SVM Classification of Neonatal Facial Images of Pain

Sheryl Brahnam[1], Chao-Fa Chuang[2], Frank Y. Shih[2], and Melinda R. Slack[3]

[1] Missouri State University, Computer Information Systems, 901 South National,
Springfield MO 65804, USA
Shb757f@smsu.edu
[2] New Jersey Institue of Technology, University Heights, Newark, NJ 07102
{cxc1235, shih}@njit.edu
[3] Medical Director of Neonatology, St. John's Hospital, 1235 E. Cherokee,
Springfield, MO 65894, USA
Melinda_slack@pediatrix.com

Abstract. This paper reports experiments that explore performance differences in two previous studies that investigated SVM classification of neonatal pain expressions using the Infant COPE database. This database contains 204 photographs of 26 neonates (age 18-36 hours) experiencing the pain of heel lancing and three nonpain stressors. In our first study, we reported experiments where representative expressions of all subjects were included in the training and testing sets, an experimental protocol suitable for intensive care situations. A second study used an experimental protocol more suitable for short-term stays: the SVMs were trained on one sample and then evaluated on an unknown sample. Whereas SVM with polynomial kernel of degree 3 obtained the best classification score (88.00%) using the first evaluation protocol, SVM with a linear kernel obtained the best classification score (82.35%) using the second protocol. However, experiments reported here indicate no significant difference in performance between linear and nonlinear kernels.

1 Introduction

Accurate assessment of pain in neonates is a difficult yet crucial task. The clinical definition of pain assumes the person experiencing pain has the ability to articulate the location, duration, quality, and intensity of their pain experience. Although nonverbal self reporting methods have been devised that allow preverbal children to indicate their pain levels by pointing to abstract renditions of facial expressions expressive of increasing levels of discomfort, neonates must rely exclusively on the proxy judgments of others [3].

Several pain assessment measures have been developed to assist clinicians in diagnosing neonatal pain. Most of these instruments rely on the neonate's facial displays. Facial displays are considered the gold standard of pain assessment [4] because they are the most specific and consistent indicators of pain. The facial characteristics of neonatal pain displays include prominent forehead, eye squeeze, naso-labial furrow, taut tongue, and an angular opening of the mouth [5]. Despite the fact that neonatal facial displays of pain are the most reliable source of pain assessment, instruments

I. Bloch, A. Petrosino, and A.G.B. Tettamanzi (Eds.): WILF 2005, LNAI 3849, pp. 121 – 128, 2006.

based on facial displays are unsatisfactory because clinicians tend to underrate pain intensity [6] and often fail to utilize all the information available to them in the infants facial signals [7].

In an attempt to bypass the unreliable observer, our research group is investigating the potential benefits face recognition technology would offer pediatric clinicians in diagnosing neonatal pain. Applying face recognition techniques to medical problems is a novel application area. Gunaratne and Sato [17] have used a mesh-based approach to estimate asymmetries in facial actions to determine the presence of facial motion dysfunction for patients with Bell's palsy, and Dai et al. [12] have proposed a method for observing the facial expressions of patients in hospital beds. The facial images used in the Dai et al. study, however, were not of actual patients but rather of subjects responding to verbal cues suggestive of medical procedures and conditions. Our work with neonatal pain expressions is the only other research we are aware of that uses face recognition techniques to diagnose medical problems.

We began work on this problem by developing the Infant COPE database. The facial displays of 26 neonates between the ages of 18 hours and 3 days old were photographed experiencing the pain of a heel lance and a variety of stressors, including transport from one crib to another, an air stimulus on the nose, and friction on the external lateral surface of the heel.

In our initial study [1], three face classification techniques, Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), and Support Vector Machines (SVMs), were used to classify the faces into two categories: pain and nonpain. The training and testing sets contained multiple samples of each subject in each expression category. No two samples were identical as each varied slightly in angle and facial configuration. While, ideally, as is the case with speech recognition software, samples of individual subjects would be available to personalize the classifier, in a clinical setting this is not practical as the typical newborn's stay is short-term. The evaluation protocol used in our first study would probably only be applicable in intensive care situations where neonates have longer stays that present opportunities for collecting facial samples. It is more realistic to assume that the classifier will need to be trained on one set of subjects and then applied out of the box to future newborns. In [2], an evaluation protocol was developed that evaluated trained classifiers using unknown subjects.

Results of the two studies were contradictory in terms of the best kernel to use with SVM. An SVM with polynomial kernel of degree 3 obtained the best classification score (88.00%) in the first study, and an SVM with a linear kernel obtained the best classification score (82.35%) in the second study. Sampling error caused by the small number of images in the sample pool is one possible explanation for this discrepancy. A set of new experiments using the first protocol was designed to explore sample error. The results of these experiments, reported in section 4, suggest that there is no significant difference in the performance of an SVM with a linear kernel and an SVM with a polynomial kernel of degree 3.

