions needs to be further verified and validated.

Dynamic 3D spontaneous facial expression representation with advanced 3D imaging systems opens up a great alternative to address the aforementioned limitations [12][13]. Essentially, face shape and actions are represented as deformation of 3D surface silhouette. The surface can be represented by a geometric mesh with resolution as high as 50,000 ~ 100,000 vertices. High-resolution 3D mesh representation allows for pain expression analysis in a high level of details. At same time, it poses a new challenge to efficiently process and represent this level of detail. To meet this challenge, we propose a three-dimensional binary edge (3D-BE) approach to extract and represent facial features in a 3D space. This method automatically detects patterns of pain in 3D dynamic sequences. The approach was developed and evaluated in the newly released Binghamton-Pittsburgh 3D Dynamic (4D) Spontaneous Facial Expression Database (BP4D-Spontaneous) [11]. The contributions of our work are listed as follows:

- A novel 3D edge feature descriptor of geometric surface has been proposed for facial expression representation;
- An optimal set of Action Units is explored among pain expression and the other non-pain expressions;
- A novel system with a unique combination of 3D edge features, a temporal model, and the optimal AUs has been developed for recognition of authentic pain expressions and their intensity. Such a system utilizes 3D dynamic sequences without using any texture information, which alleviates the influence of variable imaging conditions.

In Section 2, we describe the new approach of extracting 3D binary edge features from the geometric surface, followed by an evaluation of such a new 3D edge feature descriptor through comparing it to the other state of the art 3D/2D feature descriptors in Section 3. In Section 4, the procedure of AUs selection for pain intensity detection is illustrated. In section 5, we compare feature representation and learning methods between our design and previous designs to show the importance of temporal information. Then how we use the system to evaluate pain intensity is illustrated. Finally, a concluding remark and discussion are given in Section 6.

II. 3D Binary Edge Feature Descriptor

A 3D facial model represents the 3D topographic surface of a face. The geometric edges of the 3D surface can be used as unique features to represent facial expression since the facial muscle deformation drives the change of the geometric edges. The binary edge feature extracts facial components with large normal changes on the surface meanwhile it filters out smooth and non-informative regions. As will be shown, the binary edge contains important

expression information while keeps the low dimensionality. To construct the new feature descriptor, the following three steps are needed: (1) geometric edge feature extraction; (2) face normalization; and (3) descriptor composition.

A. Feature Extraction

Motivated by the conventional sketch drawing approaches for non-photorealistic rendering [1], we propose to detect geometric edges using the adjacent surface relationship and a 3D-to-2D mapping approach (as shown in Fig. 1). When an artist draws a sketch of a person, he/she draws lines specifically to represent the facial shape. Silhouettes (or contours) from a 3D surface depict the important geometric features explicitly. The 3D model can also be rendered (or shaded) with different hatching strokes to simulate non-photorealistic effects such as pencil drawing or crayon drawing [1]. The existing technique has been applied usually to the artificial models with “clean” surfaces and “clear” contours. Little work has been done on real-world 3D objects with noise, nor has any study on classification using such 3D edges been done.

In order to extract the representative edges from the 3D geometry, we first render the 3D model using their surface normal vectors to generate the color intensities in a projective space (Fig. 1). As a result, every vertex can be accessed easily through 2D array indexing in the rendered projective space.

Each pixel in the rendered projective space is evaluated according to the edge sharpness. An edge is counted if the two shared planes satisfy the condition:

$$N_1 \bullet N_2 = |N_1||N_2|\cos\alpha < 0 \tag{1}$$

where N_1 and N_2 are the normal vectors of the two triangles in the 3D viewing space, respectively; α is the angle between the two normal vectors. The larger the angle is, the sharper the shared edge exhibits. In order to make the 3D edge detection resilient to noise as well as scalable to various levels of detail, we further take the neighboring vertices into account. As shown in Eq. (2), the sum of the dot products of a vertex normal (N_0) and its four neighbors’ normal (N_i) is regulated to a positive range. The larger the sum is, the sharper the corresponding edge displays. The threshold t determines the amount of edges to be detected. Fig. 2.a illustrates the influence of threshold t on the edge extraction. In the first row, the third image reduces the noise as compared to the second one; however, the line of lip also partially disappears. A small threshold t generates binary edges with expected features yet tangled by a significant amount of geometric noise. If the threshold t is higher, the noise can be reduced with the risk of features loss.

$$\sum_{i=1}^{4}(1 - N_o \bullet N_i) \geq t \tag{2}$$

In order to further reduce the noise and enhance the feature, we apply a smoothing operation prior to the edge detection. Blurring the surface normal can reduce the dramatic change of

the normal vector due to the noise. Meanwhile, it also extends the influence of sharp edges to their neighbors, hence resulting in thicker edges. Fig. 2.b illustrates the effect of the smoothing procedure.

