Artificial Intelligence in Medicine (2005) **xxx**, xxx–xxx

ELSEVIER

ARTIFICIAL INTELLIGENCE IN MEDICINE

http://www.intl.elsevierhealth.com/journals/aiim

Machine recognition and representation of neonatal facial displays of acute pain

Sheryl Brahnam [a,*], Chao-Fa Chuang [b], Frank Y. Shih [b], Melinda R. Slack [c]

[a] *Department of Computer Information Systems, Missouri State University, 3rd Floor Glass Hall, 901 South National, Springfield, MO 65804, USA*
[b] *Computer Vision Laboratory, College of Computing Sciences, New Jersey Institute of Technology, University Heights, Newark, NJ 07102, USA*
[c] *Medical Director of Neonatology, St. John's Hospital, 1235 E. Cherokee, Springfield, MO 65804, USA*

Received 10 June 2004; received in revised form 1 December 2004; accepted 6 December 2004

KEYWORDS
Neonatal pain recognition;
Medical face classification;
Support vector machines;
Linear discriminant analysis;
Principal component analysis

Summary

Objective: It has been reported in medical literature that health care professionals have difficulty distinguishing a newborn's facial expressions of pain from facial reactions to other stimuli. Although a number of pain instruments have been developed to assist health professionals, studies demonstrate that health professionals are not entirely impartial in their assessment of pain and fail to capitalize on all the information exhibited in a newborn's facial displays. This study tackles these problems by applying three different state-of-the-art face classification techniques to the task of distinguishing a newborn's facial expressions of pain.
Methods: The facial expressions of 26 neonates between the ages of 18 h 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 (a disturbance that can provoke crying that is not in response to pain), an air stimulus on the nose, and friction on the external lateral surface of the heel. Three face classification techniques, principal component analysis (PCA), linear discriminant analysis (LDA), and support vector machine (SVM), were used to classify the faces.
Results: In our experiments, the best recognition rates of pain versus nonpain (88.00%), pain versus rest (94.62%), pain versus cry (80.00%), pain versus air puff (83.33%), and pain versus friction (93.00%) were obtained from an SVM with a polynomial kernel of degree 3. The SVM outperformed two commonly used methods in face classification: PCA and LDA, each using the L_1 distance metric.
Conclusion: The results of this study indicate that the application of face classification techniques in pain assessment and management is a promising area of investigation.

* Corresponding author. Tel.: +1 417 836 4932; fax: +1 417 836 6907.
E-mail address: shb757f@smsu.edu (S. Brahnam).

0933-3657/$ – see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.artmed.2004.12.003

1. Introduction

The assessment of pain in newborns is considered one of the most challenging problems in neonatology [1]. Pain assessment is difficult because neonates cannot articulate their pain experiences and vary in their responses to pain and other stimuli [2,3]. Since pain is a major indicator of medical problems [4] and the quality of patient care depends on the quality of pain management [5], it is vital that methods be developed that accurately distinguish an infant's signals of pain from a host of minor distress signals [3].

At present, assessment of neonate pain takes into consideration a number of physiological and behavioral factors. Among the many physiological indicators of pain are changes in heart and respiratory rates, blood pressure, vagal tone, and palmar sweating [6]. The consensus in the reference literature, however, is that physiological measures are insufficient and unreliable indices of pain. Physiological measures vary significantly from newborn to newborn, and they fail to reflect the intensity of pain [4]. Moreover, the physiological parameters associated with pain are not easily distinguishable from parameters associated with fear and anxiety [7].

Significant neonate behavioral responses to pain include body movement, crying, and facial expressions [4]. Facial expressions play a central role in pain assessment, as attested by the fact that most pain instruments developed for infants, toddlers, and older children, including COMFORT [8], CRIES (crying, oxygen requirement, increased vital signs, expression, and sleeplessness) [9], FLACC (face, legs, activity, cry, consolability) [10], MIPS (modified infant pain scale) [11], and CFACS (child facial coding system) [12], rely in whole or in part on facial displays. The facial characteristics associated with pain in infants include prominent forehead, narrowed eyes, deepening of the nose–lip furrow, and an angular opening of the mouth [13]. Facial expressions are a critical factor in the assessment of infant pain because they are the most specific and frequent indicators of pain [14]. Body movement and crying are behaviors that are associated with other states, such as hunger, fright, and discomfort; and neonates do not always respond to pain by crying and moving. Sleep, for instance, inhibits bodily movement; yet, in sleep, an infant's face will often register the experience of pain [15]. This is an issue that is particularly relevant to neonates since they spend between 14 and 17 hours a day sleeping [16].

Even though the facial characteristics of infant expressions of pain have been studied extensively [13], there are serious problems with pain assessment instruments that utilize facial displays. The primary problem is that these tools rely on the observations of health professionals, and health professionals have been shown to be biased in their observations and less competent than nonprofessionals in recognizing facial expressions of pain [17,18]. Xavier Balda et al. [17] theorize that health professionals become desensitized because of their constant exposure to suffering. The findings of Xavier Balda et al. [17] corroborate other studies demonstrating that the greater the clinical experience of the health professional the more likely he or she is to underestimate patient pain [18]. A repeated refrain in the reference literature, therefore, is that neonate pain assessment tools need to be developed that alleviate or circumvent the problem of observer desensitization and bias [17,19].

The objective of this study is to tackle these problems by applying state-of-the-art face classification techniques to the task of distinguishing a newborn's facial expressions of pain from facial expressions that are similar but not triggered by pain. Since the assessment of pain by machine is based on pixel states, the development of a machine classification system of pain will offer the following advantages: it will remain objective, it will exploit the full spectrum of information available in a neonate's facial expressions, and it will not degrade over time. A machine classification system of pain will offer the additional benefit of monitoring a neonate's facial expressions when the patient is left unattended.

As described more fully in Section 2, the face classification techniques used in this study are principal component analysis (PCA), linear discriminant analysis (LDA), and support vector machines (SVM). Although these techniques have succeeded in classifying faces according to identity [20,21], gender [22,23], age [24], race [25], and emotions [26,27], they have yet to be applied to medical problems that involve the face. Dai et al. [28] have proposed a method for observing the facial expressions of patients in hospital beds, but their facial images were not of actual patients but rather of subjects responding to verbal cues suggestive of medical procedures and conditions. To date, no work has employed face classification techniques to the task of classifying actual facial expressions of pain.

