MUSCLE International Workshop on Computational Intelligence for Multimedia Understanding
Antalya, Turkey, October 3-4, 2013

Workshop Booklet
<table>
<thead>
<tr>
<th>Time</th>
<th>Day 1</th>
<th>Day 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>9:00-9:15</td>
<td>Registration</td>
<td></td>
</tr>
<tr>
<td>9:15-9:30</td>
<td>A. E. Cetin-E. Salerno Welcome Speech</td>
<td>2D and 3D object registration without point correspondences</td>
</tr>
<tr>
<td>9:30-10:30</td>
<td>M. Sezgin Intelligent user interfaces: practical considerations in design and implementation</td>
<td>Z. Kato</td>
</tr>
<tr>
<td>10:30-10:45</td>
<td>Coffee Break</td>
<td>Coffee Break</td>
</tr>
<tr>
<td>10:45-12:45</td>
<td><strong>Regular Session (Page 3)</strong></td>
<td><strong>Regular Session (Page 64)</strong></td>
</tr>
<tr>
<td></td>
<td>B.U Toreyin Sub-band Decomposition Based Supervised Feature Extraction for Hyperspectral Image Classification</td>
<td>M. Belahcene Face Recognition Optimization by EFM-PSO</td>
</tr>
<tr>
<td></td>
<td>J. Filip A Probabilistic Approach to Rough Texture Compression and Rendering</td>
<td>Z. Zdziarski Extension of GBVS to 3D media</td>
</tr>
<tr>
<td></td>
<td>J. G. Alaydin Graph-Cut-based Compression Algorithm for Compressed-Sensed Image Acquisition</td>
<td>P. Perner Our View To Novel Forensic Multimedia Data Analysis</td>
</tr>
<tr>
<td></td>
<td>M. Tofghi Denoising Using a Framework Based On Projections Onto Convex Sets (POCS)</td>
<td>M. Leszczuk Key Indicators for Monitoring of Audiovisual Quality</td>
</tr>
<tr>
<td>Time</td>
<td>Session</td>
<td>Speaker</td>
</tr>
<tr>
<td>-----------</td>
<td>-------------------------------------------------------------------------</td>
<td>--------------------------------</td>
</tr>
<tr>
<td>12:45-14:00</td>
<td>Lunch</td>
<td></td>
</tr>
<tr>
<td>14:00-15:00</td>
<td>Breakfast</td>
<td>M.N. Gurcan</td>
</tr>
<tr>
<td>14:00-15:15</td>
<td>Regular Session (Page 93)</td>
<td></td>
</tr>
<tr>
<td>15:00-16:45</td>
<td>MIRACLE Session (Page 28)</td>
<td>P. Belhomme</td>
</tr>
<tr>
<td>15:00-16:45</td>
<td></td>
<td>A. Bozkurt</td>
</tr>
<tr>
<td>15:00-16:45</td>
<td></td>
<td>P. Pemer</td>
</tr>
<tr>
<td>15:00-16:45</td>
<td></td>
<td>N. Grammalidis</td>
</tr>
<tr>
<td>15:00-16:45</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16:45-17:00</td>
<td>Coffee Break</td>
<td></td>
</tr>
<tr>
<td>16:45-17:00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17:00-18:15</td>
<td>Regular Session (Page 93)</td>
<td></td>
</tr>
<tr>
<td>17:00-18:15</td>
<td></td>
<td>N. Kiryati</td>
</tr>
<tr>
<td>17:00-18:15</td>
<td></td>
<td>A. Eleyan</td>
</tr>
<tr>
<td>17:00-18:15</td>
<td></td>
<td>I. Cinaroglu</td>
</tr>
</tbody>
</table>
Regular Session 1
Abstract—Hyperspectral image processing is an important research topic in remote sensing. One of the research topics in hyperspectral image processing is feature extraction. In this paper, one-dimensional sub-band decomposition based supervised feature extraction approach is proposed for hyperspectral image classification. In the proposed method, one-dimensional sub-band decomposition is applied before principal component analysis (PCA) to increase the classification accuracy. First, the hyperspectral signal is decomposed into its wavelet sub-bands and a new one-dimensional signal is reconstructed from these sub-bands based on the determined weights in a supervised manner. It is experimentally shown that the classification performance of the proposed method is better than the comparative method, especially for higher PCA dimensions.

I. INTRODUCTION

The hyperspectral imaging started with the development of a two dimensional imager that captures spectral data with a large number of bands in the mid-1980s [1]. Its application areas has grown rapidly especially for the last 10 years. Hyperspectral imagery has much richer information than multispectral and single-band imagery, since it includes usually hundreds of contiguous bands. On the other hand, huge data volume in hyperspectral imagery results in data storage and data transmission problems. Processing time of this huge data also has to be considered.

Additionally, sample size for training a specific classification method increases exponentially with increasing number of spectral bands. This is known as curse of dimensionality [2]. Curse of dimensionality leads to Hughes phenomenon [3]. If the required training data size for a specific classifier is insufficient then the estimation of the parameters used to calculate classification accuracy becomes inaccurate and unreliable. In order to handle this problem, either the training data size has to be enlarged or dimensionality of hyperspectral images has to be reduced with some dimension reduction techniques. The former, enlarging the training data size, is usually not possible for hundreds of spectral bands in practice. For the case of hyperspectral images, dimension reduction can be realized in two ways: Feature selection and feature extraction [4]. In feature selection techniques, a subset from the original input data which is optimum for the decided criteria is selected. On the other hand, in feature extraction methods, input data is transformed to a different space where a set of new features are produced without sacrificing the discriminative ability much.

In this study, the data is reduced with feature extraction techniques rather than feature selection ones. The aim of feature extraction is to achieve reduction in the dimensionality of the original data while maintaining significant information. Reduced number of features may result in a decreased generalization error and increased power and classification accuracy. Therefore, preserving the significant information is crucial in dimension reduction methods.

Various supervised and unsupervised feature extraction methods are implemented to reduce the hyperspectral data size [5]. One of the most popular of them is Principal Component Analysis (PCA) method [6]. In [6], the hyperspectral data is divided into several parts and hierarchical PCA algorithm is applied to each part of the data then few principal components are combined to be classified. Results of [6] indicate that the method realizes a remarkable reduction in data size without much degradation in classification accuracy.

The only constraint for the PCA basis vectors is orthonormality and it does not account for localization information of bands. Besides, a few of the principal components are considered in most of the implementations. Therefore, PCA used for dimension reduction does not include all the information data.

Unlike PCA, Wavelet Decomposition coefficients contain all the info about the original signal. Moreover, wavelet based feature extraction methods give local information about spectral variation of a hyperspectral signal in separate bands and at each scale. Therefore, while using Wavelet Transform (WT) it is possible to analyze data at different scales and resolutions since WT decomposes the original signal into a series of shifted and scaled versions of the determined mother wavelet.

Wavelet decomposition is applied to hyperspectral data several times in the literature. In [7], 1D Automatic Discrete Wavelet Transform is applied in the spectral domain for dimension reduction. This technique is compared with PCA and it is observed that classification accuracy of the former is better than the latter. Automatic decomposition level is determined by calculating the correlation between the original data and the decomposed sub-band in the specified decomposition level on the training data. After that, the determined decomposition level is applied to the whole data.

In [8], best wavelet packet basis which is an adaptive wavelet decomposition algorithm is performed as feature
extraction method for hyperspectral imaging. Best wavelet packet basis is determined by minimizing the cost function based on between-class distances of the training data.

In [9], DWT is performed for feature extraction and extracted features are evaluated by automated maximum likelihood classifier for their classification accuracy in two agricultural fields. Subset of 1D best wavelet coefficients is used and it is examined that the technique performs better than PCA. The results indicate that larger scale wavelet coefficients are more useful and utilizing these coefficients gives better results than observing discrete spectral bands.

The above studies on wavelet decomposition emphasize that Wavelet Decomposition applied for dimension reduction has higher classification rate than PCA while preserving its simplicity advantage over PCA. That is why; Wavelet is combined with PCA in [10]. In the study, PCA is applied to the hyperspectral image such that the selected PCA basis elements are not only orthogonal to each other but wavelets as well. In order to compare the results, supervised classification is applied to both original and reduced data. It is observed that data dimension is reduced with this technique and further improvement on classification accuracy is obtained when compared with PCA method. Therefore, we thought that combining PCA with Wavelet decomposition can improve the PCA performance while preserving efficient data reduction of PCA algorithm.

In our work, we propose a method that first decomposes the hyperspectral signal into its wavelet sub-bands and reconstructs a new signal from these sub-bands based on the determined weights in a supervised manner. Then PCA is implemented to the reconstructed signal and Support Vector Machine (SVM) [11] is utilized as a classifier of hyperspectral data.

Our contribution is to use all the branches of the wavelet decomposition although in most wavelet-based feature extraction methods only the low-pass coefficients of the decomposition are used [10]. Instead of selecting a set of the wavelet coefficient determined based on a set of criteria explained in [7]-[8], all of the wavelet coefficients are used for classification. Since the aim is to preserve the classification accuracy while reducing the data size, the weights of the sub branch of wavelet decompositions are determined by the classification accuracies of the small-size training data.

The rest of the paper is organized as follows: In Section II, the proposed one-dimensional sub-band decomposition based feature extraction approach is introduced. Section III demonstrates the experimental results. Finally, in Section IV, conclusions are given in addition to future research directions.

II. ONE-DIMENSIONAL SUB-BAND DECOMPOSITION BASED FEATURE EXTRACTION

In this work, we propose a method that first decomposes the hyperspectral signal into its wavelet sub-bands and reconstructs a new signal from these sub-bands based on the determined weights. Then PCA is implemented to the reconstructed signal and Support Vector Machine (SVM) method is utilized as a classifier of hyperspectral data. Flowchart of the proposed algorithm is presented in Figure 1.

Wavelet decomposition has two stages: filtering and down sampling. For filtering purposes 2-level Daubechies 4 wavelet transform [12] is applied. The noisy bands and water absorption bands are eliminated from the whole data.

Let $X$ be a hyperspectral image with $N_x$ spectral bands, $R_x$ rows and $C_x$ columns. A pixel located at $(r,c)$, where $1 \leq r \leq R_x$ and $1 \leq c \leq C_x$ has a spectral signature represented by the following vector:

$$X(r,c) = [X_1(r,c) \ X_2(r,c) \ ... \ X_{N_x}(r,c)].$$

(1)

The 2-level DWT blocks in Figure 1 are implemented using a 2-level Daubechies 4 filter-bank comprising of low-pass and high-pass filters followed by decimation. This way, the input signal is decomposed into 4 quarter-length sub-signals each of which covers one-fourth of the original signal’s frequency content.