We choose an optimal scale t (=0.08) according to the trade-off between noise and features. We apply the 3D edge detection algorithm to 41 subjects' models of the BP4D-Spontaneous database. As an example shown in Fig. 2, the amount of 3D edges as well as noises is degenerated from a fine level to a coarse level when the threshold t increases. The percentage of this change (Cp) is computed by the ratio of the difference versus the total amount of pixels in the image. Further, the rate of the change (i.e., ratio of Cp versus threshold increase δt) accounts for the speed of change for the detected details. The rate of change can also indicate how fast the noise decreases along with the unit δt. Fig. 3 plots the speed of change with increasing t value. It is clear that the noise decreases fast when t is small, while the speed is slower when t is large. This is reasonable since no-edge vertices are ought to be excluded quickly using our algorithm and the useful edge information can be preserved efficiently. In our case, we choose the first t value ($=80\times10^{-3}$) that results in speed slower than 0.5% to maintain the balance between noise and features. In terms of different dataset, the corresponding t value can be found based the same speed criterion.

B. Face Normalization

A general automatic 3D alignment and cropping pipeline is employed to keep the correspondence across 3D models and normalize the size of facial features.

Fig. 4 illustrates the procedure of feature registration and face normalization for one frame of 3D face model. We detected 83 3D feature points and pose information of each model automatically, and align each frame to a generic 3D face model with three reliable feature points, which are invariant to facial expression change and robust to possible occlusion during spontaneous facial activities. First, based on the middle point of the inner eye corners and pose information, the target frame is translated to the origin and rotated to the front view. Second, the target frame is scaled to the generic model through the horizontal and vertical normalization according to the inner eye corner and philtrum. Fig. 4.a and 4.b illustrate the results before and after the alignment. Fig. 4.c shows the result after the model is cropped based on the 3D face boundary defined by the generic model. A 2D mask from the frontal view of the generic model is used to clean up the fringe of the 3D rendered image to maintain a constant number of informative image pixels. Then we use a bounding box to extract the cropped face region (as shown in Fig. 4.d). The data inside the region is scaled into a 144 by 144 matrix (Fig. 4.e).

C. 3D Edge Feature Descriptor

The binary edge reveals the important information on location and scale of the facial features. As shown in Eq. (3), we subdivide the matrix into n grids. For each grid, the number of edge pixels c_i is counted. All these numbers are divided by the total edge pixel number of this frame to obtain the percentage of information p_i each grid holds. The grid

represents the information of location while the percentage represents the size of the local edge regions.

$$V = [p_1, p_2, \ldots p_n], p_i = c_i \Big/ \sum_{i=1}^{n} c_i \quad (3)$$

III. Feature Descriptor evaluation

To show the merit of our features, we evaluate the 3D edge feature descriptor by comparing it to the other three state-of-the-art 3D feature descriptors (e.g., shape index [13], primitive nebula feature [15] and LBP-TOP depth [16]), and the two conventional 2D feature descriptors (e.g., LBP-TOP texture [16] and Gabor wavelet [8]).

To extract 3D binary edge features sequentially, the method proposed in Section 2 is implemented in parallel on the GPU. The 144×144 matrix feature is divided into grids and produces vector with a size of (36 × 36 =) 1,296.

Shape index [13] describes the shape of the local surface, where each vertex belongs to, by the curvature. During expression change, the face deformation leads to the variation of curvature distribution. We computed the shape index of each vertex on GPU and exported the rendered result into 144 ×144 matrix feature. The quality of the mesh is important to this feature representation.

Nebula feature [15] extends the primitive topographical features [17] with temporal information by dividing and fitting multiple frame data to the cubic polynomial, thus different spatio-temporal deformation can be observed. It uses 15 curvature categories and the major axis angles as features.

LBP-TOP is a temporal extension of local binary pattern features [16] by considering LBP on three orthogonal planes: XY, XT and YT. In each plane, LBP feature and the histogram can be computed and concatenated to describe the dynamic feature. Appling LBP-TOP algorithm on the depth/texture rendered face to derive the LBP-TOP depth/texture feature.

In addition, a 2D video based approach using Gabor wavelet features [8] is also selected for comparison. We apply the Gabor filter to extract features from the 2D texture video sequences, with three spatial scales and eight orientations (2, 8, 32 pixels per cycle at 1/2 octave steps). The cropped face texture is scaled to 36×36 image, thus a 31,104 (=36×36×3×8) dimensional feature is generated.

The optimal dimension of feature vector of 3D-BE, shape index and Gabor wavelet is searched by PCA.

We compared the above peer methods by testing the recognition accuracy on 12 different AUs (listed in Table I) with a binary SVM classifier on 16 subjects from the BP4D-Spontaneous database [11]. Fifty positive frames and fifty negative ones were randomly

selected from each of the eight coded sequence of these subjects. Leave-one-subject-out test was conducted with the selected frames (51,153 samples). Table II shows the result in terms of accuracy.

On average, the 3D-BE feature descriptor achieves the best performance in detecting the 12 AUs among the compared descriptors of both 3D domain and 2D domain. Indeed, the 3D-BE, which contains approximately only 20% pixels of the face region, preserves the important shape information very well. It even outperforms the 3D temporal based Nebular and LBP-TOP Depth features. The 2D result using normalized texture approach (in Table II) is no better than the 3D edge based approach even when the head poses are adjusted to be frontal (which is believed to be an optimal condition for the 2D case), not to mention the performance degradation of the 2D case when the head pose change alters the appearance due to the prospective projection, whereas the 3D case is not affected at all as long as the model contains the completed edge feature. Moreover, the normalized texture approach with Gabor features generates the data with a very high dimensionality, which is a drawback for data training and dimension reduction. Therefore, our approach is advantageous to represent the 3D facial features with low dimensionality while being robust to pose changes.

