TOWARDS VIDEO LARYNGOSTROBOSCOPY-BASED AUTOMATED SCREENING FOR LARYNGEAL DISORDERS

A. Gelzinis\(^2\), A. Verikas\(^1,2\), M. Bacauskiene\(^2\), E. Vaiciukynas\(^2\), E. Kelertas\(^2\), V. Uloza\(^3\), A. Vegiene\(^3\)

\(^1\)Intelligent Systems Laboratory, Halmstad University, Halmstad, Sweden
\(^2\)Department of Electrical and Control Equipment, Kaunas University of Technology, Kaunas, Lithuania
\(^3\)Department of Otolaryngology, Kaunas University of Medicine, Kaunas, Lithuania

Abstract: This paper is concerned with kernel-based techniques for automated categorization of laryngeal colour image sequences obtained by video laryngostroboscopy. Features used to characterize a laryngeal image are given by the kernel principal components computed using the \(N\)-vector of the 3-D colour histogram. The least squares support vector machine (LS-SVM) is designed for categorizing an image sequence (video) into the healthy, cancerous and noncancerous classes. The kernel function employed by the LS-SVM is defined over a pair of matrices, rather than over a pair of vectors. The classification accuracy of over 85\% was obtained when testing the developed tools on data recorded during routine laryngeal videostroboscopy.

Keywords: Larynx pathology, Image sequence, Classification, Support vector machine

I. INTRODUCTION

Video, laryngeal still images, voice signal, and patient’s questionnaire data are considered as the main information sources to characterize human larynx. Nowadays, automated analysis of voice is increasingly used for detecting and screening laryngeal pathologies [1], [2], [3].

However, there were very few attempts to create systems for automated analysis of still colour laryngeal images. Ilgner et al. [4] proposed a CCD camera-based technique for automated categorization of manually marked suspect lesions into healthy and diseased classes. The categorization is based on textural features extracted from co-occurrence matrices [5], [6] computed from manually marked areas of vocal fold images. The classification accuracy of 81.4\% was reported when testing the technique on a very small set of 35 images. A set of 785 colour laryngeal images obtained by direct microlaryngoscopy has been used in studies presented in [7], [8], [9]. The classification accuracy of over 95\% was achieved when categorizing the images into one healthy and two pathological (nodular and diffuse) classes. When categorizing the same set of images into seven classes (one healthy and six pathological), the classification accuracy of over 80\% was reported [10]. Image texture, distribution of colour, and geometry of edges of vocal folds are the types of features used for the categorization. It was found that colour is amongst the most discriminative types of features.

This study was supported by COST Action 2103 Advanced Voice Function Assessment.

II. THE DATA

A. Video laryngostroboscopy

Video laryngostroboscopy is used extensively for inspecting vocal folds and in the clinical practice for diagnosing voice disorders [11]. Video laryngostroboscopy is a well-established technique for measuring the glottal gap or examining the glottic closure [12]. Videostroboscopy is one of the standard methods used to examine moving objects. Flashing light is used to illuminate an object in stroboscopy. When the flashes are synchronized with the vocal fold vibrations, a stationary view of the vocal folds is obtained.

However, the single-flash-timing video laryngostroboscopy has a limitation that it is effective only when vocal fold vibrations exhibit only one single fundamental frequency. Multiple tones (fundamental frequencies) may be recorded in the case of some diseases, such as polyps, nodules, and cysts [13]. In such cases, a clear view of the vibrating vocal folds can not be obtained with the single-flash-timing video laryngostroboscopy. A multiple-flash-timing technique of video laryngostroboscopy was proposed by Deguchi et al. [13] to deal with such cases.

In [14] image sequences recorded with the stroboscopy system have been used to measure the glottic angle and the angular velocities of vocal fold abduction and adduction. The authors point out that semi-automated edge tracking would be an important improvement of the technique.