Training data is selected randomly from a uniform distribution over the available data. Randomly selected training data from each class is applied to 2-level DWT and this forms four sub-branches as LL, LH, HL and HH. Suppose that $P$ samples are selected randomly for each class and the number of classes is $C$ so the training data size is defined as a $(P.C) \times N_B$ matrix where $P.C$ represents the pixel numbers located in $P$ classes. Therefore, $m$ and $n$ are selected randomly to define $P$ classes. Thus it is observed that Wavelet Decomposition applied for dimension reduction has higher classification rate than PCA while preserving its simplicity advantage over PCA.

Then PCA is performed on each sub-branch and only few of the PCA components are contributed to the classification of wavelet sub-bands.

![Figure 1. Flow-chart of the proposed method.](image-url)
Suppose that PCA dimension is k then the outputs of PCA applied to each four sub-branch of the two level WT becomes:

\[ \text{PCA}_{LL_{mn}} = [w_{LL_{mn,1}}, \ldots, w_{LL_{mn,k}}] \]  
\[ \text{PCA}_{LH_{mn}} = [w_{LH_{mn,1}}, \ldots, w_{LH_{mn,k}}] \]  
\[ \text{PCA}_{HL_{mn}} = [w_{HL_{mn,1}}, \ldots, w_{HL_{mn,k}}] \]  
\[ \text{PCA}_{HH_{mn}} = [w_{HH_{mn,1}}, \ldots, w_{HH_{mn,k}}] \]

where \( w_{LL_{mn,j}} \) is the jth component of PCA for \( 1 \leq j \leq k \) for the LL_{mn} sub-branch.

As a classification method SVM is utilized. After the utilization of SVM, four classification accuracies (CA) for each sub-band appear.

\[ \text{acc}_{LL} = \text{SVM}(\text{PCA}_{LL_{mn}}), \text{CA of LL sub-branch} \]  
\[ \text{acc}_{LH} = \text{SVM}(\text{PCA}_{LH_{mn}}), \text{CA of LH sub-branch} \]  
\[ \text{acc}_{HL} = \text{SVM}(\text{PCA}_{HL_{mn}}), \text{CA of HL sub-branch} \]  
\[ \text{acc}_{HH} = \text{SVM}(\text{PCA}_{HH_{mn}}), \text{CA of HH sub-branch} \]

Consider the linear mapping method explained in the next section is performed.

After calculating the weights (weight_{LL}, weight_{LH}, weight_{HL}, weight_{HH}) for each subband from the training data, 2-level DWT is performed on the overall hyperspectral data. Therefore m and n combinations describe all the data. 400 bands between 400nm and 2500nm. After removing water absorption and low SNR bands, 200 bands are used in the experiments. The ground truth of the test image is also available. After removal of noisy and water absorption bands, 200-band 145x145 pixel hyperspectral data is formed for experiments. The non-classified parts and classes that have small number of samples are eliminated from the ground truth. As a result, the hyperspectral data is evaluated over nine discriminative classes. 100 samples are selected randomly for each class so the training data size is defined as 900x200 where 900 represents the pixel numbers located in nine classes and 200 represents the number of bands.

Cross validation is applied to training samples for SVM training before calculating classification accuracies of each wavelet sub-band. Then linear mapping shown in Figure 2 is performed to classification accuracies for calculating the weights of each four wavelet sub-bands. Linear mapping is formed based on two reference points: max_par and min_par. Maximum parameter is the reference classification accuracy point whose weight is determined as 2 whereas the minimum parameter is the reference classification accuracy point whose weight is determined as 1. As it is known, if it is not preferred to change the weights of the outputs of the DWT then all the weights of sub-branches should be taken as 1. However, in this study it is preferred that the sub-branch that has the largest classification accuracy dominates the reconstructed signal from the four sub-branches. Moreover, it is required to take all the sub-branches into reconstruction process. Consequently, max_par is taken as 0.80 and min_par is taken as 0.50. This means that if the classification accuracy of the sub-band is 0.80 its weight is 2 and if the classification accuracy of the sub-band is 0.50 its weight is 1.

![Figure 2. Linear mapping of classification accuracy.](image)

The weights of the four sub-bands are calculated as follows:

\[ \text{weight}_{LL} = J \cdot \text{acc}_{LL} + K, \quad (2) \]  
\[ \text{weight}_{LH} = J \cdot \text{acc}_{LH} + K, \quad (3) \]  
\[ \text{weight}_{HL} = J \cdot \text{acc}_{HL} + K, \quad (4) \]  
\[ \text{weight}_{HH} = J \cdot \text{acc}_{HH} + K, \quad (5) \]

where \( J = \frac{1}{(\text{max\_par} - \text{min\_par})} \) and \( K = 2 - J \cdot \text{max\_par} \)

In the experiments, the effect of the number of PCA components on the proposed algorithm is investigated. In order to achieve this, Monte Carlo simulations are run for different PCA dimensions ranging from 4 to 30. It is thought that if the PCA dimension is below 4 then the selected principal components can be insufficient to represent the information of the hyperspectral data. Also, it is thought that if the PCA dimension is above 30 then the data size cannot be reduced much.

In order to evaluate the performance of the proposed method, the results of the method are compared with the results of PCA algorithm. Figure 3 indicates the classification
accuracies of the proposed wavelet based sub-band decomposition method and PCA with respect to the PCA dimensions. It is investigated that the classification accuracies of PCA and the proposed method are nearly the same for lower PCA dimensions. However, for higher PCA dimensions, the classification performance of the proposed method is better than that of PCA.

![Figure 3: Performance of PCA and Proposed Method with respect to PCA dimensions](image)

IV. CONCLUSIONS

PCA provides remarkable reduction in hyperspectral imagery. Unlike DWT, it cannot eliminate arbitrary anomalies in a band. Wavelet Transform contains both the variations in amplitude and spectral content. PCA is optimal when the joint distribution of the noise and the clutter is Gaussian. However, this is not the case in practice. It is often hard to distinguish between the clutter coming from the background and the target signal.

In this paper, one-dimensional sub-band decomposition is applied before PCA to increase the classification accuracy. It is experimentally investigated that classification performance of the supervised method that is proposed here is better than PCA especially for higher PCA dimensions larger than 6 while maintaining the reduction efficiency of PCA on hyperspectral imagery. It should be noticed that in this study, we did not concentrate on finding an optimal mapping between reconstruction weights and classification accuracies. We simply prefer to use of a linear mapping that does not result in so much difference between the original signal and the reconstructed signal. Therefore, it is possible to obtain better classification accuracies than the ones in Figure 3.

As a future work, we would like to concentrate on the mapping of classification accuracies in order to determine the weights of wavelet sub-bands for reconstruction. Instead of linear mapping, quadratic or another non-linear mapping can be used. Additionally, instead of the reconstruction of the sub-bands, a fusion approach of the results of the sub-bands can be examined.

ACKNOWLEDGEMENT

We would like to thank to Mustafa Teke and Ali Ömer Kozal for their help on software support.

REFERENCES

A PROBABILISTIC APPROACH TO ROUGH TEXTURE COMPRESSION AND RENDERING

Jiří Filip    Jiří Grim    Michal Haindl

Institute of Information Theory and Automation of the ASCR, Prague

ABSTRACT

Rough textures describe a general visual appearance of real-world materials with regard to view and illumination directions. As the massive size and dimensionality of such representations is a main limitation of their broader use, efficient parameterization and compression methods are needed. Our method is based on estimating the joint probability density of the included nine spatial, directional, and spectral variables in the form of a Gaussian mixture of product components. Our reflectance prediction formula can be expressed analytically as a simple continuous function of input variables and allows fast analytic evaluation for arbitrary spatial and directional values without need for a lengthy interpolation from a finite grid of angular measurements. This method achieves high compression ratio increasing linearly with texture spatial resolution.

Index Terms— rough texture, BTF, compression, mixture model

1. INTRODUCTION

Real-world natural or man-made materials change, due to rough and structured material surface, their appearance with respect to actual viewing and illumination conditions. A proper visualization, classification, or retrieval of such rough materials with respect to viewing and illumination conditions is a challenging task [1]. Although representations exist that can comprehend a material’s appearance variability, due to the massive size of the corresponding high-dimensional data set, it is not practical to use them. Probably the most often, but still not widely, used is a bidirectional texture function (BTF), introduced by Dana et al [2]. BTF is a seven-dimensional function representing the appearance of a material sample’s surface for variable illumination \( \omega_i(\theta_i, \phi_i) \) and view \( \omega_v(\theta_v, \phi_v) \) directions (see Fig. 1-a), resulting in a function \( BTF(\lambda, x, y, \theta_i, \phi_i, \theta_v, \phi_v) \), where \( \theta \) and \( \phi \) are elevation and azimuthal angles, respectively, \( \lambda \) is the spectral channel, and \( (x, y) \) is the spatial location on the texture.

The behavior of each view- and illumination-dependent BTF pixel at position \( x, y \) can be approximated as a five-dimensional bidirectional reflectance distribution function (BRDF) [4] \( BRDF_{x,y}(\lambda, \theta_i, \phi_i, \theta_v, \phi_v) \). A typical size of BTF is several gigabytes. Hence, it has been approximated by various compression and modeling approaches so far [5]. Most of them were focused on creating a highly compact parametric representation retaining as much of the original visual fidelity as possible.

In this paper we attempt to create an analytic BTF model based on a very compact set of parameters. Unlike most of the previous approaches, we want to design a model in ours that would allow fast, ideally graphics hardware supported, analytic evaluation for arbitrary input data variables without need for a lengthy interpolation from a predefined grid of angular measurements during the appearance rendering.

2. RELATED WORK

A number of rough texture compression and modeling techniques exist [5].

Factorization and BRDF Fitting Approaches – They are either based on BTF data linear factorization, clustering, pixel-wise modeling by analytical BRDF models, or by their combinations. A parametric representation of most of them is not fully analytic and thus the model evaluation requires a time-consuming interpolation for non-measured illumination and viewing directions in individual pixels on the basis of the compressed data known for the measured directions. The factorization approaches are most often based on PCA [6, 7] or tensors [8, 9, 10]. The clustering approaches usually accomplish factorization into a predefined set of clusters [3, 11]. Analytical pixel-wise reflectance BRDF models have also been used, but due to inherited limitations of BRDF (e.g., reciprocity) [12] they have had to be further extended [13, 14]. Approaches combining reflectance fitting with estimated meso-structure geometry have also been presented [15, 16]. Although the techniques mentioned above are

![Fig. 1. Parameterization of illumination and view directions within the rough texture: (a) standard parameterization using spherical angles, (b) onion-slice parameterization [3].](image-url)
able to reproduce material’s appearance in high visual quality and reconstruction speed, they are due to storing some sort of pixel-wise parametric information limited to compression ratios \( \approx 1 : 2000 \).