In this study, PCA, LDA, and SVM are trained and tested using facial photographs of neonates experiencing four noxious stimuli: transport from one crib to another, air puff on the nose, friction from cotton and alcohol rubbed on the lateral surface of the heel, and the puncture of a heel lance. The three classifiers are reviewed in Section 2, and the experi-

mental design is described more completely in Section 3. Section 4 discusses the methods and the procedures used in the classification experiments, and Section 5 presents the experimental results. The paper is concluded, in Section 6, by noting some of the contributions and limitations of this study and by offering directions for future research.

2. Basic concepts of three face classification techniques: PCA, LDA, and SVM

In this paper, three types of classifiers are used to train and to test infant facial expressions: PCA, LDA, and SVM. The basic concepts behind these classifiers are presented in this section.

2.1. PCA

The central idea behind PCA is to find an orthonormal set of axes pointing in the direction of maximum covariance in the data. In terms of facial images, the idea is to find the orthonormal basis vectors, or the eigenvectors, of the covariance matrix of a set of images, with each image treated as a single point in a high dimensional space. It is assumed that the facial images form a connected subregion in the image space. The eigenvectors map the most significant variations between faces and are preferred over other correlation techniques that assume every pixel in an image is of equal importance (see, for instance [29]). Since each image contributes to each of the eigenvectors, the eigenvectors resemble ghostlike faces when displayed. For this reason, they are oftentimes referred to in the literature as *holons* [30] or *eigenfaces* [20], and the new coordinate system is referred to as the *face space* [20]. Examples of eigenfaces are shown in Fig. 1. Individual images can be projected onto the face space and represented exactly as weighted combinations of the eigenface components (see Fig. 2).

The resulting vector of weights that describe each face can be used both in face classification and in data compression. Classification is performed by projecting a new image onto the face space and comparing the resulting weight vector to the weight vectors of a given class [20]. Compression is achieved by reconstructing images using only those few eigenfaces that account for the most variability [31]. PCA classification and compression are discussed in more detail below.

2.1.1. PCA classification

The principal components of a set of images can be derived directly as follows. Let $\mathbf{I}(x, y)$ be a two-dimensional array of intensity values of size $N \times N$. $\mathbf{I}(x, y)$ may also be represented as a single point, a one-dimensional vector $\mathbf{\Gamma}$ of size N^2. Let the set of face images be $\mathbf{\Gamma}_1, \mathbf{\Gamma}_2, \ldots, \mathbf{\Gamma}_M$. Let

$$\mathbf{\Phi}_k = \mathbf{\Gamma}_k - \mathbf{\Psi} \tag{1}$$

represent the mean normalized column vector for a given face $\mathbf{\Gamma}_k$, where

$$\mathbf{\Psi} = \frac{1}{M}\sum_{k=1}^{M}\mathbf{\Gamma}_k \tag{2}$$

is the average face of the set.

PCA seeks the set of M orthonormal vectors, $\mathbf{u}_k$, and their associated eigenvalues, $\boldsymbol{\lambda}_k$, which best describes the distribution of the image points. The vectors $\mathbf{u}_k$ and scalars $\boldsymbol{\lambda}_k$ are the eigenvectors and eigenvalues, respectively, of the covariance matrix:

$$\mathbf{C} = \frac{1}{M}\sum_{k=1}^{M}\mathbf{\Phi}_k\mathbf{\Phi}_k^{\mathrm{T}} = \mathbf{A}\mathbf{A}^{\mathrm{T}} \tag{3}$$

where the matrix $\mathbf{A} = [\mathbf{\Phi}_1, \mathbf{\Phi}_2, \ldots, \mathbf{\Phi}_M]$ [20].

The size of $\mathbf{C}$ is N^2 by N^2, which for typical image sizes is an intractable task [20]. However, since typically $M < N^2$, that is, the number of images is less than the dimension, there will only be $N - 1$ nonzero eigenvectors. Thus, the N^2 eigenvectors can

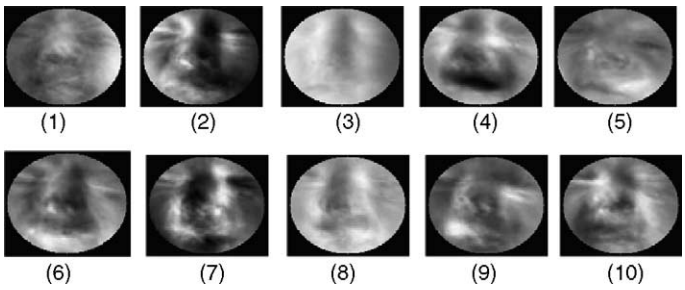

Figure 1 The first 10 eigenfaces of the 200 neonate images, with the eigenfaces ordered by magnitude of the corresponding eigenvalue.

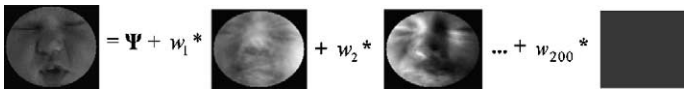

Figure 2 Illustration of the linear combination of eigenfaces. The face to the left can be represented as a weighted combination of eigenfaces plus (Ψ), the average face (see Equation (2)).

be solved, in this case, by first solving for the eigenvectors of an $M \times M$ matrix, followed by taking the appropriate linear combinations of the data points Φ (see [20]).

PCA is closely associated with the singular value decomposition of a data matrix and can be decomposed as:

$$\Phi = \mathbf{USV}^{\mathsf{T}} \tag{4}$$

where **S** is a diagonal matrix whose diagonal elements are the singular values, or eigenvalues, of Φ, and **U** and **V** are unary matrices. The columns of **U** are the eigenvectors of $\Phi\Phi^{\mathsf{T}}$, and are referred to as *eigenfaces*. The columns of **V** are the eigenvectors $\Phi^{\mathsf{T}}\Phi$ and are not used in this analysis.

Faces can be classified by projecting a new face Γ onto the face space as follows:

$$\omega_k = \mathbf{u}_k^{\mathsf{T}}(\Gamma_k - \Psi) \tag{5}$$

for $k = 1, \ldots, M'$ eigenvectors, with $M' \ll M$, if reduced dimensionality is desired. The weights form a vector $\Omega_k^{\mathsf{T}} = [\omega_1, \omega_2, \ldots \omega_{M'}]$, which contains the projections onto each eigenvector. Classification is performed by calculating the distance of Ω_k from Ω, where Ω represents the average weight vector defining some class [20].