In section 2, we describe of the facial displays in the infant COPE database more completely. In section 3, we outline the two experimental protocols, designated A and B, used in the earlier studies. In section 4, we compare SVM classification rates reported in the two studies, along with baseline PCA and LDA rates. We then present the results of a new study that varies the size of the sample pool. We conclude the paper, in section 5, by pointing out some limitations in our current work and by offering suggestions for future research.

2 The Infant COPE Database

The Infant COPE Database, described more completely in [1] and [2], contains 204 facial images of 26 neonates experiencing the pain of a heel lance and three nonpain stressors: transport from one crib to another (a stressor that triggers crying that is not in response to pain), an air stimulus on the nose (a stressor that provokes eye squeeze), and friction on the surface of the heel (a stressor that produces facial expressions of distress that are similar to the expressions of pain). In addition to these four facial displays, the database includes images of the neonates in the neutral state of rest.

Fig. 1 provides two example sets of the five neonatal expressions of rest, cry, air stimulus, friction, and pain included in the Infant COPE database. Of the 204 images in the database, 67 are rest, 18 are cry, 23 are air stimulus, 36 are friction, and 60 are pain.

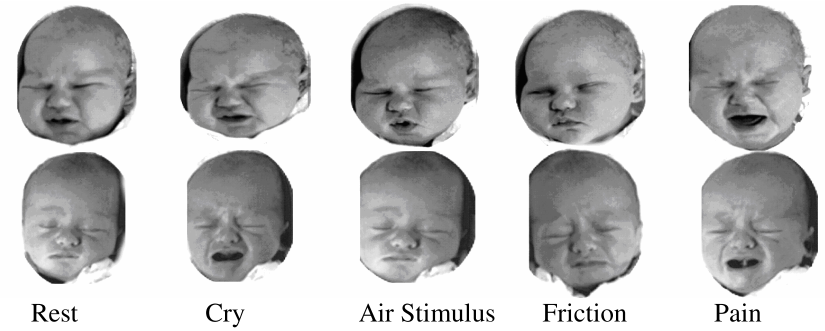

Fig. 1. Examples of the five facial expressions in the Infant COPE database

The data collection process complied with the protocols and ethical directives for research involving human subjects at Missouri State University and St. John's Health System, Inc. Informed consent was obtained from a parent, usually the mother in consultation with the father. Most parents were recruited in the neonatal unit of a St. John's Hospital sometime after delivery. Only mothers who had experienced uncomplicated deliveries were approached. The subjects were born in a large Midwestern hospital in the United States of America. All neonates used in the study were Caucasian, evenly split between genders (13 boys and 12 girls), and in good health. The interested reader is referred to [1] and [2] for more information on the data collection design.

3 Evaluation Protocols

In [1] and [2], images of the five facial expressions in the Infant COPE database were grouped into two categories: pain and nonpain. The set of nonpain images combined the rest, cry, air stimulus, and friction images and contained a total of 144 images. The set of pain images consisted of the remaining 60 images.

The evaluation protocol used in the first study, designated here as protocol A, focused on facial expression representation. The two classes of pain and nonpain facial expressions included representative images of all 26 subjects. Using a cross-validation technique, classification was a four step process. In step 1, the images were randomly divided into ten segments. In step 2, nine out of the ten segments were used in the training session. The remaining segment was used in testing, and an average classification score was obtained from the testing set of images. In step 3, steps 1 and 2 were repeated ten times. Finally, in step 4, the ten classification scores were averaged to obtain a final performance score for each classifier.

In the second study, we trained the classifiers on one set of subjects and tested them on another. Using protocol B, twenty-six experiments were performed, one for each subject. The facial images of 25 subjects formed the testing set, and the images of the remaining subject formed the testing set. The 26 classification scores were averaged to obtain a final performance score for each classifier.

4 Experimental Results

In this section, we compare the SVM performance results reported in the first two studies. We also introduce a new set of experiments designed to determine whether the performance differences in the earlier studies are due to sampling error.

SVMs with five kernels (linear, RBF, polynomial degree 2, polynomial degree 3, and polynomial degree 4) were assessed using protocols A and B. The regularization parameter, C, used in the SVM experiments was determined using a grid search. Since the recognition rates in our experiments were not significantly different in terms of different values for C, we adopted the regularization parameter C=1. The bandwidth parameter in SVM using RBF kernels was also optimized using a grid search. For comparison purposes, baseline PCA and LDA using the sum of absolute differences, or L1 distance metric, were also evaluated.