**Probabilistic Models** – The approaches that allow us to achieve of such compression ratios are based on probabilistic Markov random field (MRF) BTF models that are closely related to our method. Several different MRF models have been published in the past, based either on causal autoregressive models [17, 18] or on a Gaussian MRF model [19]. Due to the stochastic nature of the MRF models, they are less successful at reproducing regular or near-regular structures in BTF samples [5]. Hence these methods combine an estimated range map with a synthetic multi-scale smooth texture.

The Gaussian mixture (GM) models have been applied to static texture synthesis. The method described in [20], [21] was based on a multivariate GM model of the local statistical texture properties in a moving contextual neighborhood.

In this paper we consider the problem of general modeling and rendering of rough textures in full complexity of viewing and lighting conditions – in a nine-dimensional space (i.e., spatial (2D), directional (4D), and color (3D) dependencies). We suggest a solution based on simultaneous modeling BTF data using the nine-dimensional GM model. Such a model, contrary to some of the previous approaches, allows full-color modeling of arbitrary materials as well as analytic evaluation from a compact parametric set. Therefore, the model offers a compression potential that outperforms most of the BTF compression factorization approaches published so far.

### 3. PROPOSED ROUGH TEXTURE MODEL

**Method Overview** – The proposed method starts with a mean BRDF computation by averaging view- and illumination-dependent reflectance across individual BTF images. Then these mean values are subtracted from individual BTF images and the resulting data are subject to the fitting using a Gaussian mixture model. The other remaining inputs to the model are the number of mixture components used \( M \), and the number of iterations or minimal increment \( \varepsilon \) of the fitting quality evaluation function. Fitting of the Gaussian mixture is performed by means of the EM algorithm, resulting in a very compact parametric set. After the model’s parameters are fitted, their pixel-wise reconstruction is combined with mean BRDF value to obtain final BTF.

**Rough Texture Data Preprocessing** – Let \( \xi \) be the nine-dimensional vector, where \( \xi_1, \xi_2 \) are the spatial pixel coordinates of the source rough texture; \( \xi_3, \xi_4 \) and \( \xi_5, \xi_6 \) define the information concerning view and illumination directions, respectively; and \( \xi_7, \xi_8, \xi_9 \) denote the color RGB values.

First, the input illumination and viewing directions are transformed from common spherical angles \( (\theta_i, \varphi_i, \theta_r, \varphi_r) \) to “onion-cut” parameterization [3] \((\xi_3, \xi_1, \xi_5, \xi_6)\) (see Fig. 1), avoiding \( 0 \approx 2\pi \) discontinuity of azimuth angles in the spherical parameterization:

\[
\xi_4 = \arcsin(\sin \theta_i \cdot \cos \varphi_i), \quad \xi_3 = \arccos(\cos \theta_i / \cos \xi_4), \\
\xi_6 = \arcsin(\sin \theta_r \cdot \cos \varphi_r), \quad \xi_5 = \arccos(\cos \theta_r / \cos \xi_6).
\]

Finally, to simplify the prediction problem, we centralize the color reflectance values by subtracting the mean BRDF color values \( \bar{\xi} \), i.e., obtaining \( x_n = \xi_n - \bar{\xi}, \ n = 7, 8, 9 \) and \( x_n = \xi_n, \ n = 1 \ldots 6 \) forming \( S = \{ x^{(1)}, \ldots, x^{(|S|)} \} \).

**Gaussian Texture Model** – For the sake of predicting the output color values \( x_1, x_9, x_9 \) for any given input variables \( x_1, \ldots, x_9 \), we approximate the joint multivariate density function of \( x \) in the form of Gaussian mixture

\[
P(x) = \sum_{m \in M} w_m F(x | \mu_m, \sigma_m), \ x \in R^9, \quad (1)
\]

Here \( M = \{ 1, 2, \ldots, M \} \) is the index sets of components, \( N = \{ 1, 2, \ldots, 9 \} \) denotes the index sets of variables, \( w_m \) are probability weights and \( F(x | \mu_m, \sigma_m) \) denote the mixture components defined as products of univariate Gaussian densities [20], [21]:

\[
F(x | \mu_m, \sigma_m) = \prod_{n \in N} f_n(x_n | \mu_{mn}, \sigma_{mn}), \quad (2)
\]

\[
f_n(x_n | \mu_{mn}, \sigma_{mn}) = \frac{1}{\sqrt{2\pi \sigma_{mn}}} \exp \left\{ -\frac{(x_n - \mu_{mn})^2}{2\sigma_{mn}^2} \right\}.
\]

From the computational point of view the product components (2) avoid the risk of ill-conditioned covariance matrices and simplify the evaluation of marginal densities [cf. later Eq. (10)]. Note that the above-described GM model based on the product components does not imply independence of variables (i.e., it is not defined by marginal probability distributions alone). As there is no risk of over-fitting in the case of approximation problems like rough texture rendering, the chosen number of mixture components can be arbitrarily large. As it is well known, the weights of redundant components will be suppressed by the EM algorithm, and the computational time is the only practical limitation. Typically, we used \( \approx 500 \) components, providing a reasonable trade-off between visual quality and computational complexity.

**Parameter Estimation** – We use the data set \( S \) introduced above to estimate the mixture model. The EM algorithm maximizes the corresponding log-likelihood function

\[
L = \frac{1}{|S|} \sum_{s \in S} \log \left( \sum_{m \in M} w_m F(x | \mu_m, \sigma_m) \right) \quad (3)
\]

by means of the well-known EM iteration equations [20]:

\[
q(m | x) = \frac{w_m F(x | \mu_m, \sigma_m)}{\sum_{j \in M} w_j F(x | \mu_j, \sigma_j)} \quad x \in S, \quad (4)
\]

\[
w^*_m = \frac{1}{|S|} \sum_{s \in S} q(m | x), \ m \in M, \quad (5)
\]
\[ p_{mn}' = \frac{1}{\sum_{x \in S} q(m|x)} \sum_{x \in S} x_n q(m|x), \quad n \in \mathcal{N}, \]  
\[ (\sigma_{mn}')^2 = \frac{1}{\sum_{x \in S} q(m|x)} \sum_{x \in S} x_n^2 q(m|x) - (\mu_{mn}')^2. \]

Here, the apostrophe denotes the new parameter values in each iteration.

Considering large sample size and a large number of components, we may expect numerous local maxima of the log-likelihood function (3) but, according to our experience, the corresponding mixture estimates are of comparable quality. The frequently discussed implementation points of EM algorithm are therefore less relevant: we choose the number of components in hundreds ($M \approx 10^2$) according to problem complexity and initialize the parameters randomly. After the mixture model is estimated, it is used for prediction of the functional values as we explain in the following section.

**Reflectance Prediction from Parameters** – Let us suppose that the first six input variables corresponding to pixel coordinates, illumination and view directions are known. Denoting \( x_I = (x_1, x_2, \ldots, x_6) \in \mathcal{X}_I \) the subvector of input variables, \( I = \{1, 2, \ldots, 6\} \subset \mathcal{N} \), we can estimate the output color reflectance values \( x_7, x_8, x_9 \) by means of the conditional densities

\[ p_{n|z}(x_n|x_I) = \frac{P_{n|z}(x_n, x_I)}{P_2(x_I)} = \sum_{m \in \mathcal{M}} W_m(x_I) f_n|x_I|\mu_{mn}, \sigma_{mn}. \]

Here

\[ W_m(x_I) = \frac{w_m F(x_I|\mu_m, \sigma_m)}{\sum_{j \in \mathcal{M}} w_j F(x_I|\mu_j, \sigma_j)} \]

are the conditional weights given \( x_I \in \mathcal{X}_I \) and

\[ F(x_I|\mu_m, \sigma_m) = \prod_{n \in \mathcal{N}} f_n(x_n|\mu_{mn}, \sigma_{mn}) \]

denotes the marginal component functions corresponding to the subspace \( \mathcal{X}_I \). Note that the simple plug-in formula (8) is formally enabled by a simple evaluation of the marginal densities \( P_{n|z}(x_n, x_I) \) and \( P_2(x_I) \).

Equation (8) can be applied to predict the output color variables \( x_n, n \in \mathcal{N} \setminus I \), e.g., by computing the conditional expectations for indices \( n = 7, 8, 9 \)

\[ E\{x_n|x_I\} = \int x_n p_{n|z}(x_n|x_I)dx_n = \sum_{m \in \mathcal{M}} W_m(x_I) \mu_{mn}, \]

Note that the final estimated reflectance values \( \xi_7, \xi_8, \xi_9 \), are obtained as a sum of the previously subtracted mean BRDF values and GM model contributions:

\[ \xi_n = E\{x_n|x_I\} + \xi_n, \quad n = 7, 8, 9. \]

As the proposed model is fully parametric, the rough texture values for any spatial and angular coordinates are obtained analytically from parameters of individual components. Contrary to MRF BTF models [17, 18, 19], the proposed reconstruction of individual pixels is completely independent and can be easily implemented directly in graphics hardware (GPU) for fast visualization purposes.

### 4. Testing and Results

**Test Datasets** – We have used five data sets from the BTF Database Bonn\(^1\) (aluminum profile, corduroy, dark and light fabrics, and knitted wool). Four of the tested material samples are fabrics, exhibiting challenging visual interactions between the light and the material’s surface. These data have illumination and viewing directions \((n_x, n_y) = (81 \times 81)\) producing uniform sampling of a hemisphere above a material sample. A spatial resolution of the data sets is 256×256 pixels; but for the sake of faster and more convenient data processing we cut BTF tiles [22] that can be freely repeated without visually disruptive seams. The mean BRDFs obtained from the tested samples by means of averaging these BTF pixels are shown in Fig. 2. The rows and columns in the BRDF images are indices of individually measured illumination- and viewing directions uniformly covering the hemisphere above the sample, by spiral movement starting at its pole.

[Fig. 2] The mean BRDF computed for the five tested materials (rows illumination directions, columns viewing directions). "Diamond"-like patterns correspond to material’s anisotropic behavior for fixed elevation angles.

**Experimental Results** – The proposed parametric representation has broad application potential and its view- and illumination-dependent descriptive qualities are best illustrated in compressed data visualization. Therefore, we compared renderings from the original data (6561 images) with renderings obtained as reconstructions from the proposed model. Our implementation in the OpenGL evaluates the model at each pixel of the rough texture mapped on a triangulated 3D model. After conducting the practical experiments we set the numbers of components to \( M = 600 \). Note that \( M \) is related to spatial visual complexity of material structure rather than to pixel count. Results on the five BTF samples are shown in Fig. 3, where the first row (a) shows original data rendering and the second row (b) shows reconstruction by the model. Fig. 3 also shows tile sizes and compression ratios for the tested BTF tiles achieved by the proposed GM model’s parametric representation. Note that the ratios include the model’s parameters as well as mean BRDF data. Additionally, in Fig. 3-c we show differences between original and modeled images using RMSE, PSNR [dB] and SSIM metrics, as well as the difference image in the selected inset window. It is apparent that although the model tends to reliably match rough texture structure, in its current state it does not reproduce well all of the fine high-frequency details.

\[1\]http://btf.cs.uni-bonn.de/
as is visible, e.g., on the corduroy material. The current im-
gages and averaged RMSE / PSNR[dB] / SSIM values.