It is worth mentioning that not only edge tracking but also other tasks usually carried out when analyzing video data need automated or semi-automated analysis. Decision making is one of such tasks. In clinical practice, decision making is quite often based on subjective evaluation of video data. Quantitative measures of motion, colour distribution and geometry of vocal folds can provide objective information and be useful in medical treatment planning and greatly facilitate tracing progress over time.

The long-term goal of this work is a decision support system to facilitate screening for laryngeal disorders. A voice signal, sequences of colour vocal fold images obtained from video laryngostroboscopy, and questionnaire data [15] are the information sources to be used in the analysis. This paper is concerned with automated categorization of image sequences obtained from laryngeal videostroboscopy into a healthy class and two classes of disorders, namely cancerous and noncancerous.
video laryngostroboscopy into three decision classes, namely a healthy class and two pathological classes—mass lesions of vocal folds. We distinguished two groups of mass lesions of vocal folds i.e. noncancerous lesions—nodules, polyps, papillomatous plaques, and cysts—and cancerous lesions—carcinoma. The diagnosis was confirmed by histological examination of laryngeal specimens removed during endolaryngeal microsurgical intervention. To illustrate the three decision classes, Fig. 1 presents examples of vocal fold images obtained by the direct micro-laryngoscopy.

Fig. 1. Images from the noncancerous (left), cancerous (middle), and healthy (right) classes.

The data have been recorded at the Department of Otolaryngology, Kaunas University of Medicine, Lithuania. The image sequences were acquired during routine videostroboscopy, using the "EndoSTROB" device. The duration of one image sequence was equal to 8 s. The resolution of 720 × 576 pixels was used to record the image sequences. Data from 87 patients were available. Among those, 63 patients belong to the noncancerous class, 18 to the cancerous class and 6 to the healthy class.

III. FEATURES

Various types of features characterizing colour, texture, and geometry of the biological structures seen in colour images of vocal folds can be extracted [8]. Features characterizing the distribution of image colour are used in this study. The approximately uniform distribution of image colour is employed to represent colours. We characterize the colour content of an image by the probability distribution of the colour represented by the 3-D colour histogram of \( N = 4096 \) (16 × 16 × 16) bins and consider the histogram as an \( N \)-vector. Most of bins of the histograms were empty or almost empty. Therefore, to reduce the number of components of the \( N \)-vector, the histograms built from a set of training images were summed up and the \( N \)-vector components corresponding to the bins containing less than \( N_\alpha \) hits in the summed histogram were left aside. Hereby, when using \( N_\alpha = 50 \) we were left with 918 bins—a \( \Phi \) vector of measurements with 918 components.

Having a vector of measurements \( \psi \), the feature vector \( x \) is computed in the following way. We assume that \( x = \kappa(\psi) \) and \( \Phi \) is a mapping of \( \psi \) onto the feature space \( F \), such that \( \kappa(\psi_i, \psi_j) = \langle \Phi(\psi_i), \Phi(\psi_j) \rangle \), where \( \langle \cdot, \cdot \rangle \) stands for the inner product. Let \( \Phi(\psi_i) := \Phi(\psi_i) - \frac{1}{M} \sum_{i=1}^{M} \Phi(\psi_i) \) (1) with \( M \) being the number of data points. The features \( x \) are then given by the kernel principal components computed as projections of \( \Phi(\psi) \) onto the eigenvectors

\[
\mathbf{v} = \sum_{i=1}^{M} \alpha_i \Phi(\psi_i) \tag{2}
\]

of the covariance matrix \( K_{ij} = \langle \Phi(\psi_i), \Phi(\psi_j) \rangle \), where the expansion coefficients \( \alpha_i \) of the eigenvector are found from the eigenvalue problem

\[
\lambda \alpha = K \alpha \tag{3}
\]

where, the solutions \( \alpha \) are normalized by requiring \( \lambda(\alpha, \alpha) = 1 \). Thus, the feature \( x \) is given by

\[
x = \langle \mathbf{v}, \Phi(\psi) \rangle = \sum_{i=1}^{M} \alpha_i \langle \Phi(\psi_i), \Phi(\psi) \rangle \tag{4}
\]

The optimal number of components (features) used is determined experimentally.