Also notice as compared to the posed facial expressions, the spontaneous facial expressions are more challenging to recognize due to their variety and complexity in facial appearances in terms of how, when, and what expressions could appear. We will show in the following section (i.e., Section 5) that the performance will be improved in recognizing the target expression when the temporal information is taken into account using the 3D-BE and the latent-dynamic conditional random field (LDCRF).

IV. Target Action Units Selection

In BP4D-Spontaneous database, eight elicitation methods (Task 1 ~ 8) have been applied to elicit eight spontaneous emotions, *i.e.*, *happiness, sadness, surprise/startle, embarrassment, fear/nervous, physical pain, anger/upset, disgust,* respectively [11]. We are focusing on the physical pain, where the pain expression was elicited by submerging the subject's hand in ice water.

Before we apply the 3D binary edge feature for spontaneous pain expression recognition, the question "What kind of expression does represent spontaneous physical pain?" needs to be answered. Previous psychological research demonstrates that a specific group of Action Units carries the bulk of information about pain. Prkachin and Solomon selected a study pool of possible Action Units related to pain: brow-lowering (AU 4), cheek-raising (AU 6), eyelid tightening (AU 7), nose wrinkling (AU 9), upper-lip raising (AU 10), oblique lip raising (AU 12), horizontal lip stretch (AU 20), lips parting (AU 25), jaw dropping (AU 26), mouth stretching (AU 27) and eye closing (AU 43) [6]. Then they collected video data regarding shoulder pain of patients on the affected and unaffected sides. With the most intense facial expression as study sample and anatomic relation of the AUs, the Prkachin and Solomon pain intensity scale (PSPI) is defined as below:

$$\text{Pain} = \text{AU4} + (\text{AU6}\|\text{AU7}) + (\text{AU9}\|\text{AU10}) + \text{AU43} \quad (4)$$

PSPI has been used to label the intensity of pain based on AUs. Patrick et al. used the PSPI to automatically label the shoulder pain intensity of the videos [7]. They also showed interesting results that AU4 and 10 are poor in detecting pain and including AU12 increase the performance. On the other hand, Littlewort et al. investigated the detection of genuine pain (by cold pressor) and posed fake pain [8] using 2D videos and twenty linear SVM classifiers. They find that the genuine pain is highly correlated with AU 6, 9, 10, 12, 25 and 26.

Choosing a proper AU subset is crucial for the reliable pain analysis. We take into account both the AU codes of the database and the existing work [6], to determine an optimal set of AUs for pain expression classification. Based on the work [6] and the coded AUs of pain in the BP4D-Spontaneous, AU4, 6, 7, 9, 10, 12, 20, and 27 are chosen as candidates. The AU histogram (Fig. 5 (top)) shows the candidate AUs with significant amount of frame counts are AU10, 7, 6, 12, 4, and 9. The histogram (Fig. 5 (bottom)) also shows that the most frequently occurred AU combinations from the candidate set are AU6+7+10+12 and AU4+7. Therefore, AU 4, 6, 7, 10, 12, and 9 appear to be the good choice to represent the pain expression.

In order to further investigate whether the selected AU subset is sufficient and accurate for distinguishing pain expression from the other expressions, we conducted paired t-test on the selected AUs and AU-combinations for all the tasks of the database. This study is to assess whether pain and other expressions are different in terms of the AU quantity. Based on the locations of the AUs in a face we also consider combining AU 6 and 7, AU 10 and 12 as two pairs. For each possible AU set, the percentage of frames with this AU set is reported for each task per subject. For AU combinations in the AU set, only the frames showing all elements count. As we can see in the p score listed in Table III, the differences between pain and all other expressions regarding AU quantity are only significant in the case of the AU sets of 6&7, 9 and 10&12 ($p < 0.05$). The traditional AU set for PSPI, which excludes AU43, as shown in the first two rows, does not show preeminent discriminative power.

Given the above study, we select AU6&7, 9 and 10&12 (row 6 of Table III) as the optimal AU set. Binary classifiers are built for each of them and the log-likelihood output is averaged in order to evaluate the pain level in our experiment.

V. Pain Detection Experiment and Evaluation

A. Data

We use the well-annotated BP4D-Spontaneous database [11] in our experiment. The database contains 41 subjects with eight tasks (i.e., emotions) including pain. For each task over 20 seconds, a 12-second video segment is selected for AU coding with binary labels on each frame. Otherwise, the entire sequence is AU coded. We use the data of 41 subjects for

our experiment, including 52,578 frames of 3D models. Fig. 6 illustrates an example of an original 2D video and the corresponding 3D video.

B. Classification Models

AdaBoost—AdaBoost uses weak classifiers to learn from the training set [14]. In each round, it builds a weak classifier to assign a label based on the votes from the classifier set. It gives more weight to the training data which is misclassified in the next iteration. In order to boost the performance, it only requires the initial classification to be better than random guess. The dimensions associated with the weak classifiers give us the most important feature to AdaBoost in our feature vector space.