Table 1. Performance of the LPCA method [11].

<table>
<thead>
<tr>
<th>Material</th>
<th>Tile Size</th>
<th>C.R.</th>
<th>RMSE/PSNR/SSIM</th>
</tr>
</thead>
<tbody>
<tr>
<td>alu</td>
<td>21×26</td>
<td>1:18</td>
<td>1.97/42.5/0.99</td>
</tr>
<tr>
<td>corduroy</td>
<td>36×46</td>
<td>1:55</td>
<td>4.6/34.9/0.94</td>
</tr>
<tr>
<td>fabric dark</td>
<td>21×23</td>
<td>1:16</td>
<td>7.3/30.9/0.85</td>
</tr>
<tr>
<td>fabric light</td>
<td>19×23</td>
<td>1:10</td>
<td>2.0/42.2/0.98</td>
</tr>
<tr>
<td>knitted wool</td>
<td>25×25</td>
<td>1:21</td>
<td>3.3/37.9/0.95</td>
</tr>
</tbody>
</table>

Table 2. Compression ratios achieved by the GM model in comparison with previous approaches to the rough texture compression.

<table>
<thead>
<tr>
<th>Method (BTF tile:25×25 pix.)</th>
<th>C.R.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gaussian mixture model</td>
<td>195.6</td>
</tr>
<tr>
<td>Polynomial texture maps (per-view) PTM RF</td>
<td>13.5</td>
</tr>
<tr>
<td>Polynomial Lafortune (per-view) PLM RF</td>
<td>14.3</td>
</tr>
<tr>
<td>PCA factorization (per-view) PCA RF</td>
<td>11.2</td>
</tr>
<tr>
<td>PCA factorization (entire data) PCA BTF</td>
<td>23.8</td>
</tr>
<tr>
<td>Local PCA clustering (entire data) LPCA BTF</td>
<td>21.4</td>
</tr>
<tr>
<td>Probabilistic GMRF model (tilted range-map)</td>
<td>600.0</td>
</tr>
<tr>
<td>Probabilistic 2D CAR model (tilted range-map)</td>
<td>800.0</td>
</tr>
<tr>
<td>Probabilistic 3D CAR model (tilted range-map)</td>
<td>100.0</td>
</tr>
</tbody>
</table>

The graph in Fig. 4 shows different methods’ compression ratios dependency on number of BTF pixels. Although this dependency is linear in the Local PCA approach [11], which is also true for the proposed model, the compression ratios of our model are far higher given the same pixel count. Note, that the compression ratios of GMRF, 2D CAR, and 3D CAR methods have this dependency linear as well beyond the size of tiled height-map. Although their compression ratio is higher their visual quality is compromise due to height simulation of regular structures, that cannot handle non-local effects such as inter-reflections, translucency, or scattering.

Note that 500 components (M) require us to store M × 9 (variables) × 2 (μ and σ) = 9000 floating point values, in addition to the mean BRDF image represented by 81×81×3 = 6561×3 byte values, which in the case of a modest size of BTF data (256 × 256 × 81 × 3 = 1.3·10^9) represent a BTF sample compression ratio of 1:45 323 (for HDR 32-bits/channel) and 1:22 989 (for LDR 8-bits/channel).

5. CONCLUSIONS

This paper outlines a novel method for high-dimensional rough texture parameterization and compression. The method starts with data normalization and their modeling with the aid of a Gaussian mixtures model. This model allows full-color BTF reconstruction based on analytic evaluation for any combination of spatial and directional variables. Moreover, the size of the model's parametric representation is very compact, and its compression ratio is linearly dependent on a rough texture resolution. We are not aware of any parametric BTF model that would provide such a compact parametric representation together with the similar visual performance as the model proposed here. Although the model’s visualizations do not convey all fine details in a material’s structure yet, its main visual features are preserved accurately. In our future work we will attempt to improve implementation and input data processing steps to enhance speed and quality of the
model’s fitting and apply them to adaptive data measurement.

6. REFERENCES


Graph-Cut-based Compression Algorithm for Compressed-Sensed Image Acquisition

Julide Gulen Alaydin  
Seden Hazal Gulen  
Cankaya University  
Eskisehir Yolu 29.km Yenimahalle  
Ankara 06810, Turkiye  
{juledegulen, s.hazalgulen}@gmail.com

Maria Trocan  
Institut Superieure  
d’Electronique de Paris  
28 rue Ndc, Paris 75006, France  
Email: maria.trocan@isep.fr

Behcet Ugur Toreyin  
Cankaya University  
Eskisehir Yolu 29.km Yenimahalle  
Ankara 06810, Turkiye  
Email: toreyin@cankaya.edu.tr

Abstract—The purpose of the paper is to find the best quantizer allocation for compressed-sensed acquired images, by using a graph-cut quantizer allocation method. The compressed sensed acquisition is realized in a block-based manner, using a random projection matrix, and on the obtained block measurements a graph-cut-based quantizer allocation method is applied, in order to further reduce the bitrate associated to the measurements. Finally, the quantized measurements are reconstructed using a Smooth Projected Landweber recovery method. The proposed compression method for compressed sensed acquisition shows better results when compared to JPEG2000.

I. INTRODUCTION

Images in today’s era of digital media have become the creative convergence of human expression, communication, social interaction, education and promotion of digital arts. The problem that we face today is with their transmission i.e., they generally occupy much more space in a hard disk, or bandwidth in a transmission system, than their proverbial counterpart, thus compression of an image becomes necessary to reduce storage and transmission resources and to avoid data redundancy. Nowadays many compression techniques are presented like JPEG, JPEG2000 etc.to compress images efficiently. In this paper, we will compare the proposed method with JPEG2000 which has some disadvantages. For instance, it produces ringing artifacts within the image, and it is not adapted for noisy signals, as the compressed-sensed acquired images can be seen. The proposed method which we will present below gives visually better results and better compression ratio when it is compared with JPEG2000. In addition to this, it reduces cost and power consumption by using fewer sensors than what is used in traditional acquisition methods.

As it is known, Nyquist-Shannon sampling theory has a big role in signal processing area. It says that the number of samples necessary to reconstruct a signal without error is determined by its bandwidth, and the number of samples should be half of the bandwidth to prevent data loss. However, an alternative theory Compressed Sensing (CS) is presented in recent years by F.D. Donoho, E. Candes, T. Tao and J. Romberg [1][2][9][11] which shows that signals and images can be reconstructed from fewer data than it is considered necessary in Nyquist-Shannon sampling theory, by compressing and sampling data simultaneously. It says that the entire process of acquiring the full signal, computing all the transform coefficients, encoding the largest coefficients and discarding all the others and then applying compression is unnecessary, exerts much power and wastes time. Because of all these disadvantages, it is not economic. Therefore, sampling and compressing data simultaneously makes more sense. However, most of the studies in Compressed Sensing remain at the theoretical level and it is realized that compressed sensing is not suitable for real-time sensing of images because the measurement process requires access to the entire signal at once [2][3]. Thus, L. Gan developed an alternative method called Block Compressed Sensing [4] where the original image is divided into small blocks and each block is sampled independently using the same measurement operator. Since each block is processed independently, the initial solution can be easily obtained and the reconstruction process can be considerably accelerated.

As it is mentioned above, reducing the storage is substantial in digital media. In order to do this each pixel is labeled before they are stored. If the pixels of images have near or same values, then they get the same label quantity. The task of assigning a minimum and appropriate number of labels to the pixels of an image, in order to compress efficiently, is a challenge in the image compression area. Thus, in order to find a better solution to the labeling problem, a Graph-Cut minimization algorithm is developed and proposed by Y. Boykov and O. Veksler [5]. The Graph-Cut algorithm can also be employed in another areas like image segmentation, motion estimation, etc., in a flexible manner, the challenging part being the definition of the energies (or costs) to be minimized within the graph. In this work we will use the algorithm in [5] for the minimization of the quantization noise.

In our paper, we employ Compressed Sensing technique for image acquisition and the graph-cut based quantization cost minimization on the compressed-sensed measurements in order to further reduce the transmission cost, avoid redundant operations and get better compression results. At first, original image is divided into $B \times B$ blocks and sampled using a random projection matrix. Then, a graph-cut based quantizer
Consider a graph, defined by a set of vertices $V$ and a set of edges $E$; in our approach, the graph is designed in a grid-like manner, where each value within the measurement matrix (or the image pixels, in an image quantization context) represents a vertex, and the edges are given by the neighboring values, in the grid. A cut is the separation of nodes into two regions according to the cost values associated to vertices and edges. If one region of vertices $u$ is denoted by $S$ (source) and the latter one region of vertices $v$ by $T$ (sink), the edges connecting the source and sink represent the cut. The minimal cut is thus obtained by minimizing the cost of all edges that go from $S$ to $T$ [7][8], i.e.:

$$\min C = (S, T) = \min_{u \in S, v \in T} C(u, v)$$

Generally, in a graph-cut minimization problem, several minimal cuts are computed, cuts that minimize a functional defined as:

$$E(q) = \sum_{w \in P} D_u(q_u) + \sum_{(u, v) \in E} V_{u,v}(q_u, q_v)$$

$D_u(q_u)$ is a function obtained from the data that measures the cost of assigning the quantizers $q_u$ and $q_v$ to two neighboring vertices, $u$ and $v$, and represents the direct quantization noise, further denoted by ($E_{\text{data}}$); $V_{u,v}(q_u, q_v)$ is a function that measures the penalty for associated different quantizer values, thus introducing discontinuity, between neighboring vertices $u$ and $v$; this function will be denoted in the followings by ($E_{\text{smooth}}$). Smoothness cost and data cost are jointly minimized in order to assign the most appropriate quantizers to the measurement matrix. The graph-cut minimization in Eq. 2 has proven to be a useful multidimensional optimization tool which can enforce piecewise smoothness while preserving sharp discontinuities [7] within the dataset (an example of quantizer repartition - 2 cuts - can be seen in Fig.II-1).