Two commonly used distance measures are the sum of absolute differences, also known as the L_1 metric and the Euclidean distance, also known as the L_2 metric. If we have two points, $A(x_1, y_1)$ and $B(x_2, y_2)$, the L_1 distance between A and B is $abs(x_1 - x_2) + abs\ (y_1 - y_2)$. The L_2 metric is $\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$.

2.1.2. PCA data compression

Since the eigenfaces are ordered, with each one accounting for a different amount of variation among the faces, images can be reconstructed using only those few eigenfaces, $M' \ll M$ in Equation (4), that account for the most variability [31]. Because PCA results in a dramatic reduction of dimensionality and maps the most significant variations in a dataset, it is typically used to represent faces when performing other classification procedures.

2.2. LDA

While PCA is optimal for reconstructing images from a low dimensional space, it is not optimal for discrimination. PCA yields projection directions that maximize the total scatter across all classes. LDA, or Fisher's linear discriminants, in contrast, is a supervised learning procedure that projects the images onto a subspace that maximizes the between-class scatter and minimizes the within-class scatter of the projected data. A classical technique in pattern recognition, LDA is an example of a *class specific method* in that it shapes the scatter in order to make it more reliable for classification [21]. There has been a tendency to prefer LDA to PCA because LDA deals directly with discrimination between classes, whereas PCA aims at faithfully representing the data. It has been shown that LDA outperforms PCA only when large and representative training data sets are given [32].

2.3. SVM

SVM, introduced by Vapnik [33], is a learning system that separates a set of input pattern vectors into two classes with an optimal separating hyperplane. The set of vectors is said to be optimally separated by the hyperplane if it is separated without error and the distance between the closest vectors to the hyperplane is maximal. SVM produces the pattern classifier by applying a variety of kernel functions (linear, polynomial, radial basis function, and so on) as the possible sets of approximating functions, by optimizing the dual quadratic programming problem, and by using structural risk minimization as the inductive principle, as opposed to classical statistical algorithms that maximize the absolute value of an error or of an error squared.

Originally, SVM was designed to handle dichotomic classes. Recently, research has concentrated on expanding two-class classification to multiclass classification. Since the objective in this paper has been to distinguish neonate facial displays of pain from other facial displays, this discussion of SVM will be limited to dichotomic classification.

Different types of SVM classifiers are used depending upon the type of input patterns: a linear maximal margin classifier is used for linearly separable data, a linear soft margin classifier is used for linearly nonseparable, or overlapping, classes, and a nonlinear classifier is used for classes that are overlapped as well as separated by nonlinear hyperplanes. All three classifiers are discussed in more detail below. It should be noted, however, that the

linearly separable case is rare in real world problems and was not explored in our experiments.

2.3.1. Linear maximal margin classifier

The case where the training patterns can be linearly separated by a hyperplane, $\mathbf{w}\cdot\mathbf{x} + b = 0$, is the simplest case and provides a good foundation for the other two cases. The purpose of the SVM is to find the optimal values for $\mathbf{w}$ (e.g., $\mathbf{w}_0$) and b (e.g., b_0). After finding the optimal separating hyperplane, $\mathbf{w}_0\cdot\mathbf{x} + b_0 = 0$, an unseen pattern, $\mathbf{x}_t$, can be classified by the decision rule $f(x) = \text{sign}(\mathbf{w}_0\cdot\mathbf{x}_t + b_0)$, as shown below.

Suppose, there is a set of training data, $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k$, where $\mathbf{x}_i \in \mathbf{R}^n$ and $i = 1, 2, \ldots, k$. Each $\mathbf{x}_i$, belonging as it does to one of two classes, has a corresponding value y_i, where $y_i \in \{-1, 1\}$. The goal in this case is to build the hyperplane that maximizes the minimum distance between the two classes. Because the hyperplane is $\mathbf{w}\cdot\mathbf{x} + b = 0$, the training data can be divided into two classes such that

$$\begin{aligned} &\mathbf{w}\cdot\mathbf{x}_i + b \geq 1, \quad \text{if } y_i = 1, \\ &\mathbf{w}\cdot\mathbf{x}_i + b \leq -1, \quad \text{if } y_i = -1 \end{aligned} \tag{6}$$

where $\mathbf{w} \in \mathbf{R}^n$ and $b \in \mathbf{R}$.

Combining the equations in (6), we obtained the following:

$$y_i(\mathbf{w}\cdot\mathbf{x}_i + b) \geq 1 \quad \forall \mathbf{x}_i, \quad i = 1, 2, \ldots, k. \tag{7}$$

The distance between a point $\mathbf{x}$ and the hyperplane is $d(\mathbf{w}, b; \mathbf{x}) = |\mathbf{w}\cdot\mathbf{x} + b|/||\mathbf{w}||$.

According to Equation (6), the minimum distance between one of the two classes and the hyperplane is $\frac{1}{\|\mathbf{w}\|}$. The margin, M, which is the distance between the two classes, is $\frac{2}{\|\mathbf{w}\|}$.

Finding the optimal separating hyperplane having a maximal margin requires that the following minimization problem be solved:

$$\begin{aligned} &\text{Minimize :} && \frac{1}{2}\mathbf{w}\cdot\mathbf{w}, \\ &\text{Subject to :} && y_i(\mathbf{w}\cdot\mathbf{x}_i + b) \geq 1 \quad \forall \mathbf{x}_i, \quad i = 1, 2, \ldots, k. \end{aligned}$$

This nonlinear optimization problem with inequality constraints can be solved by the *saddle point* of the Lagrange function:

$$\mathbf{L}(\mathbf{w}, b, \alpha) = \frac{1}{2}\mathbf{w}\cdot\mathbf{w} - \sum_{i=1}^{K}\alpha_i(y_i(\mathbf{w}\cdot\mathbf{x}_i + b) - 1) \tag{8}$$

where $\alpha_i \geq 0$ are the Lagrange multipliers.