The SVM, PCA, and LDA experiments were processed in the MATLAB environment under the Windows XP operating system using a Pentium 4 – 2.80 GHz processor. SVM was implemented using the OSU SVM Classifier MATLAB Toolbox developed by Ohio State University.

The general experimental procedures used in all our experiments can be divided into the following stages: preprocessing, feature extraction, and classification. In the preprocessing stage, the original images were cropped, rotated, and scaled. Eyes were aligned roughly along the same axis. The original 204 images, size 3008 x 2000 pixels, were also reduced to 100 x 120 pixels. In the feature extraction stage, facial features were centered within an ellipse and color information was discarded. The rows within the ellipse were concatenated to form a feature vector of dimension 8383 with entries ranging in value between 0 and 255. PCA was then used to reduce the dimensionality of the feature vectors further. The first 70 principle components resulted in the best classification scores. Finally, in the classification stage, the feature vectors were used as inputs to the classifiers.

Table 1 compares the average classification scores obtained using the two protocols. Referring to Table 1, the average classification score for PCA was 80.36% and for LDA 80.32%. SVM, as expected, outperformed both PCA and LDA, except in the case of RBF kernel. Given previous reports in facial expression classification using SVM (see, for instance, [8]), we did not expect the RBF kernel performance to be as low as it was. An SVM with polynomial degree 3 provided the best recognition rate of 88.00% in the experiments using protocol A. An SVM with linear kernel provided the best recognition rate of 82.35% using protocol B.

Table 1. Comparison of SVM classification rates using protocol A and B

Type of svm	Protocol A	Protocol B	Average (A & B)
Linear	83.67%	**82.35%**	83.01%
Polynomial degree = 2	86.50%	79.90%	83.20%
Polynomial degree = 3	**88.00%**	80.39%	84.20%
Polynomial degree = 4	82.17%	72.06%	77.12%
RBF	70.00%	70.10%	70.05%
PCA with L1 distance	80.33%	80.39%	80.36%
LDA with L1 distance	83.67%	76.96%	80.32%

There are several possible explanations for the kernel performance differences in the two studies. The most likely cause for the discrepancy is sampling error due to the small number of images in the sample pool. The average performance of the SVMs using the two kernels, for instance, is very close, the difference being only 1.18%. However, since the data in the training sets used in the two sets of experiments differ only in a few inputs (approximately 15%), we questioned this assumption.

To determine if the difference in kernel performance is the result of sampling error, we performed new experiments that varied the size of the sample pool. We did this by comparing SVM classification of pain expressions to each of the other four facial displays. This resulted in pool sizes of 83 images for pain versus air stimulus, 78 images for pain versus cry, 96 images for pain versus friction, and 127 images for pain versus rest. Only protocol A was used in these experiments, as splitting expressions for each subject (protocol B) resulted in pool sizes that were too small for training.

Tables 2-4 present the results of the new set of experiments. The average performance of the four experiments using SVM with a linear kernel is 85.51%, and the average performance of SVM with a polynomial kernel of degree 3 is 87.74%. The difference in kernel performance (2.23%) is half that in [1] (4.33%), which also used protocol A. This leads us to believe that sample error is most likely the cause of kernel performance differences. As far as neonatal facial expressions are concerned, the results of the new set of experiments suggest that there is no significant classification difference in SVMs using a linear kernel versus a polynomial kernel of degree 3

Table 2. Pain vs. Air stimulus

Method	Classification score
Linear	90.00%
Polynomial degree = 2	77.78%
Polynomial degree = 3	83.33%
Polynomial degree = 4	78.89%
RBF	66.67%

Table 3. Pain vs. Cry

Method	Classification score
Linear	71.25%
Polynomial degree = 2	78.75%
Polynomial degree = 3	80.00%
Polynomial degree = 4	76.25%
RBF	75.00%

Table 4. Pain vs. Friction

Method	Classification score
Linear	90.00%
Polynomial degree = 2	96.00%
Polynomial degree = 3	93.00%
Polynomial degree = 4	92.00%
RBF	60.00%

Table 5. Pain vs. Rest

Method	Classification score
Linear	90.77%
Polynomial degree = 2	84.62%
Polynomial degree = 3	94.62%
Polynomial degree = 4	86.15%
RBF	53.85%

Kernel. This conclusion is consistent with [8], which examined SVM expressing classification performance using a number of adult facial databases.