SVM—As a second baseline, a binary SVM is trained with one label per frame using a Radial Basis Function (RBF) kernel. The SVM does not use temporal information hence the training and test is based on features at the isolated frame. The output margin to the SVM hyper plane is used to plot the receiver operating characteristic (ROC) curves. Two parameters are validated: C, the penalty parameters of the error term in the SVM objective function, and γ, the RBF kernel parameter. Both parameters were validated with values 2×10^k, $k \in \{-2, -1, \ldots, 2\}$.

LDCRF—In order to take the temporal information of 3D model sequences into consideration, we propose to apply the Latent-Dynamic Conditional Random Field (LDCRF) [3] theory on the 3D binary edge feature. LDCRF was designed for recognition on un-segmented sequential data. Previous research has shown the contribution of LDCRF to recognize specific Action Unit correlated to smile in spontaneous video sequence [5]. Given a sequence of observation $\boldsymbol{X} = \{x_1, x_2, \ldots, x_m\}$ and corresponding labels $\boldsymbol{Y} = \{y_1, y_2, \ldots, y_m\}$, a latent conditional model is defined as in Eq. (5). A set of hidden variables $\boldsymbol{h} = \{h_1, h_2, \ldots, h_m\}$ monitors the "sub-structure" which is not observed in the training set. Suppose the hidden states are disjoint, i.e. for each class label y_i, a set H_{yi} defines its possible hidden states formed from members of $\boldsymbol{h}$, then we have $\bigcap_{i=1}^{m} H_{yi} = \phi$. In the case of our training set, each y_i is a class label for the i^{th} frame of 3D sequence, and it is in the set of {0, 1} for the binary classification problem. Each frame observation x_i is represented by the feature, which is the vectorization of the image matrix.

$$P(Y \mid X, \theta) = \sum_{h:\, \forall h_i \in H_{yi}} P(Y \mid h, X, \theta) P(h \mid X, \theta) \tag{5}$$

$P(h \mid X, \Theta)$ is defined using the CRF formulation[2].

The optimal weights Θ^* is searched by gradient ascent which looks for the maximization of Eq. (6) for n training sequences, in which the first term is the conditional log-likelihood of the training data, and the second term is the norm-l_2 regularization. To estimate the label y_i^* of a frame in the test sequence, $P(h_j \mid X, \Theta^*)$ is computed for all possible hidden states, the summation of all the members of each H_{yi}, is obtained, and the label associated with the maximum sum is chosen.

$$L(\theta) = \sum_{i=1}^{n} \log P(Y_i | X_i, \theta) - \frac{\|\theta\|^2}{2\sigma^2} \quad (6)$$

The sum also represents the confidence score of the prediction. It is used as the intensity indicator of the class.

To training LDCRF with consideration of long-range dependencies in the sequence, window size $\in \{0, 1, 2\}$ is validated. And various regularization parameter values are tested $\delta = 2 \times 10^k$, $k \in \{-2, -1, \ldots, 2\}$. Different number of hidden states are also considered, $h_n \in \{1, 2, 3, 4\}$.

C. Feature Vector Generation

3D binary edge feature is generated in the same way as described in Section 2. When using AdaBoost as a classifier, we apply it for dimension reduction as well. When using the other classifiers, e.g., SVM and LDCRF, the PCA is applied for dimension reduction. Eventually, the dimension of 3D-BE features is reduced to 40 in our experiment.

D. Methodology

A 10-fold cross validation is used to select the parameters for each AU combination. The performance is evaluated with the leave-one-subject-out method on 41 subjects.

After selecting the training subjects, for AU6&7, and AU10&12, positive sequence segments are selected from their *pain* sequences (Task 6), and the negative samples are from the *sad* sequences (Task 2) since in Table III this category displays the most significant difference compared to Task 6 with this AU set. If the subject does not have any positive segments, it is used in test but is discard from the training set.

For AU9, because only 15 subjects out of 41 show this AU in pain sequence, while 27 subjects show it during Task 8 when they experience feeling of disgust, we decide to select positive samples from it, and test on both task 6 and 8. This setting could illustrate the potential to conduct training with the target AU from any emotional categories.

To test the model, the original two AU coded sequences of *pain* task and *sad* task from the remaining subject are passed to the classifier. The log-likelihood of each frame regarding the AU labels is given and the ROC curve can be drawn. The average Area Under Curve (AUC) weighted by positive sample number is used to evaluate the performance.

E. Result and Evaluation

(1) Comparison with different learning approaches—Based on the weighted AUC, we choose the best model parameters and compare the three models as shown in Fig. 7. Due to the use of temporal information, LDCRF generates the best result (green line) among the three classifiers for AU recognition. For AU9, if we train and test with disgust, the weighted AUC=0.93877 (the right plot in Fig. 7), which also shows that our proposed

3D edge features and the classification models are also applicable to the other expression recognition.

(2) Evaluation of consistency between AU intensity and classification score—Based on the work reported by Lucey et al. in [7], the AU classification scores can be fused to represent the PSPI. Their result shows that the fused score is a good indicator of pain intensity. By intuition, a higher level of pain-related AU intensity should signify a stronger pain expression, and vice versa. Here we apply the classification result generated by our system trained with AU binary codes to describe the AU intensity level, which could be potentially used to infer the pain level.