IV. THE CLASSIFIER

We use a support vector machine (SVM) as a classifier in this work. Assuming that \( \Upsilon(x) \) is the non-linear mapping of \( x \) into the new space, the 1-norm soft margin SVM can be constructed by solving the following problem:

\[
\min_{w,b,\xi} \frac{1}{2} w^T w + \gamma \sum_{i=1}^{M} \xi_i \tag{5}
\]

subject to

\[
y_i(\langle w, \Upsilon(x_i) \rangle + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad i = 1, ..., M \tag{6}
\]

where \( w \) is the weight vector, \( y_i = \pm 1 \) is the desired output, \( M \) is the number of training data, \( \langle \cdot, \cdot \rangle \) stands for the inner product, \( \xi_i \) are the slack variables, \( b \) is the threshold, and \( \gamma \) is the regularization constant controlling the trade-off between the margin and the slack variables. The discriminant function \( f(x) \) for a new data point \( x \) is given by:

\[
f(x) = \Upsilon \left[ \sum_{i=1}^{M} \alpha_i^* y_i k(x, x_i) + b^* \right], \tag{7}
\]

where \( k(x, x_i) \) stands for the kernel and the Heaviside function \( \Upsilon[y(x)] = -1 \), if \( y(x) \leq 0 \) and \( \Upsilon[y(x)] = 1 \) otherwise. The optimal values \( \alpha_i^* \), \( b^* \) of the parameters \( \alpha_i \) and \( b \) are found during training.

A. Least squares SVM

Suykens and Vandewalle [17] have introduced a least squares version of the SVM classifier (LS-SVM). We use this type of SVM in this work. Parameters of the LS-SVM are estimated by solving the following optimization problem:

\[
\min_{w,b,e} \frac{1}{2} w^T w + \frac{\gamma}{2} \sum_{i=1}^{M} e_i^2 \tag{8}
\]

subject to

\[
y_i(\langle w, \Upsilon(x_i) \rangle + b) = 1 - e_i, \quad i = 1, ..., M \tag{9}
\]
The main difference between the LS-SVM and SVM is the equality constraints (Eq.9) used in the LS-SVM instead of unequally constraints defined by Eq.6. Due to the equality constraints, the optimal parameter values can be found by solving a set of linear equations, instead of quadratic programming applied in the case of SVM. The solution is given by [17]

\[
\begin{bmatrix}
0 \\
y^T \\
Z + \gamma \mathbf{I}
\end{bmatrix}
\begin{bmatrix}
b \\
\alpha
\end{bmatrix}
= \begin{bmatrix}
0 \\
1
\end{bmatrix}
\tag{10}
\]

where \( Z_{ij} = y_i y_j \kappa(x_i, x_j) \), \( I \) is the identity matrix, \( 1 = \{1, \ldots, 1_M\} \), \( y = \{y_1, \ldots, y_M\} \), and \( \alpha = \{\alpha_1, \ldots, \alpha_M\} \).

Since an SVM is a binary classifier while the task is to distinguish between three classes, the one-against-one scheme is used to make the categorization in this work.

B. Kernel function

For \( \kappa(x_i, x_j) \), one usually uses the linear: \( x_i^T x_j \), Gaussian: \( \exp[-||x_i - x_j||^2/\sigma] \) or polynomial: \( (x_i^T x_j + 1)^d \) kernel. The kernel is defined over a pair of vectors.