2) Block Compressed Sensing with Smooth Projected Landweber Reconstruction (BCS-SPL): Compressive Sampling Theory [2] emphasizes that entire process of acquiring the full signal, computing all the transform coefficients, encoding the largest coefficients and discarding all the others, then compressing the signal is unnecessary and wastes time. Instead, sampling and compressing data simultaneously is more reasonable because it performs less computations and uses fewer sensors. However, CS has some disadvantages, as it is not suitable for real-time sensing of image because the measurement operation requires access to the full signal at once, and the reconstruction algorithms are generally very computational expensive. Instead, Block CS which is presented by L.Gan in [3] is more efficient for real-time applications because each block is acquired separately without waiting to sense the entire image. In this theory, original image is divided into small blocks and each block is sampled independently using the same measurement operator. This method has considerable advantages, as the measurement operator can be easily stored. Besides, the initial solution is easily obtained and the reconstruction process is speeded up since each block is processed independently [4][6].

Consider a $M \times N$ image which has $P=M \times N$ pixels resolution and assume it is required to take $n$ CS measurements. In block-based CS, the image is divided into small blocks with size of $B \times B$ each, and sampled with the same operator. Suppose that $x_j$ is a vector representation of signal’s $j^{th}$ block via raster scanning. The corresponding $y_j$ is then,

$$y_j = \Phi B x_j$$

where $\Phi_B$ is an $n_B \times B^2$ orthonormal measurement matrix ($\Phi^T \Phi = I$), with $n_B = \frac{n}{B} \times B^2$. Block CS is memory efficient as we need to store an $n_B \times B^2$ ensemble, $\Phi_B$, rather than a full $n \times P$ matrix [4][6]. Obviously, small $B$ requires less memory in storage and faster implementation, while large $B$ performs better reconstruction. In our implementation framework we have chosen $B = 32$, as it has been shown in [9] to be the best trade-off in terms of storage and recovery complexity.

As proposed in [6], the image $\hat{x}$ is reconstructed using a Smoothed Projected-Landweber (SPL) operator (4) and the artifacts caused by block-based acquisition and also further quantization, in our approach, are reduced by a smoothing.

\[ \hat{x} = SPL(\Phi_B, \tilde{\Psi}, \Psi), \]  

(4)

where \( \tilde{\Psi} \) represents the quantized measurement, after the quantizer allocation. Several transforms \( \Psi \) can be used for the final image recovery: the discrete cosine and wavelet transforms (DCT and DWT), or directional transforms as contourlets (CT) and complex valued dual trees DDWT [6]. The experimental results show that the directional transforms in the SPL-based CS reconstruction combined with quantization are efficient and fully operational.

3) Proposed Algorithm: In this work, the image is acquired by \( B \times B \) (i.e., \( B = 32 \)) block projections, sampled using a measurement matrix \( \Phi_B = B^2 \times (B^2 \times \text{rate}) \) by using the formula (3), where \( 0 < \text{rate} < 1 \) represents the sampling rate. Each \( y_j \) block , \( \forall j \in 1, \ldots, \frac{B^2}{\text{rate}} \) is then seen as a grid-graph, on which the graph-cut minimization algorithm in [5] is applied. In our approach, we have used uniform quantization values, \( q \in \{0, 1, \ldots, 2^{20}\} \) and the impact of quantization has been measured in terms of mean squared error (MSE). The energy (2) to be minimized becomes in this case:

\[ E(q) = \min \sum_j \left\| y_j - \left( \frac{y_j}{q_i} \right) q_i \right\| + \Delta_{izp}(q_i, q_p), \]  

(5)

where \( \Delta \) function represents a smoothness constraint, being equal to 0 if \( i = p \), e.g., same quantizer has been assigned to two neighbouring values \( y_j \) and \( y_p \), or a penalty constant \( \delta = \|q_i - q_p\| \), if the quantization steps differ. Once the minimization in (5) has been performed for all the blocks \( y_j \) and the quantized measurement blocks \( \tilde{y}_j \) have been obtained, the SPL algorithm in [4] is used for recovering the entire image.

III. Experimental Results

In our simulation framework, we have considered three representative test images: Baboon (512×512 pixels), Goldhill (512×512 pixels) and Peppers (512×512 pixels), which have been selected for their different texture characteristics. In order to prove the efficiency of the proposed quantization approach, expressed in terms of PSNR, the results obtained using the proposed algorithm have been compared to the ones obtained using JPEG2000 [12]. As previously mentioned, in this work we have considered scalar quantization, where the number of reconstruction levels is \( Q = 32 \). In Table I, II and III, the results obtained with the different \( \Psi \)-transforms used in the SPL reconstruction are denoted by \( GC - \text{Transform Name} - PSNR \). As it can be seen in Fig. 3 and Tables I-III, our scheme has an average gain of \( \approx 0.5\text{dBs} \) with respect to the JPEG2000, for the recovery with DDWT. This is due to the enhanced directionality given by the dual-tree transform. Moreover, better results are expected for more advanced quantization techniques, coupled with graph-cut minimization, like vector quantization or a Max-Lloyd-based quantization.

IV. Conclusion

In this paper we have proposed a graph-cut based quantizer allocation method for the images acquired in a compressed sensed environment. The compressed-sensed environment ensures a high cost reduction for the device in terms of required sensors, and coupled with an efficient direct quantization, the scheme can be easily deployed on low-cost acquisition tools. Moreover, the proposed scheme outperforms the current image compression standard JPEG2000 in terms of compression quality. This work can be further extended to cope with more efficient quantization methods within the graph-cut minimization framework, like vector quantization, etc.

REFERENCES

Fig. 3. Visual comparison for "Baboon" and "GoldHill" images: (a,c) Proposed scheme (b,d) JPEG2000.

**TABLE I**
PSNR (dB) performances and associated bitrates for Baboon

<table>
<thead>
<tr>
<th>bitrate</th>
<th>0.1</th>
<th>0.2</th>
<th>0.3</th>
<th>0.4</th>
<th>0.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bitrate</td>
<td>0.0675</td>
<td>0.1318</td>
<td>0.1835</td>
<td>0.2609</td>
<td>0.3049</td>
</tr>
<tr>
<td>GC-DWT-PSNR</td>
<td>20.4048</td>
<td>21.5179</td>
<td>22.4595</td>
<td>23.3633</td>
<td>24.3738</td>
</tr>
</tbody>
</table>

**TABLE II**
PSNR (dB) performances and associated bitrates for Peppers

<table>
<thead>
<tr>
<th>bitrate</th>
<th>0.1</th>
<th>0.2</th>
<th>0.3</th>
<th>0.4</th>
<th>0.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bitrate</td>
<td>0.0654</td>
<td>0.1258</td>
<td>0.2025</td>
<td>0.2189</td>
<td>0.2912</td>
</tr>
<tr>
<td>JPG2K-PSNR</td>
<td>27.0770</td>
<td>30.2925</td>
<td>31.4083</td>
<td>32.9541</td>
<td>33.5246</td>
</tr>
<tr>
<td>GC-DWT-PSNR</td>
<td>26.8593</td>
<td>30.2659</td>
<td>30.9880</td>
<td>32.0992</td>
<td>32.2072</td>
</tr>
<tr>
<td>GC-DDWT-PSNR</td>
<td>27.5086</td>
<td>30.7310</td>
<td>31.5162</td>
<td>32.6717</td>
<td>32.2512</td>
</tr>
<tr>
<td>GC-CT-PSNR</td>
<td>25.7390</td>
<td>29.7742</td>
<td>30.8005</td>
<td>31.8353</td>
<td>32.0051</td>
</tr>
</tbody>
</table>

**TABLE III**
PSNR (dB) performances and associated bitrates for Goldhill

<table>
<thead>
<tr>
<th>bitrate</th>
<th>0.1</th>
<th>0.2</th>
<th>0.3</th>
<th>0.4</th>
<th>0.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bitrate</td>
<td>0.0654</td>
<td>0.1245</td>
<td>0.1803</td>
<td>0.2452</td>
<td>0.2949</td>
</tr>
<tr>
<td>JPG2K-PSNR</td>
<td>25.3992</td>
<td>27.6255</td>
<td>28.6356</td>
<td>29.5932</td>
<td>30.6202</td>
</tr>
<tr>
<td>GC-DCT-PSNR</td>
<td>25.8229</td>
<td>27.4554</td>
<td>28.2236</td>
<td>29.2446</td>
<td>29.7786</td>
</tr>
<tr>
<td>GC-DWT-PSNR</td>
<td>25.3399</td>
<td>27.7189</td>
<td>28.9254</td>
<td>29.9020</td>
<td>30.3231</td>
</tr>
<tr>
<td>GC-CT-PSNR</td>
<td>20.2264</td>
<td>24.8591</td>
<td>29.2865</td>
<td>29.9384</td>
<td>30.5313</td>
</tr>
</tbody>
</table>
Denoising Using a Framework Based On Projections Onto Convex Sets (POCS)

Mohammad Tofighi, Kivanc Kose*, A. Enis Cetin
Dept. of Electrical and Electronic Engineering, Bilkent University, Ankara, Turkey
* Dermatology Department, Memorial Sloan Kettering Cancer Center, New York

tofighi@ee.bilkent.edu.tr, cetin@bilkent.edu.tr, "kosek@mskcc.org

Abstract—Two new optimization techniques based on projections onto convex space (POCS) framework for solving convex optimization problems are presented. The dimension of the minimization problem is lifted by one and sets corresponding to the cost function are defined. If the cost function is a convex function in $\mathbb{R}^N$ the corresponding set is also a convex set in $\mathbb{R}^{N+1}$. The iterative optimization approach starts with an arbitrary initial estimate in $\mathbb{R}^{N+1}$ and an orthogonal projection is performed onto one of the sets in a sequential manner at each step of the optimization problem. The method provides globally optimal solutions in total-variation (TV), filtered variation (FV), $\ell_1$, and entropic cost functions. A new denoising algorithm using the TV framework is developed. The new algorithm does not require any of the regularization parameter adjustment. Simulation examples are presented.

Index Terms—Projection onto Convex Sets, Bregman Projections, Iterative Optimization, Lifting

I. INTRODUCTION

In many inverse signal and image processing problems and compressing sensing problems an optimization problem is solved to find a solution to the following problem:

$$\min_{w \in \mathbb{C}} f(w), \tag{1}$$

where $\mathbb{C}$ is a set in $\mathbb{R}^N$ and $f(w)$ is the cost function. Some commonly used cost functions are based on $\ell_1$, $\ell_2$, total-variation (TV), filtered variation, and entropic functions [1]–[5]. Bregman developed iterative methods based on the so-called Bregman distance to solve convex optimization problems which arise in signal and image processing [6]. In Bregman’s approach, it is necessary to perform a D-projection (or Bregman projection) at each step of the algorithm and it may not be easy to compute the Bregman distance in general [5], [7], [8].