In BP4D-Spontaneous, a subset of Action Units is coded with 6-point intense scale that range from 0 (absent) to 5 (maximum intensity). We evaluate the consistency of the log-likelihood generated by our classifier and the ground truth intensity coding on AU6, 10 and 12. The log-likelihood is in range of [0, 1]. We then divide the range into 5 isometric bins. For any value larger than 0, it is mapped to one intensity level from 1 to 5 depending on the bin it belongs to. This is a simple linear mapping method. In addition, considering the pain expression development in both onset and offset stages is non-linear, we apply a polynomial function with a power at order of 6 to the log-likelihood values for a better conversion.

We calculate the level difference between the automatically estimated intensity and ground truth one to one from the result of the leave-one-subject-out test based on 41 subjects. The difference is calculated based only on the correctly classified positive frames to avoid (1) the impact from large level differences due to mis-classification and (2) the bias introduced by zero difference from the negative frames. The result is shown in Table IV. Clearly, the non-linear mapping fits the conversion better than the linear mapping since both mean and standard deviation are smaller in the non-linear case. In general, the average difference of automatically estimated intensity and the ground truth is about one level, thus indicating a moderate level of consistency between the automatically estimated AU intensity and the ground truth.

(3) Pain intensity detection—Based on the findings in subsection (2), it is possible to use the log-likelihood to represent the intensity of AU we choose. And similar to PSPI, we can composite the pain intensity from the AU set. We take the average value of AU6&7, 9, and 10&12's classification output from LDCRF to illustrate the level of pain (i.e., intensity). Fig. 8 shows an example of the pain intensity rating. The upper part of Fig. 8 shows 4 frames from the sequence, including 2D texture, 3D model and the 3D binary edge features, which correspond to the individual times through the red lines in the lower part. In the lower part, we illustrate the AU coding changes along with time. The blue dashed lines are the manually coded binary AU labels, and the green lines are the probabilities of the recognition given by the automated AU detection system. Our result matches the ground truth very well. The bar represents the pain intensity calculated by the AU set. The warmer the color, the more pain the subject feels. In frame 1, all the AUs are detected, which indicates an intensive pain, while in frame 2, the face is relaxed and none of the three AUs appear. Importantly, although the AU is only binary coded, our model can detect the AU intensity change. Note in Fig. 8,

from frame 3 to frame 4, the green lines of AU6&7, 9 and 10&12 climb to the peak gradually. This illustrates the intensity of pain goes up along with the pain expression change.

VI. Conclusion

In this paper, we proposed the 3D binary edge on the 3D range models to represent facial expressions, particularly for application in AU detections for pain expression analysis. First we define our novel feature descriptor. It outperforms other state of the art 3D/2D features in the comparison. Then based on the study of different spontaneous expression categories containing 52,578 samples, we find that the AU set including AU6&7, 9, and 10&12 is a good indicator of genuine pain expression. Taking the temporal information into account, we apply the latent-dynamic conditional random field model for AU classification on 3D dynamic facial sequences. Then we discuss on the consistency of classification score and AU intensity. Finally an automated pain intensity rating system based on 3D dynamic videos illustrates the possibility and essentiality to monitor the pain in a less restricted environment regarding pose and lighting change.

Our future work will collect different modality data and analyze the signals from multiple sources of data for verification and/or fusion in order to infer the fine level of pain intensity. The 3D binary edge feature will be assessed with more AUs based on the entire database. The AU set correlated with other genuine expressions will also be investigated. The current RGB-D imaging sensors offer a great potential to acquire depth videos in real time and our GPU based implementation of 3D binary edge feature is able to directly integrated with the data stream. And since we use the projective space to efficiently access the neighbor normal information, triangulation of the point cloud data is not necessary at all. We plan to extend our new 3D edge features and classification models to different quality 3D videos for real-time application.

Acknowledgment

This material is based upon the work supported in part by the National Science Foundation under grants IIS-1051103, IIS-1051169, CNS-1205664, and CNS-1205195. We would like to thank Dean Rosenwald and Shawn Zuratovic for FACS coding, Nicki Siverling and Jeffrey Girard for technical assistance. We would also like to thank Dr. Peter Gerhardstein for his help in data collection.

References

[1]. Sander Pedro V., et al. “Silhouette clipping.” SIGGRAPH 2000.

[2]. Lafferty J, McCallum A, and Pereira F “Conditional random fields: Probabilistic models for segmenting and labeling sequence data.” Proceedings of the 18th ICML 2001.

[3]. Morency L-P, Quattoni A, and Darrell T “Latent-dynamic discriminative models for continuous gesture recognition.” CVPR 2007.

[4]. Ambadar Z, et al. (2009). All smiles are not created equal: Morphology and timing of smiles perceived as amused, polite, and embarrassed/nervous. Journal of Nonverbal Behavior, 33(1), 17–34. [PubMed: 19554208]

[5]. Bartlett M, Littlewort G, et al. (2014, in press). Automatic decoding of facial movements reveals deceptive pain expressions. Current Biology.

[6]. Prkachin K, Solomon P, et al. “The structure, reliability and validity of pain expression: Evidence from patients with shoulder pain.” Pain, 139.2 (2008): 267–274. [PubMed: 18502049]

[7]. Lucey P, et al. “Automatically detecting pain using facial actions.” In ACII 2009.

[8]. Littlewort G, Bartlett M, et al. “Faces of pain: automated measurement of spontaneous all facial expressions of genuine and posed pain.” ACM ICMI 2007.