By minimizing the Lagrange function with respect to $\mathbf{w}$ and b, as well as by maximizing with respect to α_i, the minimization problem above can be transformed to its dual problem, called the quadratic programming problem:

$$\frac{\partial L(\mathbf{w}, b, \alpha_i)}{\partial \mathbf{w}}|\mathbf{w} = \mathbf{w}_0 = \left(\mathbf{w}_0 - \sum_{i=1}^{K}\alpha_i y_i \mathbf{x}_i\right) = 0, \tag{9}$$

$$\frac{\partial L(\mathbf{w}, b, \alpha_i)}{\partial b}|b = b_0 = \sum_{i=1}^{K} y_i\alpha_i = 0. \tag{10}$$

Equation (11) can be obtained by plugging (9) and (10) into (8):

$$\mathbf{L}(\alpha) = \sum_{i=1}^{K}\alpha_i - \frac{1}{2}\sum_{i=1}^{K}\sum_{j=1}^{K}\alpha_i\alpha_j y_i y_j \mathbf{x}_i\mathbf{x}_j \tag{11}$$

The dual problem can be described as follows:

$$\begin{aligned} &\text{Maximize :} && \sum_{i=1}^{K}\alpha_i - \frac{1}{2}\sum_{i=1}^{K}\sum_{j=1}^{K}\alpha_i\alpha_j y_i y_j \mathbf{x}_i\mathbf{x}_j \\ &\text{Subject to :} && \sum_{i=1}^{K} y_i\alpha_i = 0, \qquad \alpha_i \geq 0. \end{aligned}$$

By solving the dual problem, the optimal separating hyperplane is determined by Equations (12) and (13).

$$\mathbf{w}_0 = \sum_{i=1}^{K}\alpha_i y_i \mathbf{x}_i \tag{12}$$

$$b_0 = y_i - \mathbf{w}_0\cdot\mathbf{x}_i \tag{13}$$

where $\mathbf{x}_i$ belongs to support vectors, $y_i \in \{-1, 1\}$.

The unseen test data $\mathbf{x}_t$ can be classified, therefore, by simply computing Equation (14):

$$f(x) = \text{sign}(\mathbf{w}_0\cdot\mathbf{x}_t + b_0) \tag{14}$$

By examining Equation (12), it can be seen that the hyperplane is determined by all the training data, $\mathbf{x}_i$, that have the corresponding attributes of $\alpha_i > 0$. We call this kind of training data *support vectors*. Thus, the optimal separating hyperplane is not determined by the training data per se but rather by the support vectors.

2.3.2. Linear soft margin classifier

As mentioned above, patterns that are linearly separable are rare in real world problems. In this section, we expand SVM to handle input patterns that are overlapping, or linearly nonseparable. In this case, our objective is to separate the two classes of training data with a minimal number of errors. To accomplish this we introduce some nonnegative slack variables ξ_i, $i = 1, 2, \ldots, k$ to the system. Thus, Equations (6) and (7), in the linearly separable case above, can be rewritten as Equations (15) and (16):

$$\begin{aligned} &\mathbf{w}\cdot\mathbf{x}_i + b \geq 1 - \xi_i, \quad \text{if } y_i = 1, \\ &\mathbf{w}\cdot\mathbf{x}_i + b \leq -1 + \xi_i, \quad \text{if } y_i = -1, \end{aligned} \tag{15}$$

$$y_i(\mathbf{w}\cdot\mathbf{x}_i + b) \geq 1 - \xi_i, \quad i = 1, 2, \ldots, k \tag{16}$$

Just as we obtained the optimal separating hyperplane in the linearly separable case, obtaining the soft margin hyperplane in the linear nonsepar-

able case requires that the following minimization problem be solved:

$$\text{Minimize}: \quad \frac{1}{2}\mathbf{w}\cdot\mathbf{w} + C\left(\sum_{i=1}^{K}\xi_i\right)$$

$$\text{Subject to}: \quad y_i(\mathbf{w}\cdot\mathbf{x}_i + b) \geq 1 - \xi_i, \quad i = 1, 2, \ldots, k.$$
$$\xi_i \geq 0, \qquad i = 1, 2, \ldots, k$$

where C is a penalty or regularization parameter.

By minimizing the Lagrange function with respect to $\mathbf{w}$, b, and ξ_i, as well as by maximizing with respect to α_i, the minimization problem above can be transformed to its dual problem, described as follows:

$$\text{Maximize}: \quad \sum_{i=1}^{K}\alpha_i - \frac{1}{2}\sum_{i=1}^{K}\sum_{j=1}^{K}\alpha_i\alpha_j y_i y_j \mathbf{x}_i\mathbf{x}_j$$

$$\text{Subject to}: \quad \sum_{i=1}^{K} y_i\alpha_i = 0, \quad 0 \leq \alpha_i \leq C$$

Solving the dual problem, the soft margin hyperplane is determined by Equations (17) and (18):

$$\mathbf{w}_0 = \sum_{i=1}^{K}\alpha_i y_i \mathbf{x}_i \tag{17}$$

$$b_0 = y_i - \mathbf{w}_0\cdot\mathbf{x}_i \tag{18}$$

where $\mathbf{x}_i$ belongs to margin vectors, $y_i \in \{-1, 1\}$.

By examining Equation (17), it can be seen that the hyperplane is determined by all the training data, $\mathbf{x}_i$ that have the corresponding attributes of $\alpha_i > 0$. These support vectors can be divided into two categories. The first category has the attribute of $\alpha_i < C$. In this category, $\xi_i = 0$, and these support vectors lie at the distance $\frac{1}{\|\mathbf{w}\|}$ from the optimal separating hyperplane. We called these support vectors *margin vectors*. The second category has the attributes of $\alpha_i = C$. In this category, the support vectors are correctly classified with either a distance smaller than $\frac{1}{\|\mathbf{w}\|}$ from the optimal separating hyperplane (if $0 < \xi_i \leq 1$) or they are misclassified (if $\xi_i > 1$). The support vectors in the second category are regarded as errors.

2.3.3. Nonlinear classifier

Sometimes, the input vectors cannot be linearly separated in the input space. In this case, kernel functions, such as the polynomial or radical basis function (RBF), are used to transform the input space to a feature space of higher dimensionality. In the feature space, a linear separating hyperplane is sought that separates the input vectors into two classes.

If $\mathbf{x} \in \mathbf{R}^n$ is in the input space, we can map the input vector $\mathbf{x}$ from the n-dimension input space to a corresponding N-dimensions feature space through a function, ϕ. After the transformation, we know $\phi(\mathbf{x}) \in \mathbf{R}^N$. Following the steps described in the case of linearly separable training patterns (Section 2.3.1) and the case of linearly nonseparable training patterns (Section 2.3.2), the hyperplane and decision rule for nonlinear training patterns can be established.