In this article Bregman’s older projections onto convex sets (POCS) framework [9], [10] is used to solve convex and some non-convex optimization problems instead of the Bregman distance approach. Bregman’s POCS method has also been widely used for finding a common point of convex sets in many inverse signal and image processing problems [10]–[33]. In the ordinary POCS approach the goal is simply to find a vector which is in the intersection of convex sets. In each step of the iterative algorithm an orthogonal projection is performed onto one of the convex sets. Bregman showed that successive orthogonal projections converge to a vector which is in the intersection of all the convex sets. If the sets do not intersect iterates oscillate between members of the sets [34]–[36]. Since there is no need to compute the Bregman distance in standard POCS, it found applications in many practical problems.

In our approach the dimension of the minimization problem is lifted by one and sets corresponding to the cost function are defined. This approach is graphically illustrated in Fig. 1. If the cost function is a convex function in $\mathbb{R}^N$ the corresponding set is also a convex set in $\mathbb{R}^{N+1}$. As a result the convex minimization problem is reduced to finding a specific member (the optimal solution) of the set corresponding to the cost function. As in ordinary POCS approach the new iterative optimization method starts with an arbitrary initial estimate in $\mathbb{R}^{N+1}$ and an orthogonal projection is performed onto one of the sets. After this vector is calculated it is projected onto the other set. This process is continued in a sequential manner at each step of the optimization problem. The method provides globally optimal solutions in total-variation, filtered variation, $\ell_1$, and entropic function based cost functions because they are convex cost functions.

The article is organized as follows. In Section II, the convex minimization method based on the POCS approach is introduced. In Section III, a new denoising method based on the convex minimization approach introduced in Section II, is presented. This new approach uses supporting hyperplanes of the TV function and it does not require a regularization parameter as in other TV based methods. Since it is very easy to perform an orthogonal projection onto a hyperplane this method is computationally implementable for many cost functions without solving any nonlinear equations. In Section IV, we present the simulation results and some denoising examples.

II. CONVEX MINIMIZATION

Let us first consider a convex minimization problem

$$\min_{w \in \mathbb{R}^N} f(w), \tag{2}$$

where $f : \mathbb{R}^N \rightarrow \mathbb{R}$ is a convex function. We increase the dimension by one to define the following sets in $\mathbb{R}^{N+1}$ corresponding to the cost function $f(w)$ as follows:

$$C_f = \{ w = [w^T y]^T : y \geq f(w) \}, \tag{3}$$

which is the set of $N+1$ dimensional vectors whose $(N+1)^{st}$ component $y$ is greater than $f(w)$. This set $C_f$ is called the
epigraph of \( f \). We use bold face letters for \( N \) dimensional vectors and underlined bold face letters for \( N+1 \) dimensional vectors, respectively.

The second set that is related with the cost function \( f(w) \) is the level set:

\[
C_s = \{ w = [w^T \ y]^T : y \leq \alpha, \ w \in \mathbb{R}^{N+1} \}, \tag{4}
\]

where \( \alpha \) is a real number. Here it is assumed that \( f(w) \geq \alpha \) for all \( f(w) \in \mathbb{R} \) such that the sets \( C_f \) and \( C_s \) do not intersect. They are both closed and convex sets in \( \mathbb{R}^{N+1} \). Sets \( C_f \) and \( C_s \) are graphically illustrated in Fig. 1 in which \( \alpha = 0 \).

The POCS based minimization algorithm starts with an arbitrary \( w_0 = [w_0^T \ y_0]^T \in \mathbb{R}^{N+1} \). We project \( w_0 \) onto the set \( C_s \) to obtain the first iterate \( w_1 \) which will be,

\[
w_1 = [w_0^T \ 0]^T, \tag{5}
\]

where \( \alpha = 0 \) is assumed as in Fig. 1. Then we project \( w_1 \) onto the set \( C_f \). The new iterate \( w_2 \) is determined by minimizing the distance between \( w_1 \) and \( C_f \), i.e.,

\[
w_2 = \arg \min_{w \in C_s} \| w - w_1 \|. \tag{6}
\]

Equation 6 is the ordinary orthogonal projection operation onto the set \( C_f \in \mathbb{R}^{N+1} \). To solve the problem in Eq. 6 we do not need to compute the Bregman’s so-called D-projection. After finding \( w_2 \), we perform the next projection onto the set \( C_s \) and obtain \( w_3 \) etc. Eventually iterates oscillate between two nearest vectors of the two sets \( C_s \) and \( C_f \). As a result we obtain

\[
\lim_{n \to \infty} w_n = [w^* f(w^*)]^T, \tag{7}
\]

where \( w^* \) is the \( N \) dimensional vector minimizing \( f(w) \). The proof of Eq. 7 follows from Bregman’s POCS theorem [9], [34]. It was generalized to non-intersection case by Gubin et. al [12], [34], [35]. Since the two closed and convex sets \( C_s \) and \( C_f \) are closest to each other at the optimal solution case, iterations oscillate between the vectors \( [w^* f(w^*)]^T \) and \( [w^* 0]^T \) in \( \mathbb{R}^{N+1} \) as \( n \) tends to infinity. It is possible to increase the speed of convergence by non-orthogonal projections [24].

If the cost function \( f \) is not convex and have more than one local minimum then the corresponding set \( C_f \) is not convex in \( \mathbb{R}^{N+1} \). In this case iterates may converge to one of the local minima.

### III. Denoising Using POCS

In this section, we present a new method of denoising, based on TV and FV. Let the noisy signal be \( y \), and the original signal or image be \( w_0 \). Suppose that the observation model is the additive noise model:

\[
y = w_0 + v, \tag{8}
\]

where \( v \) is the additional noise. In this approach we solve the following problem for denoising:

\[
w^* = \arg \min_{w \in C_f} \| y - w \|^2, \tag{9}
\]

where, \( y = [y^T \ 0] \) and \( C_f \) is the epigraph set of TV or FV in \( \mathbb{R}^{N+1} \). The minimization problem is essentially the orthogonal projection onto the set \( C_f \). This means that we select the nearest vector \( w^* \) on the set \( C_f \) to \( y \). This is graphically illustrated in Fig. 2.

In current TV based denoising methods [37], [38] the following cost function is used:

\[
\min \| y - w \|^2 + \lambda \text{TV}(w). \tag{10}
\]

The solution of this problem can be obtained using the method that we discussed in Section II. One problem with this approach is the estimation of the regularization parameter \( \lambda \). One has to determine the \( \lambda \) in an ad hoc manner or by visual inspection. On the other hand we do not require any parameter adjustment in (9).

The denoising solution in (9) can be found by performing successive orthogonal projection onto supporting hyperplanes of the epigraph set \( C_f \). In the first step we calculated \( \text{TV}(y) \). We also calculate the surface normal at \( y = [y^T \ \text{TV}(y)] \) in \( \mathbb{R}^{N+1} \) and determine the equation of the supporting hyperplane at \( [y^T \ \text{TV}(y)] \). We project \( y = [y^T \ 0] \) onto this...
hyperplane and obtain \( w_1 \) as our first estimate as shown in Fig. 3. In the second step we project \( w_1 \) onto the set \( C_s \) by simply making its last component zero. We calculate the TV of this vector and the surface normal, and the supporting hyperplane as in the previous step. We project \( y \) onto the new supporting hyperplane, etc.

The sequence of iterations obtained in this manner converges to a vector in the intersection of \( C_s \) and \( C_f \). In this problem the sets \( C_s \) and \( C_f \) intersect because \( \text{TV}(w) = 0 \) for \( w = [0, 0, ..., 0]^T \) or for a constant vector. However, we do not want to find a trivial constant vector in the intersection of \( C_s \) and \( C_f \). We calculate the distance between \( y \) and \( w_i \) at each step of the iterative algorithm described in the previous paragraph. This distance \( \|y - w_i\|_2^2 \) initially decreases and starts increasing as \( i \) increases. Once we detect the increase we perform some refinement projections to obtain the solution of the denoising problem. A typical convergence graph is shown in Fig. 4 for the “note” image. Simulation examples are presented in the next section.

IV. SIMULATION RESULTS

Consider the “Note” image shown in Fig. 5. This is corrupted by a zero mean Gaussian noise with \( \lambda = 45 \) in Fig. 6. The image is restored using our method and Chambolle’s algorithm [37] and the denoised images are shown in Fig. 7 and 8, respectively. The \( \lambda \) parameter in (10) is manually adjusted to get the best possible results. Our algorithm not only produce a higher SNR, but also a visually better looking image. Solution results for other SNR levels are presented in Table I. We also corrupted this image with \( \epsilon \)-contaminated Gaussian noise (“salt-and-pepper noise”). Denoising results are summarized in Table II.

In Table III, denoising results for 10 other images with different noise levels are presented. In almost all cases our method produces higher SNR results than the denoising results obtained using [37].

V. CONCLUSION

A new denoising method based on the epigraph of the TV function is developed. The solution is obtained using POCS. The new algorithm does not need the optimization of the regularization parameter.

<table>
<thead>
<tr>
<th>Noise level, ( \text{std} )</th>
<th>Input SNR</th>
<th>POCS</th>
<th>Chambolle</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>21.12</td>
<td>30.63</td>
<td>29.48</td>
</tr>
<tr>
<td>10</td>
<td>15.12</td>
<td>25.93</td>
<td>24.20</td>
</tr>
<tr>
<td>15</td>
<td>11.56</td>
<td>22.91</td>
<td>21.05</td>
</tr>
<tr>
<td>20</td>
<td>9.06</td>
<td>20.93</td>
<td>18.90</td>
</tr>
<tr>
<td>25</td>
<td>7.14</td>
<td>19.27</td>
<td>17.17</td>
</tr>
<tr>
<td>30</td>
<td>5.59</td>
<td>17.89</td>
<td>15.78</td>
</tr>
<tr>
<td>35</td>
<td>4.21</td>
<td>16.68</td>
<td>14.69</td>
</tr>
<tr>
<td>40</td>
<td>3.07</td>
<td>15.90</td>
<td>13.70</td>
</tr>
<tr>
<td>45</td>
<td>2.05</td>
<td>15.08</td>
<td>12.78</td>
</tr>
<tr>
<td>50</td>
<td>1.12</td>
<td>14.25</td>
<td>12.25</td>
</tr>
</tbody>
</table>

REFERENCES


Fig. 9. Sample images used in our experiments (a) House, (b) Jet plane, (c) Lake, (d) Lena, (e) Living room, (f) Mandrill, (g) Peppers, (h) Pirate.