[9]. Ekman P, Friesen WV, & Hager JC (2002). “Faical action coding system” Research Nexus, Network Research Information.

[10]. Lucey P, Cohn J, Matthews I, Lucey S, et al. “Automatically detecting pain in video through facial action units.” IEEE Trans. on SMC, Part B, 41(3):664–674, 2011.

[11]. Zhang X, Yin L, Cohn JF, et al. “A High-resolution spontaneous 3D dynamic facial expression database.” FG 2013.

[12]. Sandbach G, Zafeiriou S, Pantic M, and Yin L “Static and dynamic 3D facial expression recognition: A comprehensive survey.” Image and Vision Computing, 30(10), 2012 P683–679.

[13]. Sun Y and Yin L “Facial expression recognition based on 3D dynamic range model sequences.” ECCV 2008.

[14]. Viola Paul, and Jones Michael. “Fast and robust classification using asymmetric AdaBoost and a detector cascade.” Advances in Neural Information Processing Systems 2 (2002): 1311–1318.

[15]. Reale M, Zhang X, and Yin L “Nebula feature: a space-time feature for posed and spontaneous 4D facial behavior analysis.” FG 2013.

[16]. Zhao G and Pietikainen M “Dynamic texture recognition using local binary patterns with an application to facial expressions.” IEEE Trans. PAMI, 29(6): 915–928, 6 2007.

[17]. Wang J, Yin L, Wei X, and Sun Y “3D facial expression recognition based on primitive surface feature distribution.”, IEEE CVPR 2006.

[18]. Lucey P, Cohn JF, et al. “Recognizing emotion with head pose variation: Identifying pain segments in video.” IEEE Transactions on Systems, Man, and Cybernetics - Part B, 41(3), 664–674.

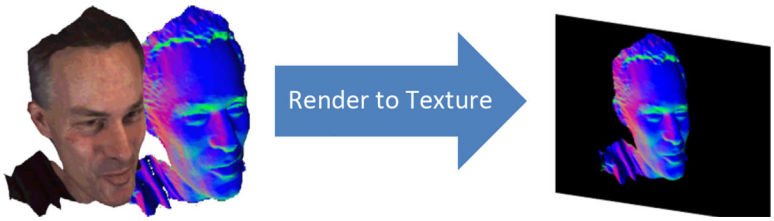

Fig. 1.
Render 3D geometry shape to texture pipeline.

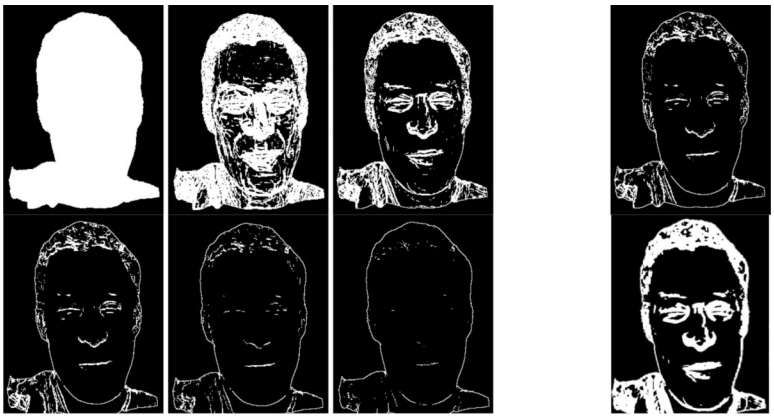

Fig. 2.

(a) Edge variation along with different thresholds. From left to right, top to bottom, the threshold is 0, 0.004, 0.02, 0.08, 0.16, and 0.32. (We denote it as 3D binary edge: 1 representing edges, and 0 for non-edges). (b) Effect of Gaussian blurring as a pre-process step. The threshold t for both images is 0.08. Top one is without Gaussian blur, bottom one is with Gaussian blur at a kernel size of 9 pixels.

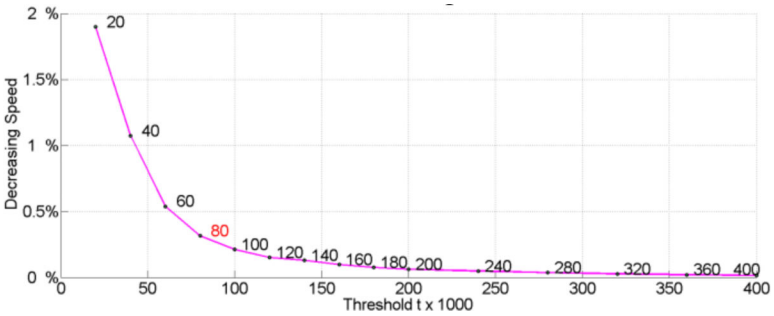

Fig. 3.
Noise reduction speed change with different threshold *t.*

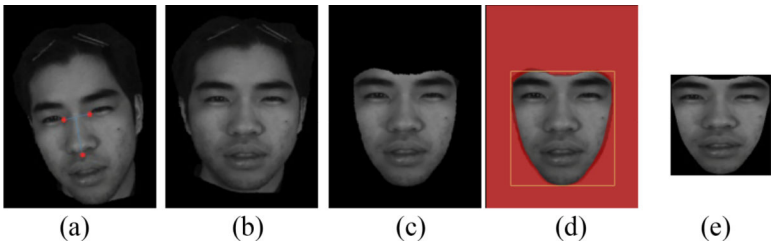

Fig. 4.
Diagram of feature registration. For readability, textured model is used as illustration.