In a manner similar to obtaining the hyperplane for the linearly separable training patterns in Equations (12) and (13) and the hyperplane for the linearly nonseparable training patterns in Equations (17) and (18), we can obtain the hyperplane for the nonlinear training pattern as in Equation (19):

$$\mathbf{w}_0\cdot\phi(\mathbf{x}) + b_0 = \sum_{i=1}^{K}\alpha_i y_i \phi(\mathbf{x}_i)\cdot\phi(\mathbf{x}) + b_0 \tag{19}$$

In Equation (19), we see that the original dot products of input variables can be replaced by a function, ϕ. That is, kernel function $K(\mathbf{x}_i, \mathbf{x}) = \phi(\mathbf{x}_i)\cdot\phi(\mathbf{x})$. The decision rule for nonlinear training patterns can be established as shown in Equation (20):

$$f(x) = \text{sign}\left(\sum_{i=1}^{K}\alpha_i y_i K(\mathbf{x}_t, \mathbf{x}) + b\right) \tag{20}$$

where $K(\mathbf{x}_t, \mathbf{x})$ is a kernel function.

3. Study design

The most important consideration in the design of this study was the choice of stimuli. Warnock and Sandrin [3] have recently stressed the importance of including a variety of contrasting stimuli in studies on infant pain assessment. Early research focused mostly on neonate responses to two stimuli: a pain inducing stimulus (pin prick or puncture of a lancet) and friction on the heel [34,35]. Contemporary studies tend to include additional stressors, such as exposure to bright light [17] and diaper change [3].

This study follows contemporary research by including two stressors in addition to the puncture of a lancet. Since PCA, LDA, and SVM classifiers are sensitive to light changes, an air puff stimulus was introduced in lieu of a bright light stimulus. Exposure to a puff of air on the face is similar to exposure to bright light in that it causes the eyes to squeeze tightly together, producing a facial expression that is similar, yet distinct, from the facial expression of pain. At the suggestion of hospital personnel, infants were transported from one crib to another before each photography session. This change was welcomed as it supplied a stressor, like diaper change that sometimes provokes crying. We were thus afforded the opportu-

nity of contrasting classifier recognition rates of neonate crying expressions that were in response to pain to those crying expressions that were in response to a less noxious stimulus.

Thus, four noxious stimuli are included in this study: (1) the puncture of a heel lance, (2) friction on the external lateral surface of the heel, (3) transport from one crib to another, and (4) an air stimulus.

3.1. Subjects

This study complied with the protocols and ethical directives for research involving human subjects at 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 neonate unit of a St. John's Hospital sometime after delivery. Only mothers who had experienced uncomplicated deliveries were approached.

A total of 200 color photographs were taken of 26 Caucasian neonates (13 boys and 13 girls) ranging in age from 18 h to 3 days old. Six males had been circumcised the day before the photographs were taken, and the last feeding time before the photography session ranged from 45 min to 5 h. All infants were in good health.

3.2. Apparatus

All photographs were taken using a Nikon D100 digital camera under ambient lighting conditions in a room separated from other newborns.

3.3. Procedure

The facial expressions of the newborns were photographed in one session while the infants were experiencing four distinct stimuli: transport from one crib to another, air puff on the nose, friction from cotton and alcohol rubbed on the heel, and the puncture of a heel lance. The state of the infant after being transported to another crib was further evaluated at the time the photographs were taken into one of two states: crying or resting.

All stimuli were administered by an attending nurse. Following the example of [17] and the requirements of standard medical procedures, photographs of the facial expressions of the four stimuli were taken in the following sequence:

1. *Transport from one crib to another (rest/cry)*: After being transported from one crib to another, the neonate was swaddled and a series of photographs was taken over the course of 1 min. The state of the neonate was noted as either crying or resting for each photograph taken in the series.
2. *Air stimulus*: After resting for at least one additional minute, the neonate's nose was exposed to a puff of air emitted from a squeezable plastic camera lens cleaner. A series of pictures of the neonate's face was taken immediately after the air puff contacted the infant's face.
3. *Friction*: After resting for at least 1 min, the neonate received friction on the external lateral surface of the heel with cotton wool soaked in 70% alcohol for 10—15 s. The face of the neonate was repeatedly photographed during the friction rubbings.
4. *Pain*: After resting for at least 1 min, the external lateral surface of the heel was punctured for blood collection. Several continuous photographs of the neonate's face were taken starting immediately after introduction of the lancet and while the skin of the heel was squeezed for blood samples.

Of the 200 facial photographs, 63 are rest, 18 cry, 23 air stimulus, 36 friction, and 60 are pain. Fig. 3 provides two example sets, with backgrounds removed, of the five neonate facial expressions of rest, cry, air puff, friction, and pain.

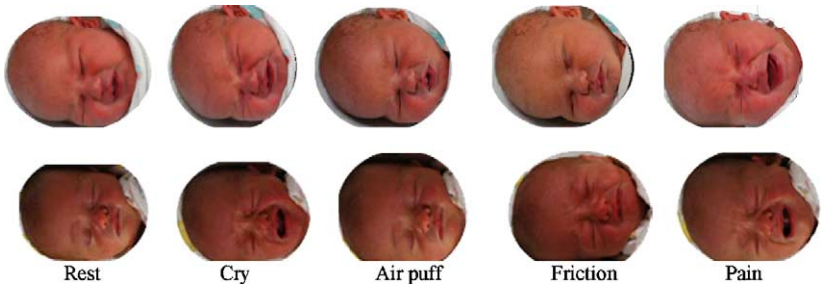

Figure 3 Examples of the five facial expressions in the dataset.