TABLE II
COMPARISON OF THE RESULTS FOR DENOISING ALGORITHMS FOR ϵ-CONTAMINATION NOISE FOR NOTE IMAGE

<table>
<thead>
<tr>
<th>ϵ</th>
<th>σ₁</th>
<th>σ₂</th>
<th>Input SNR</th>
<th>POCs</th>
<th>Chambolle</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.9</td>
<td>5</td>
<td>30</td>
<td>14.64</td>
<td>23.44</td>
<td>20.56</td>
</tr>
<tr>
<td>0.9</td>
<td>5</td>
<td>40</td>
<td>12.55</td>
<td>21.39</td>
<td>17.60</td>
</tr>
<tr>
<td>0.9</td>
<td>5</td>
<td>50</td>
<td>10.75</td>
<td>19.49</td>
<td>15.54</td>
</tr>
<tr>
<td>0.9</td>
<td>5</td>
<td>60</td>
<td>9.29</td>
<td>17.61</td>
<td>13.82</td>
</tr>
<tr>
<td>0.9</td>
<td>5</td>
<td>70</td>
<td>7.98</td>
<td>16.01</td>
<td>12.57</td>
</tr>
<tr>
<td>0.9</td>
<td>5</td>
<td>80</td>
<td>6.89</td>
<td>14.54</td>
<td>11.37</td>
</tr>
<tr>
<td>0.9</td>
<td>10</td>
<td>30</td>
<td>12.56</td>
<td>22.88</td>
<td>19.74</td>
</tr>
<tr>
<td>0.9</td>
<td>10</td>
<td>40</td>
<td>11.13</td>
<td>21.00</td>
<td>15.30</td>
</tr>
<tr>
<td>0.9</td>
<td>10</td>
<td>50</td>
<td>9.85</td>
<td>19.35</td>
<td>12.47</td>
</tr>
<tr>
<td>0.9</td>
<td>10</td>
<td>60</td>
<td>8.58</td>
<td>17.87</td>
<td>10.42</td>
</tr>
<tr>
<td>0.9</td>
<td>10</td>
<td>70</td>
<td>7.52</td>
<td>16.38</td>
<td>8.76</td>
</tr>
<tr>
<td>0.9</td>
<td>10</td>
<td>80</td>
<td>6.46</td>
<td>15.05</td>
<td>7.45</td>
</tr>
<tr>
<td>0.95</td>
<td>5</td>
<td>30</td>
<td>16.75</td>
<td>24.52</td>
<td>23.18</td>
</tr>
<tr>
<td>0.95</td>
<td>5</td>
<td>40</td>
<td>14.98</td>
<td>22.59</td>
<td>20.44</td>
</tr>
<tr>
<td>0.95</td>
<td>5</td>
<td>50</td>
<td>13.41</td>
<td>20.54</td>
<td>18.45</td>
</tr>
<tr>
<td>0.95</td>
<td>5</td>
<td>60</td>
<td>12.10</td>
<td>18.72</td>
<td>16.80</td>
</tr>
<tr>
<td>0.95</td>
<td>5</td>
<td>70</td>
<td>10.80</td>
<td>17.13</td>
<td>15.34</td>
</tr>
<tr>
<td>0.95</td>
<td>5</td>
<td>80</td>
<td>9.76</td>
<td>15.63</td>
<td>14.11</td>
</tr>
<tr>
<td>0.95</td>
<td>10</td>
<td>30</td>
<td>13.68</td>
<td>23.79</td>
<td>20.43</td>
</tr>
<tr>
<td>0.95</td>
<td>10</td>
<td>40</td>
<td>12.66</td>
<td>22.09</td>
<td>15.35</td>
</tr>
<tr>
<td>0.95</td>
<td>10</td>
<td>50</td>
<td>11.71</td>
<td>20.65</td>
<td>12.28</td>
</tr>
<tr>
<td>0.95</td>
<td>10</td>
<td>60</td>
<td>10.72</td>
<td>19.10</td>
<td>10.22</td>
</tr>
<tr>
<td>0.95</td>
<td>10</td>
<td>70</td>
<td>9.82</td>
<td>17.59</td>
<td>8.66</td>
</tr>
<tr>
<td>0.95</td>
<td>10</td>
<td>80</td>
<td>8.92</td>
<td>16.12</td>
<td>7.34</td>
</tr>
</tbody>
</table>

Fig. 5. Original “Note” image.

Fig. 6. “Note” image corrupted with Gaussian noise with λ = 45.

Fig. 7. Denoised image “Note” image, using [1 -1] filter; SNR = 15.08.

123, 2011.


TABLE III

<table>
<thead>
<tr>
<th>Images</th>
<th>Noise std</th>
<th>Input SNR</th>
<th>POCS</th>
<th>Chambolle</th>
</tr>
</thead>
<tbody>
<tr>
<td>House</td>
<td>30</td>
<td>13.85</td>
<td>27.43</td>
<td>27.13</td>
</tr>
<tr>
<td>House</td>
<td>50</td>
<td>9.45</td>
<td>24.20</td>
<td>24.36</td>
</tr>
<tr>
<td>Lena</td>
<td>30</td>
<td>12.95</td>
<td>23.63</td>
<td>23.54</td>
</tr>
<tr>
<td>Lena</td>
<td>50</td>
<td>8.50</td>
<td>21.46</td>
<td>21.37</td>
</tr>
<tr>
<td>Mandrill</td>
<td>30</td>
<td>13.04</td>
<td>19.98</td>
<td>19.64</td>
</tr>
<tr>
<td>Mandrill</td>
<td>50</td>
<td>8.61</td>
<td>17.94</td>
<td>17.92</td>
</tr>
<tr>
<td>Living room</td>
<td>30</td>
<td>12.65</td>
<td>21.21</td>
<td>20.88</td>
</tr>
<tr>
<td>Living room</td>
<td>50</td>
<td>8.20</td>
<td>19.25</td>
<td>19.05</td>
</tr>
<tr>
<td>Lake</td>
<td>30</td>
<td>13.44</td>
<td>22.19</td>
<td>21.86</td>
</tr>
<tr>
<td>Lake</td>
<td>50</td>
<td>8.97</td>
<td>20.03</td>
<td>19.90</td>
</tr>
<tr>
<td>Jet plane</td>
<td>30</td>
<td>15.57</td>
<td>26.28</td>
<td>25.91</td>
</tr>
<tr>
<td>Jet plane</td>
<td>50</td>
<td>11.33</td>
<td>23.91</td>
<td>23.54</td>
</tr>
<tr>
<td>Peppers</td>
<td>30</td>
<td>12.65</td>
<td>23.57</td>
<td>23.59</td>
</tr>
<tr>
<td>Peppers</td>
<td>50</td>
<td>8.20</td>
<td>21.48</td>
<td>21.36</td>
</tr>
<tr>
<td>Pirate</td>
<td>30</td>
<td>12.13</td>
<td>21.39</td>
<td>21.30</td>
</tr>
<tr>
<td>Pirate</td>
<td>50</td>
<td>7.71</td>
<td>19.37</td>
<td>19.43</td>
</tr>
<tr>
<td>Cameraman</td>
<td>30</td>
<td>12.97</td>
<td>24.13</td>
<td>23.67</td>
</tr>
<tr>
<td>Cameraman</td>
<td>50</td>
<td>8.55</td>
<td>21.55</td>
<td>21.22</td>
</tr>
<tr>
<td>Flower</td>
<td>30</td>
<td>11.84</td>
<td>21.97</td>
<td>20.89</td>
</tr>
<tr>
<td>Flower</td>
<td>50</td>
<td>7.42</td>
<td>19.00</td>
<td>18.88</td>
</tr>
<tr>
<td>Average</td>
<td>30</td>
<td>13.11</td>
<td>23.18</td>
<td>22.84</td>
</tr>
<tr>
<td>Average</td>
<td>50</td>
<td>8.69</td>
<td>20.82</td>
<td>20.70</td>
</tr>
</tbody>
</table>

Fig. 8. Denoised image “Note” image, using Chambolle’s algorithm; SNR = 12.78.

[29] Y. Censor and S. A. Zenios, “Proximal minimization algorithm withd-
functions,” Journal of Optimization Theory and Applications, vol. 73,
limited Functions,” Journal of Mathematical Analysis and Applications,
[31] Y. Censor, “Row-action methods for huge and sparse systems and their
entropy over linear constraints,” Applied Numerical Mathematics, vol. 7,
[34] L. Gubin, B. Polyak, and E. Raik, “The Method of Projections for
Finding the Common Point of Convex Sets,” {USSR} Computational
d’optimisation structurés,” 2012.
the FFT Algorithm,” IEEE Signal Processing Magazine, vol. 14, no. 2,
[37] A. Chambolle, “An algorithm for total variation minimization and
2004.
[38] P. L. Combettes and J.-C. Pesquet, “Proximal splitting methods in signal
processing,” in Fixed-Point Algorithms for Inverse Problems in Science
and Engineering, ser. Springer Optimization and Its Applications, H. H.
Bauschke, R. S. Burachik, P. L. Combettes, V. Elser, D. R. Luke, and
AN ADAPTIVE METHOD FOR HIERARCHICAL TEXTURE-BASED SEGMENTATION

Wafa Romdhane and Walid Barhoumi

Research team SIIVA, RIADI laboratory, Manouba University, Tunisia

ABSTRACT

Texture is a complex concept seen the diversity of its visual aspects, what makes challenging the establishment of a segmentation method while considering the different texture types. Besides, the image segmentation is strongly related to the scale observation and to the definition of an analysis window. We propose an adaptive segmentation providing two levels of texture perception. After the over-segmentation of the image, a statistical analysis is conducted to determine the nature of each region (fine texture vs. coarse texture). Then, a wavelet-based analysis is achieved by adapting the wavelet sub-band choice to the textural nature of each region. Final segmentation is obtained by merging regions, of the same nature, while exploiting an adjacency graph. Results show the accuracy of the proposed method.

Index Terms— Image segmentation, texture, wavelet.

1. INTRODUCTION

Segmentation is a challenging low-level step for various image processing applications. However, the presence of the texture makes the task more difficult. This criterion is of great importance for the analysis of several types of images (e.g. satellite and medical images). Texture-based segmentation is founded on the definition of a set of descriptive attributes and their clustering using mean shift [1], split and merge [2] or graph partitioning [3]. However, the diversity of texture definitions presents a real problem to the establishment of a segmentation method which takes into account the different types of textures. In fact, the texture may be fine, with a chaotic distribution of greyscale, or coarse, defined by a repetition of a basic pattern, called “texel”, according to a spatial arrangement rule. Moreover, the problem of image segmentation is strongly related to the scale of observation and to the definition of an analysis window. Indeed, the texture is not a punctual property and it is detectable only for a group of pixels. Thus, segmentation methods addressing the problem in a closely local way suffer from the lack of spatial coherence. Therefore, large-scale measures must be integrated. Recent works have focused on representing texture at different scales with regions trees [4], using hidden Markov trees [5] or wavelets [6]. However, while applying coarse-scale measures near segments boundaries, texture transitions between different textures will not be thus detected. Ideally, we should apply coarse measurements only within segments. Based on this idea, the principle of the proposed texture-based method is very close to the split and merge model. The first step performs an over-segmentation of the image to obtain a map of small homogeneous regions. The over-segmentation is achieved by a method based on graph partitioning providing a global view of the image, while combining information from contour and texture. At this scale, analysis is made on the obtained regions to determine their nature (fine texture vs. coarse texture). In order to effectively determine the nature of each region, the choice of descriptors and classification method was the subject of several tests. The second step merges adjacent regions of the over-segmented image while adapting the choice of the wavelet sub-band relatively to the already defined nature of each region. Tests on the “Berkley” benchmark showed the effectiveness of the proposed method compared to the state of the art methods.