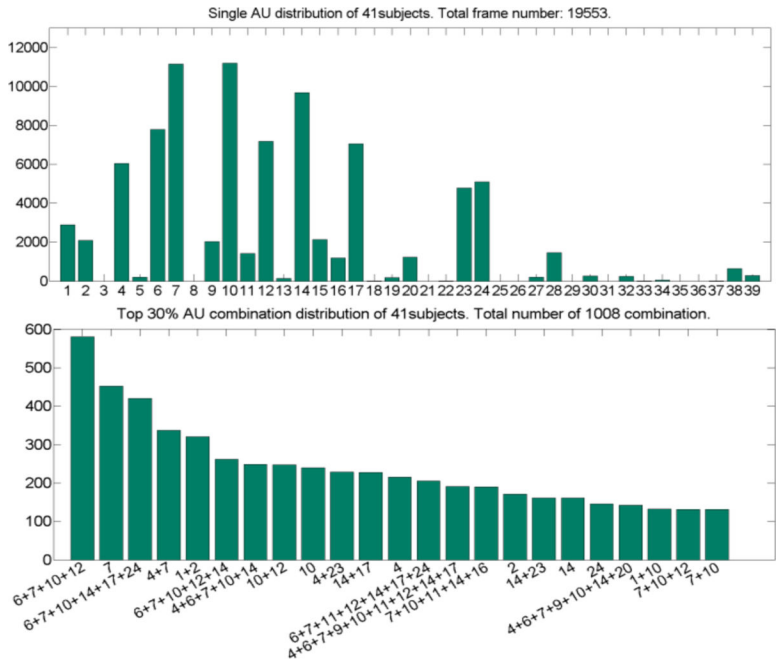

Fig. 5.

AU histogram with respect to frame count of pain sequences of BP4D-Spontaneous. The upper chart shows the result for the coded AUs. The lower chart shows the top 30% AU combinations in descend order.

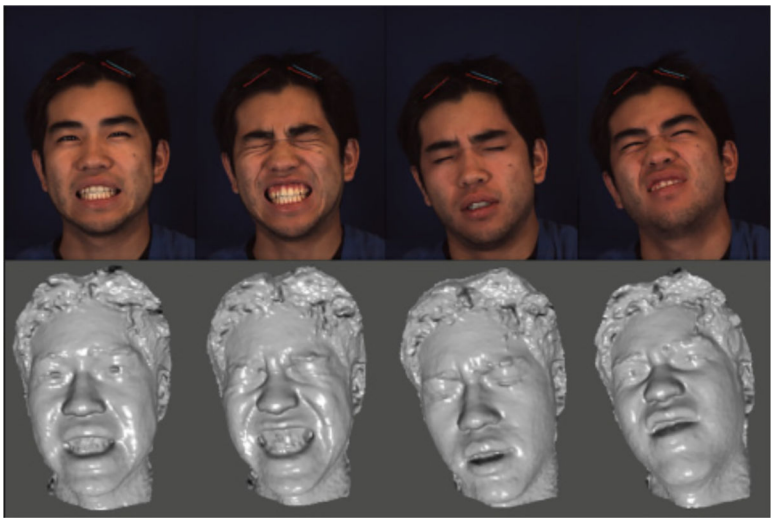

Fig. 6.
Sample frames of a pain sequence from the BP4D-Spontaneous (Upper: 2D video; Lower: 3D video)

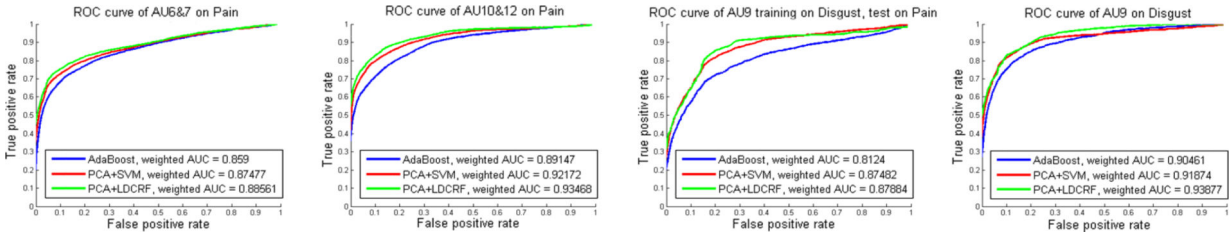

Fig. 7.

ROC curves of different classifiers. From left to right they are for AU6&7, AU10&12, AU9 test on the pain sequence (Task 6), and AU 9 test on the disgust sequence (Task 8). Green line is for PCA+LDCRF, red line is for PCA+SVM, and blue line is for AdaBoost.