5 Conclusion

This paper reports new experiments intended to explore performance differences in two pervious studies that investigated SVM classification of neonatal pain expressions using the Infant COPE database. This database contains 204 photographs of 26 neonates (age 18-36 hours) experiencing the acute pain of a heel lance and three non-pain stressors.

The SVM classifiers were trained and tested using images divided into two sets: pain and nonpain. Two separate evaluation protocols, designated in this paper as A and B, were also used. Protocol A, described in [1], assumes that samples of neonates are available for customizing the classifier. Representative expression samples of all 26 subjects were thus included in both the training and the testing sets. Protocol B, described in [2], assumes that the classifiers will be trained on one sample and tested on another. The facial images of 25 subjects formed the training set, and the images of the remaining subject formed the testing set. A total of 26 experiments were thus performed using protocol B, one for each subject. An SVM of polynomial kernel degree 3 obtained the best classification score of 88.00% using protocol A, and an SVM with a linear kernel obtained the best classification score of 82.35% using protocol B.

We assumed that the difference in kernel performance was due to sample error. A set of new experiments that varied the size of the sample pool was performed to test our assumption. In these experiments, which used protocol A, the average performance of SVM with a linear kernel was 85.51%. With polynomial kernel of degree 3 it was 87.74%. The difference in kernel performance is half that reported in [1], which also used protocol A. This leads us to believe that there is no significant performance difference using SVM with a linear kernel and polynomial kernel of degree 3.

We would like to conclude this paper with some general remarks concerning the limitations, future directions, and significance of our research in neonatal pain classification.

There are a number of limitations in our current work. First, these studies use two-dimensional still photographs and do not consider the dynamic and multidimensional nature of facial expressions. The classification rates reported in these two studies, however, are consistent with facial expression classification rates reported using video displays of adult facial expressions. For example, [9] reports classification rates between 88%-89%. Second, we have yet to explore facial shape information in the facial displays. Third, the focus thus far has been on acute pain. We have not examined facial expressions in reaction to repeated pain experiences.

In terms of future directions, we are working on addressing the limitations noted above. We are currently collecting video data of neonates experiencing additional stressors and two types of pain: acute and repeated pain. We are also working on experiments that incorporate shape information. In addition, we are examining the classification performance of a number of neural network architectures. For instance, the performance rate of NNSOA, a neural network simultaneous optimization algorithm, using protocol B is reported in [2].

Finally, in terms of significance, we expected that the performance of SVMs in the first study that used protocol A would be better than SVMs in the second study that used protocol B. What we did not know is how well SVM performance would hold up using protocol B. SVM results compare well, and the classification rates in both studies indicate a high potential for applying standard face recognition technology to this problem domain. We believe the results of the SVM experiments encourage further explorations using more sophisticated face recognition technologies.

References

1. Brahnam, S., Chuang, C., Shih, F.Y., and Slack, M.R. Machine Recognition and Representation of Neonate Facial Displays of Acute Pain. International Journal of Artificial Intelligence in Medicine (AIIM), (in press and available on publisher website)
2. Brahnam, S., Chuang, C., Sexton, R., Shih, F.Y., and Slack, M.R. Machine Assessment of Neonatal Facial Expressions of Acute Pain. Decision Support Systems, (in revision)
3. Wong, D. and Baker, C. Pain in Children: Comparison of Assessment Scales. Pediatric Nursing, Vol. 14: **1** (1988) 9017
4. Craig, K.D. The Facial Display of Pain in Infants and Children. In: Craig, K.D. (ed.). The Facial Display of Pain in Infants and Children, IASP Press, Seattle, (1998) 103-121
5. Grunau, R.E., Grunau, R.V.E., and Craig, K.D. Pain Expression in Neonates: Facial Action and Cry. Pain, Vol. 28: **3** (1987) 395-410
6. Prkachin, K.M., Solomon, P., Hwang, T., and Mercer, S.R. Does Experience Influence Judgments of Pain Behaviour? Evidence from Relatives of Pain Patients and Therapists. Pain Research and Management, Vol. 6: **2** (2001) 105-112
7. Prkachin, K.M., Berzins, S., and Mercer, S.R. Encoding and Decoding of Pain Expressions: A Judgement Study. Pain, Vol. 58: **2** (1994) 253-59
8. Littlewort, G., Bartlett, M.S., Fasel, I., Susskind, J., and Movellan, J. Dynamics of Facial Expression Extracted Automatically from Video. The First IEEE Workshop on Face Processing in Video. Washington, DC. (2004)
9. Cohen, I., Sebe, N., Garg, A., Chen, L.S., and Huang, T.S. Facial Expression Recognition from Video Sequences: Temporal and Static Modeling. Computer Vision and Image Understanding, Vol. 91: **1-2** (2003) 160-187