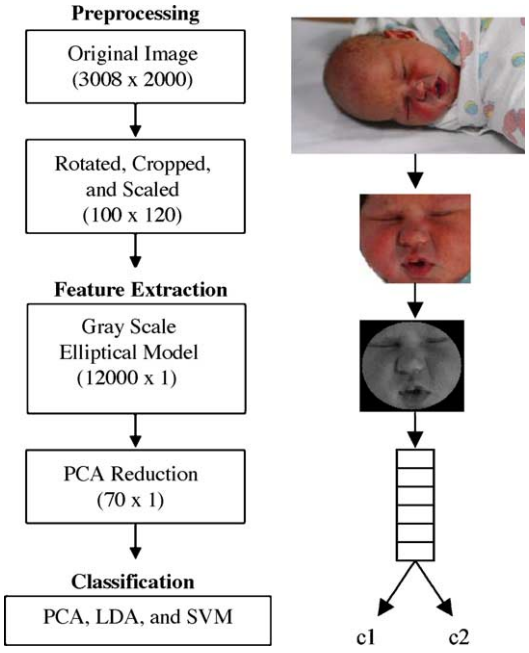

Figure 4 The experimental procedure.

4. Experimental procedures

As illustrated in Fig. 4, the experimental procedures can be divided into the following stages: preprocessing, feature extraction, and classification.

In the preprocessing stage, the original images are manually rotated and scaled using Adobe Photoshop 7 [36] such that the eyes lie roughly along the same axis. They are then cropped automatically by computer using MATLAB [37]. The original 200 images, size 3008 pixels $\times$ 2000 pixels, are also reduced to 100 pixels $\times$ 120 pixels.

In the feature extraction stage, facial features are centered within an ellipse and color information is discarded. The rows within the ellipse are concatenated to form a feature vector of dimension 12,000 with entries ranging in value between 0 and 255. PCA is used to reduce the dimensionality of the feature vectors further. In our experiments, the first 70 principle components resulted in the best classification scores (in Section 5.2, we contrast classification results and processing times, in the case of SVM, using the 70 principal components as inputs versus using the 12,000 raw inputs).

Finally, in the classification stage, PCA, LDA, and SVM are used to classify the feature vectors into the following category pairs: pain/nonpain, pain/cry, pain/air puff, and pain/friction. All experiments were processed in the MATLAB environment under Windows XP operating system using a Pentium 4—2.80 GHz processor. In addition, for SVM, we used the OSU SVM Classifier Matlab Toolbox developed by Ohio State University.

5. Experimental results

This section describes two face classification experiments. In experiment 1, PCA, LDA, and SVM classify faces into the following classification pairs: (a) pain/rest, (b) pain/cry, (c) pain/air puff, and (d) pain/friction. In experiment 2, PCA, LDA, and SVM classify faces into the classification pair of pain/nonpain. The set of nonpain images was obtained by combining the rest, cry, air puff, and friction images into one category of 140 images. The remaining 60 images were of pain.

Because the number of images in the dataset is small, a cross-validation technique was applied in all experiments. The cross-validation technique we performed was a four step process. In step 1, the images were randomly divided, in terms of the set of facial expressions being examined in the two experiments, into 10 segments. In step 2, 9 out of the 10 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 10 times. Finally, in step 4, the 10 classification scores were averaged to obtain a final performance score.

As an example of this process for experiment 2, suppose there are 60 pain images and 100 nonpain images. In step 1, we would randomly choose 54 pain and 90 nonpain images as the training set and let the remaining 6 pain and 10 nonpain images become the testing set. In step 2, we would train the classifiers and then obtain the recognition rate for the testing set. For example, if 2 images out of the 16 images in the testing set were wrongly classified, then the classification score for this run would be $(16 - 2)/16 = 87.5\%$. In step 3, we would repeat steps 1 and 2 ten times. In step 4, we would average the 10 classification scores for the final performance score.

The regularization parameter, C, for the SVM classifiers was determined using a grid search. First we set C to a default value of 1 and incremented it by some fixed value. We then computed the performance of each regularization parameter C ($C = 1, 5, 11, 16, 21, \ldots$). Finally, we determined the regularization parameter C in terms of recognition rates. 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 SVMs using RBF kernels was also optimized using a grid search.

Table 1 Experiment 1: classification rates of pain vs. rest

Type of SVM	Recognition rate (%)
Linear	90.77
Polynomial with degree = 2	84.62
Polynomial with degree = 3	94.62
Polynomial with degree = 4	86.15
RBF kernel	53.85
PCA with L1 distance	87.18
LDA with L1 distance	93.08

5.1. Experiment 1

Tables 1—4 compare the classification scores of PCA (using the L_1 metric), LDA (using the L_1 metric) and SVM (using linear, polynomials of degree 2, 3, and 4, and RBF kernel functions) for each of the following expression pairs: pain/rest, pain/cry, pain/air puff, and pain/friction. Referring to Tables 1 and 2, the best classification score of pain versus rest is 94.62% and pain versus cry is 80.00%, using an SVM with polynomial kernel of degree 3. In Table 3, the best classification score of pain versus air puff is 90.00% using a linear SVM. In Table 4, most SVM systems separate pain from friction. The best result is 96.00% using an SVM with polynomial kernel of degree 2.

Tables 1—4 demonstrate that different SVM systems have distinct effects on recognizing pain from other facial expressions. For example, a linear SVM has a stable recognition rate of 90.00% in pain versus all facial expressions except cry. In general, however, an SVM with polynomial kernel of degree 3 has the best overall classification performance.

5.2. Experiment 2

In experiment 2, PCA, LDA, SVM are used to classify faces into two categories: pain and nonpain. The nonpain set consisted of all the air puff, cry, friction, and rest images. Table 5 shows the results of the PCA, LDA and several SVM systems. The best classification score (88.00%) is obtained using SVM with polynomial kernel of degree 3. Recall that SVM with polynomial kernel of degree 3 provided the best overall classification score in experiment 1 as well. SVM with polynomial kernel of degree 3, therefore, is probably the best selection for classifying neonate facial expressions of pain.