In this work, classification is based on a set of vectors rather than on a single vector. A sequence of images is recorded from a patient. Each image is represented by a feature vector. Feature vectors are then collected into a matrix (each vector constitutes a matrix column) and used to make a decision. Therefore, a kernel function utilized by the LS-SVM classifier is defined over a pair of matrices \( (A, B) \) rather than over a pair of vectors. A positive definite kernel of such type has been recently proposed by Wolf and Shashua [18]. The authors use the principal angles between the two column spaces defined by the matrices \( (A, B) \) to assess the matching between the spaces and derive a positive definite kernel based on that concept. The "QR" factorization of the matrices \( (A, B) \) and the kernel Gram-Schmidt orthogonalization process are used to derive the kernel. Applying the "QR" factorization the matrices \( (A, B) \) can be written as \( A = Q_A R_A \) and \( B = Q_B R_B \), where \( Q \) is an orthonormal basis and \( R \) is an upper-diagonal matrix of size \( M \times M \) of the Gram-Schmidt coefficients representing the columns of the original matrix in the new basis. The principal angles \( \cos(\theta_i) \) are given by the singular values \( \sigma_i \) of the matrix \( Q_A^T Q_B \):

\[
\cos(\theta_i) = \sigma_i, \quad i = 1, \ldots, M
\]

It was shown that

\[
\kappa(A, B) = \det(Q_A^T Q_B)^2 = \prod_{i=1}^{M} \cos(\theta_i)^2
\tag{11}
\]

is a positive define kernel [18]. We use this kernel in our work. The algorithm for evaluating the kernel without explicit computation of \( Q_A \) and \( Q_B \) can be found in [18]. Only inner-products between the columns of \( A \) and the columns of \( B \) are used.

V. EXPERIMENTAL INVESTIGATIONS

A. Experimental setup

There were 200 image frames in one image sequence. However, only 20 image frames were used to estimate the kernel defined over a pair of matrices. Due to the small number of data available for the experiments, the leave-one-out test has been used to estimate the classification accuracy. The data used were normalized to zero mean and variance one. The polynomial kernel of degree \( q = 2 \) was used to extract the kernel principal components, while the Gaussian kernel was used to estimate the kernel defined over a pair of matrices. The experimental tests performed concern the influence of the LS-SVM regularization constant \( \gamma \), the Gaussian kernel width parameter \( \sigma \), and the number of the kernel principal components used on the test set data classification accuracy. The dependence of the classification accuracy on the percentage of the data variance accounted for by the number of the kernel principal components used was also studied.

B. Results

In Fig. 2, shown is the classification accuracy of the test set data as a function of the number of the kernel principal components used to characterize colour of one image frame, for given values of the regularization constant \( \gamma \) and the kernel width parameter \( \sigma \). As can be seen, nine principal components is the best choice.

![Fig. 2. The classification accuracy of the test set data as a function of the number of the kernel principal components, for given values of the regularization constant \( \gamma \) and the kernel width parameter \( \sigma \).](image)

The graph presented in Fig. 3 plots the test set data classification accuracy as a function of the percentage of the data variance accounted for by the number of the kernel principal directions used to represent colour. As can be seen from Fig. 3, the percentage of the data variance accounted for by the optimal number of the components is close to 90. Fig. 4 relates the test set data classification accuracy the regularization constant \( \gamma \), and the number of the kernel principal components used to represent colour. As can be seen from Fig. 4, a large number of principal components significantly deteriorates the classification accuracy. Fig. 5 plots the test set data classification accuracy as a function of the regularization constant \( \gamma \) and the kernel width parameter \( \sigma \).

VI. CONCLUSIONS

A technique for automated categorization of laryngeal colour image sequences obtained by video laryngostro-
bosphoscopy was developed. The LS-SVM employed to categorize an image sequence into the healthy, cancerous and noncancerous classes exploits a kernel function defined over a pair of matrices, rather than over a pair of vectors. The classification accuracy of over 85% was observed when testing the developed tools on data recorded during routine laryngeal videostroboscopy. One can expect increasing the accuracy even further by adding features of other types. A larger database needs to be collected for the comprehensive examination of the technique.

REFERENCES