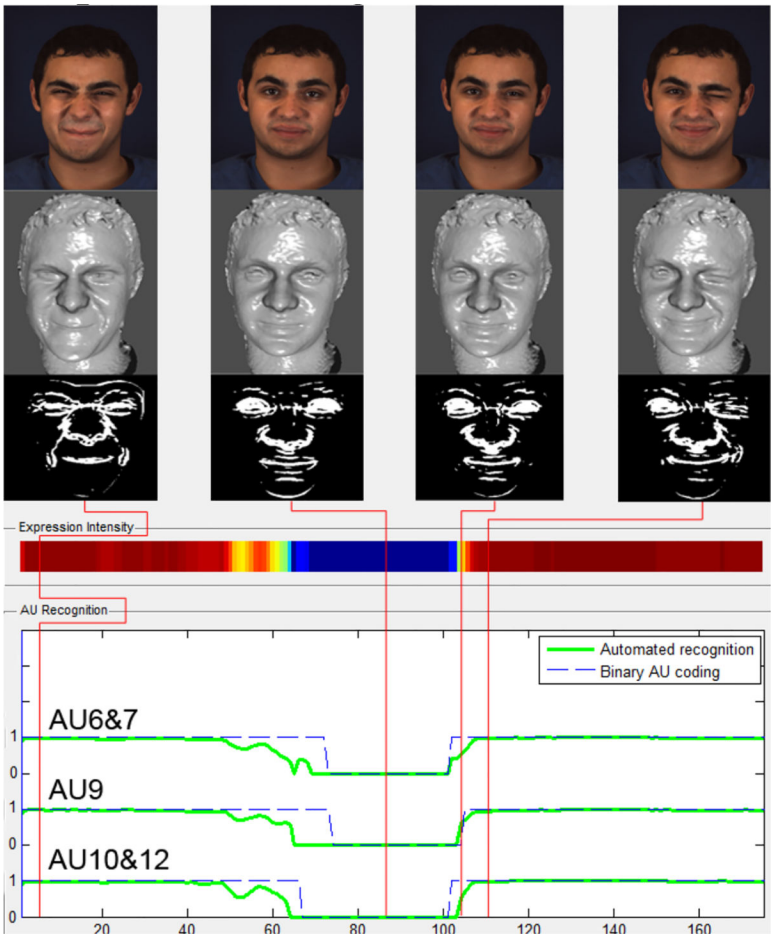

Fig. 8.

Sample AU outputs (green lines) for a 7 seconds pain sequence. Manually labeled ground truth (blue dashed lines) are superimposed for comparison. Different imaging modalities are illustrated.

TABLE I.

Twelve Action Units tested

AU1	Inner Brow Raiser	AU12	Lip Corner Puller
AU2	Outer Brow Raiser	AU14	Dimpler
AU4	Brow Lowerer	AU15	Lip Corner Depressor
AU6	Cheek Raiser	AU17	Chin Raiser
AU7	Lid Tightener	AU23	Lip Tightener
AU10	Upper Lip Raiser	AU24	Lip Pressor

TABLE II.

Accuracy for AU detection (in percentage %)

AU \ Method	3D-based Features			2D-based Features		
	3D-BE	Shape Index	Nebular	LBP-TOP Depth	LBP-TOP texture	Gabor texture
1	64.6	53.2	54.1	52.4	57.9	61.0
2	57.1	59.4	63.0	55.9	59.2	60.8
4	66.5	61.6	58.7	51.1	53.3	58.6
6	69.0	70.4	67.6	61.3	64.8	67.6
7	64.5	64.1	58.9	52.4	55.4	64.4
10	68.7	68.0	66.4	56.9	62.1	70.5
12	75.2	75.2	57.3	53.3	59.1	74.5
14	55.9	53.5	54.5	52.8	52.3	52.8
15	66.2	65.1	66.0	63.1	64.5	60.9
17	64.2	59.2	61.8	53.3	60.0	62.2
23	63.6	50.9	60.6	59.3	58.5	61.7
24	75.9	67.9	63.3	62.9	63.4	72.1
Avg.	66.0	62.3	61.3	56.2	59.2	63.9

TABLE III.

p score of paired t-test on differences between pain and other emotion tasks on AU sets. Not significant difference situation ($p>0.05$) is illustrated with *.

AU set \ Task	1	2	3	4	5	7	8
4, 6, 7, 9, 10	0.0257	0.0196	0.0014	0.0005	0.0013	0.2346*	0.0331
4, (6&7), (9&10)	0.6041*	0.2754*	0.0012	0.2113	0.8187*	0.0245	0.0343
4, 6, 7, 10, 12	0.0042	0.0067	0.0019	0.0004	0.0007	0.0992*	0.0290
4, (6&7), (10&12)	0.0052	0.0546*	0.0040	7.0879×10^{-6}	0.0125	0.3507*	0.0227
4, (6&7), 9, (10&12)	0.0074	0.0475	0.0034	9.7693×10^{-6}	0.0156	0.3951*	0.0120
(6&7), 9, (10&12)	0.0003	1.0536×10^{-8}	0.0032	1.3646×10^{-7}	8.1435×10^{-5}	0.0307	0.0043

TABLE IV.

Statistics of the difference between automatically estimated AU intensity and the ground truth.

AU \ Type	Linear mapping		6-th power mapping	
	MEAN	**STD**	**MEAN**	**STD**
6	1.949	0.998	1.312	0.924
10	1.678	0.959	1.243	0.842
12	2.130	1.053	1.436	1.033