Tables 6 and 7 illustrate the difference, in experiment 2, using SVMs with and without PCA preprocessing of inputs. As can be seen in the classification results, SVMs with PCA inputs have two advantages over SVMs with raw inputs. First, SVMs with PCA inputs have a better recognition rate than the SVMs with raw inputs. This is because the raw inputs contain more noise and irrelevant information. This degrades the performance of the SVM. Second, SVMs with PCA inputs have a faster computation time than the SVMs with raw inputs. The former is about 10 times fast than the latter. This is because the raw inputs contain all 12,000 features for each image,

Table 2 Experiment 1: classification rates of pain vs. cry

Type of SVM	Recognition rate (%)
Linear	71.25
Polynomial with degree = 2	78.75
Polynomial with degree = 3	80.00
Polynomial with degree = 4	76.25
RBF Kernel	75.00
PCA with L1 distance	68.75
LDA with L1 distance	70.42

Table 3 Experiment 1: classification rates of pain vs. air puff

Type of SVM	Recognition rate (%)
Linear	90.00
Polynomial with degree = 2	77.78
Polynomial with degree = 3	83.33
Polynomial with degree = 4	78.89
RBF kernel	66.67
PCA with L1 distance	81.48
LDA with L1 distance	89.63

Table 4 Experiment 1: classification rates of pain vs. friction

Type of SVM	Recognition rate (%)
Linear	90.00
Polynomial with degree = 2	96.00
Polynomial with degree = 3	93.00
Polynomial with degree = 4	92.00
RBF kernel	60.00
PCA with L1 distance	74.00
LDA with L1 distance	91.00

Table 5 Experiment 2: classification rates of pain vs. nonpain (nonpain includes air puff, cry, friction, and rest)

Type of SVM	Recognition rate (%)
Linear	83.67
Polynomial with degree = 2	86.50
Polynomial with degree = 3	88.00
Polynomial with degree = 4	82.17
RBF kernel	70.00
PCA with L1 distance	80.33
LDA with L1 distance	83.67

Table 6 Experiment 2: classification rates and processing times using SVM with PCA

SVM with PCA	Recognition rate (%)	Testing per image (s)
Linear	83.67	0.0008
Polynomial with degree = 2	86.50	0.0008
Polynomial with degree = 3	88.00	0.0008
Polynomial with degree = 4	82.17	0.0008
RBF kernel	70.00	0.0008

Table 7 Experiment 2: classification rates and processing time using SVM without PCA

SVM without PCA	Recognition rate (%)	Testing per image (sec)
Linear	84.75	0.00785
Polynomial with degree = 2	84.17	0.00625
Polynomial with degree = 3	85.75	0.00780
Polynomial with degree = 4	84.50	0.00910
RBF kernel	70.00	0.01657

All experiments were processed in the MATLAB environment under Windows XP operating system using a Pentium 4—2.80 GHz processor.

whereas the PCA inputs only contain 70 features for each image.

6. Conclusions

In this paper, three face classification techniques, PCA, LDA, and SVM, were applied to the classification problem of distinguishing neonate facial expressions of pain. The facial expressions of 26 neonates experiencing the puncture of a heel lance, transport from one crib to another, air stimulus to the nose, and friction on the external lateral surface of the heel were photographed. The state of the infant after being transported from one crib to another was further noted as being in one of two states: resting or crying. A series of experiments compared the recognition rates of PCA, LDA, and SVM classifying the following pairs: pain/nonpain, pain/rest, pain/cry, pain/air puff, and pain/friction. We concluded that SVM with a polynomial kernel of degree 3 produced the best overall recognition rates of pain versus nonpain (88.00%), pain versus cry (80.00%), pain versus rest (94.62%), pain versus air puff (83.33%), and pain versus friction (93.00%).

We believe this study makes a number of contributions. It is one of the first attempts at applying state-of-the-art face recognition technologies to actual medical problems. As noted in the introduction, medical applications of standard face recognition technologies have been suggested [28] but not tried with actual medical data. The results of this study are promising and suggest that face recognition technologies could prove useful in neonate pain assessment.

Moreover, even though machine classification of emotion has long been an area of active investigation, we are unaware of research that includes the machine classification of pain experiences. Our research not only addresses pain, but the dataset in this study also includes facial expressions in response to several stressors that result in expressions that are similar to the facial displays of pain. Infants typically respond to pain by crying, for instance, but they also cry in reaction to a number of minor disturbances; this study has included in the dataset expressions of crying that were not triggered by pain experiences.

Finally, this study is one of the first to investigate machine classification of neonate facial displays. Most work in facial classification has focused on adult faces. Rarely have the faces of children been included in these studies, and certainly not the faces of infants.

There are a number of limitations in the current study that also need to be addressed. First, the dataset used in this study is small and needs to be expanded. Second, only reactions to acute pain experiences were included in the dataset. This study does not address chronic pain—pain experiences that are thought to have long-term psychological and neurological consequences [4]. Third, because this study uses photographs, it does not take into account the dynamic nature of facial expressions. It is possible that temporal changes in expressions include significant information regarding a neonate's state. Fourth, this study does not compare human assessment of neonate pain with machine assessment, nor does it speculate on the practicality of implementing these technologies within a hospital setting.

In terms of future research possibilities, we are currently designing a study that will compare human recognition rates of pain with machine classification rates, and in future studies we plan on investigating the dynamic nature of facial displays. Another research direction would be to combine machine recognition of *physiological* indices with machine recognition of facial expressions. Lindh et al. [38], for instance, have had some success classifying pain as it relates to heart rate variability using PCA. Machine recognition of *behavioral* indicators, however, offers the advantage of monitoring neonates without the attachment of sensors. Since there is some evidence that the temporal frequency and intensity information in cry can discriminate pain (see [39]), combining sound classifiers with face recognition is yet another area of potential research. Given the difficulty of distinguishing an individual cry from within the robust population of the typical neonate unit, however, monitoring the facial expressions of an infant, rather than its cry, probably offers the most practical solution in a hospital setting.

Acknowledgements

The authors wish to gratefully acknowledge the partial funding of this project by Missouri State University faculty grant No. 1015-22-2181.

References

[1] Franck LS, Miaskowski C. Measurement of neonatal responses to painful stimuli: a research review. J Pain System Manage 1997;14:343—78.
[2] Fitzgerald M. The developmental neurobiology of pain. In: Bond MR, Charlton AJE, Woolf CJ, editors. Proceedings of VIth World Congress on Pain. Amsterdam: Elsevier Science Publishers; 1991. p. 253—61.
[3] Warnock F, Sandrin D. Comprehensive description of newborn distress behavior in response to acute pain (newborn male circumcision). Pain 2004;107:242—55.
[4] McGrath PA. Pain in children: nature, assessment and treatment. New York: Guildford Press, 1989.
[5] Harrison D, Evans C, Johnston L, Loughnan P. Bedside assessment of heel lance pain in the hospitalized infant. J Obstet Gynecol Neonatal Nurs 2002;31:551—7.
[6] Coffman S, Alvarez Y, Pyngolil M, Petit R, Hall C, Smyth M. Nursing assessment and management of pain in critically ill children. Heart Lung 1997;26:221—8.
[7] Van Cleve L, Johnson L, Pothier P. Pain responses of hospitalised infants and children to venipuncture and intravenous cannulation. J Pediatr Nurs 1996;11:161—8.
[8] Ambuel B, Hamlett KW, Marx CM, Blumer JL. Assessing distress in pediatric intensive care environments: the COMFORT scale. J Pediatr Psychol 1992;17:95—109.
[9] Krechel SW, Bilder J. CRIES: a new neonatal postoperative pain measurement score: initial testing of validity and reliability. Paediatr Anaesth 1995;5:53—61.
[10] Merkel SI, Voepel-Lewis T, Shayevitz JR, Malviya S. The FLACC: a behavioral scale for scoring postoperative pain in young children. Pediatr Nurs 1997;23:293—7.
[11] Buchholz M, Karl HW, Pomietto M, Lynn A. Pain scores in infants: a modified infant pain scale versus visual analogue. J Pain System Manage 1998;15:117—24.
[12] Gilbert CA, Lilley CM, Craig KD, McGrath PJ, Court CA, Bennett SM, et al. Postoperative pain expression in preschool children: validation of the child facial coding system. Clin J Pain 1999;15:192—200.
[13] Grunau RE, Grunau RVE, Craig KD. Pain expression in neonates: facial action and cry. Pain 1987;28:395—410.
[14] Stevens BJ, Johnston C, Petryshen P, Taddio A. Premature infant pain profile: development and initial validation. Clin J Pain 1996;12:13—22.
[15] Merkel S, Voepel-Lewis T, Malviya S. Pain assessment in infants and young children: the FLACC scale. Am J Nurs 2002;102:55—8.
[16] Williams RL, Karacan I, Hursch CJ. Electroencephalography (EEG) of human sleep: clinical applications. New York: Wiley, 1974.
[17] Xavier Balda R, Guinsburg R, Almeida MFBd, Araujo CPd, Miyoshi MH, Kopelman BI. The recognition of facial expression of pain in full-term newborns by parents and health professionals. Arch Pediatr Adolesc Med 2000;154:1009—16.
[18] Prkachin KM, Solomon P, Hwang T, Mercer SR. Does experience influence judgments of pain behaviour? Evidence from relatives of pain patients and therapists. Pain Res Manage 2001;6:105—12.
[19] Hultgren MS. Assessment of postoperative pain in critically ill infants. Prog Cardiovasc Nurs 1990;5:104—12.
[20] Turk MA, Pentland AP. Eigenfaces for recognition. J Cog Neurosci 1991;3:71—86.
[21] Belhumeur P, Hespanha J, Kriegman D. Eigenfaces vs. fisherfaces: recognition using class specific linear projection. IEEE Trans Pattern Analysis Mach Intell 1997;19:711—20.
[22] O'Toole AJ, Deffenbacher KA. The perception of face gender: the role of stimulus structure in recognition and classification. Mem Cognit 1997;26:146—60.
[23] Moghaddam B, Yang M-H. Learning gender with support faces. IEEE Trans Pattern Analysis Mach Intell 2002;24:306—11.
[24] Valentin D, Abdi H, O'Toole AJ, Cottrell GW. Connectionist models of face processing: a survey. Pattern Recog 1994;27:1209—30.
[25] O'Toole AJ, Abdi H, Deffenbacher KA, Bartlett JC. Classifying faces by race and sex using an autoassociative memory trained for recognition. In: Hammond KJ, editor. Proceedings of 13th Annual Conference on Cognitive Science. Hillsdale, NJ: Lawrence Erlbaum Associates; 1991. p. 847—51.
[26] Zhang Z, Lyons M, Schuster M, Akamatsu S. Comparison between geometry-based and Gabor-wavelets-based facial expression recognition using multi-layer perceptron. In: Yachida M, editor. Proceedings of Third IEEE International Conference on Automatic Face and Gesture Recognition (FG'98). Los Alamitos, CA: IEEE Computer Society Press; 1998. p. 454—9.
[27] Buciu I, Kotropoulos C, Pitas I. ICA and gabor representation for facial expression recognition. In: Horsch A, Lehmann T, editors. Proceedings of IEEE International Conference of Image Processing. Dordrecht: Kluwer Academic Publishers; 2003. p. 855—8.
[28] Dai Y, Shibata Y, Ishii T, Hashimoto K, Katamachi K, Noguchi K, et al. An associate memory model of facial expressions

and its application in facial expression recognition of patients on bed. In: Proceedings of IEEE International Conference on Multimedia and Expo. Los Alamitos, CA: IEEE Computer Society Press; 2001. p. 772—5.
[29] Kosugi M. Human-face search and location in a scene by multi-pyramid architecture for personal identification. Sys Comp Jn 1995;26:27—38.
[30] Cottrell GW, Fleming MK. Face recognition using unsupervised feature extraction. In: Proceedings of International Conference on Neural Networks. Dordrecht: Kluwer Academic Publishers; 1990. p. 322—5.
[31] Sirovich L, Kirby M. Low dimensional procedure for the characterization of human faces. J Opt Soc A 1987;4:519—24.
[32] Martinez AM, Kak AC. PCA versus LDAs. IEEE Trans Pattern Analysis Mach Intell 2001;23:228—33.
[33] Vapnik VN. The nature of statistical learning theory. New York: Springer-Verlag, 1995.
[34] Izard CE, Huebner RR, Risser D, McGinnes GC, Dougherty LM. The young infant's ability to produce discrete emotion expressions. Dev Psychol 1980;16:418—26.
[35] Grunau RVE, Johnston CC, Craig KD. Neonatal facial and cry responses to invasive and non-invasive procedure. Pain 1990;42:295—305.
[36] Adobe Creative Team. Adobe photoshop 7.0 classroom in a book. San Jose, CA: Adobe Systems Incorporated, 2002.
[37] The MathWorks. Using MATLAB: the language of technical computing. Natick, MA: The Mathworks Inc., 2000.
[38] Lindh V, Wiklund U, Håkansson S. Heel lancing in term new-born infants: an evaluation of pain by frequency domain analysis of heart rate variability. Pain 1999;80: 143—8.
[39] Levine JD, Gordon NC. Pain in prelingual children and its evaluation by pain-induced vocalizations. Pain 1982;14: 85—93.