2. RELATED WORKS

Texture-based segmentation consists on the definition of descriptive attributes, using texture analysis techniques, in order to group similar connected pixels using segmentation techniques. To define texture-based attributes, the way in which a texture is perceived should be discerned. Since there is no single texture definition and that this notion covers various aspects, the texture perception is a complex task. Some works define the texture as a set of structured basic elements (texels). Others characterize it by a random distribution of greyscales. In a more general way, [7] defines the texture as a two-dimensional phenomenon with two levels of perception. The first tonal primitive is the basic component of the texture, and the second is specific to the spatial arrangement of primitives. Thus, texture is described by the number and the type of its primitives as well as their spatial arrangement. This distribution may be random as it may have a dependency on primitives compared to its neighboring ones. This dependence can be structural, probabilistic or functional. Besides, depending on the type of tonal primitives, texture can be “fine” when the spatial structure in the tonal primitives is random and the grayscale variation between the primitives is large, or “coarse” when the spatial structure becomes more accurate and tonal primitives involve a larger number of pixels [7]. However,
the texture is usually related to a hierarchical appearance, and the fine texture can still be seen in the coarse primitive.

Texture analysis approaches can be grouped into four main classes: frequential, structural, statistical, and probabilistic. Frequencional approaches identify the different frequencies composing the texture with or without selection in scale and/or orientation. In particular, wavelet transform has the advantage of the change in spatial resolution to represent textures in the most appropriate scale. To characterize texture in the frequency domain, local energies are generally used as descriptors on each pixel. Nevertheless, texture analysis methods can be used in a transformed space in which the texture is projected. Texture descriptors can also be defined as a texton signature, representing a filter response for neighboring pixels. Structural approaches are designed to extract and describe primitives by: blobs detection, mathematical morphology, multi-thresholding, points of interest detection, filter banks and textons. This is done before defining the rules under which previously characterized primitives are spatially arranged in the image. These rules can either be defined by searching the most probable affine transformation that distorts the ideal texels network [8] or supposed to be random and then defined by heuristics, stochastic processes or Voronoi mosaics [9]. However, statistical approaches do not consider the geometric dimension of texture. They rather compute features by analyzing local distribution of grayscales. Depending on the number of pixels defining the local function, statistical methods may be first order (e.g. autocorrelation function), second order (e.g. co-occurrence matrix) or of higher order (e.g. local binary pattern LBP). Probabilistic approaches consider texture as a random process that can be described by a probability distribution (e.g. Markov Random Field models, MRF) [10]. More recently, with the propagation of wavelets, probabilistic models such as hidden Markov model [5] and the generalized Gaussian density [11], were used to characterize the distribution of the sub-bands coefficients in the wavelet domain. The concepts of textons and LBP are close to that of the co-occurrence matrix. They model the occurrence frequency of a basic pattern. However, for the first two descriptors, the basic pattern is not formed by a pair of pixels as is the case for co-occurrence matrix, but rather by a set of neighboring pixels. Moreover, texton and LBP histograms detect only the presence of patterns in a local neighborhood, while the co-occurrence matrix describes their relative positions in the image. Co-occurrence matrix is thus considered as region descriptors whose effectiveness is not approved in a local context. Besides, the accuracy of probabilistic methods is influenced by the choice of the window size in which the model parameters are estimated. When the neighborhood is too small, model parameters do not capture sufficient information to represent texture, especially for coarse texture. However, including useless neighborhood increases the computational cost while degrading the model performance, especially if the neighborhood contains contours or it is texturally heterogeneous. Thus, we can classify feature extraction methods into local methods, when descriptors are calculated for each pixel, and global ones that model the texture by spatial distribution of grayscales in a larger scale of observation. Global methods illustrate better the spatial aspect of the texture than local ones. For this, we used global methods for the texture description. Then, the segmentation can be realized by grouping pixels, using graph partitioning or active contours. The clustering methods generally reduce the scale of observation to a local neighborhood and transform the image into a feature space separating pixels of different textures. Indeed, given the spatial coherence of most images, there is often a correlation between descriptors, which can be represented by a non-oriented graph, where nodes are the points, and each edge is weighted according to nodes’ similarity. Segmentation can thus be considered as a graph partitioning problem. However, since the texture is not a punctual property and is only detectable for a group of pixels, addressing the segmentation in a punctual manner suffers from the lack of spatial coherence. Thus, several studies estimate the descriptors within a sliding window [12], such that each block represents a node. Nevertheless, windows lead to poor detection of borders. When measures at coarse scales are taken close to the segments’ borders, measures mix information from neighboring segments. Thus, it is difficult to locate the boundaries between segments and coarse measures should be applied only within segments. Hence, many works propose multi-scale methods by hierarchical aggregation [13] or by hierarchical graph partitioning [14], while texture descriptors are calculated from several scales of observation. Active contours provide a coarse texture vision while preserving boundaries by adapting the energy to the texture perception. However, this method is used to extract only foreground objects, which is not the context of our work. Indeed, in the proposed method, coarse measures are defined within regions of the over-segmented image without risk of mixing statistics from several segments or edge smoothing.

3. PROPOSED METHOD

Firstly, an over-segmentation of the image is obtained by a graph-cuts based method ensuring overall perception of the image. Then, a statistical analysis based on the co-occurrence matrix allows to determine the textural nature of each region. Next, a wavelet analysis of regions is performed by adapting wavelet sub-band choice to their natures. The final segmentation is obtained by merging adjacent regions, having the same nature, by exploiting an adjacency graph.

Since the first step prepares to the application of statistics at coarse scale within the segments, priority is given to the detection of boundaries. In fact, obtained segments should be homogeneous without including
contours. For this, we start by detecting boundaries, while combining both brightness and texture information using textons [15], in order to provide an edge map. Furthermore, the Normalized Cuts algorithm [16] is applied to group similar pixels into superpixels [17]. This algorithm depends on a parameter $N$ which represents the number of regions. After having made several tests on various natural images from the Berkeley benchmark [18], we set $N$ to 40 (Fig. 1.b).

Once the image is over-segmented, the second step defines textural descriptors for better discrimination between fine and coarse textures. All statistics performed directly on the image (statistical approaches) or in the frequency domain model the usual qualitative notions of texture. These approaches thus lead to a characterization of the texture: fine or coarse [19]. Geometric approaches and those based on a model seek to understand the overall appearance of texture and describe the texture architecture rather than its nature. In order to choose the combination descriptor/classifier ensuring optimal discrimination between the two classes of textures, we followed an experimental study involving three methods for defining descriptive vectors with different classifiers. The most commonly used statistical methods are: the co-occurrence matrix, the run lengths matrix and the LBP. The first two methods were chosen for this experiment, however, the LBP has been rejected since it describes the texture by a histogram, and classification will then require the definition of a mapping function between LBP histograms [20]. Among frequency methods, we chose the wavelet energy and three classifiers have been tested: LDA (linear discriminant analysis), SVM (Support Vector Machine) and kNN (k nearest neighbors). For the latter, three tests are performed while varying the parameter $k$. The learning dataset includes 61 images from the Brodatz album [21] which contains different textures. We have manually labeled the used textures according to theirs types: 32 textures as "fine" and 29 textures as "coarse". The test dataset contains 31 images grouping both fine and coarse textures. For each test, three descriptive vectors are calculated. The first is formed by Haralick’s attributes derived from the co-occurrence matrix. The second is formed by the energies of wavelet coefficients (EWC) from six sub-bands obtained by decomposition into three levels. The latter is formed by five attributes derived from the run length matrix. We evaluate the results of the different tests according to the error rate (the ratio of the number of misclassified samples by the total number of samples) (Tab. 1). For the three descriptors, recorded errors using the LDA and SVM are lower than those recorded with kNN. This can be justified by the fact that LDA and SVM are based on the detection of differences between observations belonging to different classes (inter-class separation), unlike kNN which is based on maximizing the intra-class similarity. When the features number is important, LDA provides better results than SVM since it applies a principal component analysis (PCA) for the classification. Indeed, LDA seeks, among all possible PCA on the characteristics space, the one providing the optimal discrimination between classes. In fact, a small error is recorded with the co-occurrence matrix attributes whose number is more important than the other two descriptors. SVM provides the same error value that LDA using wavelet energy. Since one attribute is calculated from several sub-bands, the PCA provides no improvement. In conclusion, the best discrimination between the two types of texture is recorded using the co-occurrence matrix and applying LDA classifier. This combination is adopted for learning and automatic determination of the textural nature of regions (Fig. 1.c) such that the co-occurrence matrix is computed from the largest block included in each region.

Next, the discrete wavelet transform (DWT), according to two levels of decomposition, is applied to the largest included block for each region. The multi-scale wavelet allows the representation of textures in the most appropriate sub-bands. Then, characteristics measurements of luminance and texture are defined from the DWT. Indeed, the luminance information is represented by the coefficients distribution of the approximation band. Since the luminance histograms of each region do not follow a particular parametric distribution [22], we chose the kernel density estimator (KDE) which is a non-parametric estimator. For texture descriptors, our interest was focused on the definition of attributes according to the textural nature of regions. In the case of coarse textures, it is more interesting to model the overall structure (i.e. macrotexture). Analysis at a high frequency models the eventual microtextures present in the macrotexture, this information is unnecessary and poor in terms of meaningful information. Therefore, it is preferable to model coarse textures at low-frequency. Fine textures, however, should be seen at a high frequency for a better representation of the microtexture. In fact, fine textures do not present a visible overall structure where their analysis at a low frequency is useless. The main goal of the proposed method is to maximize the representative information of the texture by adapting the choice of the sub-band according to the texture nature. Indeed, the more the frequency is high (resp. low), the more the coefficients represent a local variation around the pixel (resp. an overall change). Thus, $HH1$ and $HH2$ bands are used to characterize fine texture, while coarse textures are calculated from $LH1$, $LH2$, $HL1$ and $HL2$ bands. Then, modeling the coefficients distributions of the sub-bands will be performed by the generalized Gaussian density (GGD) (1), which allows a reliable approximation of wavelet coefficients distribution.

$$p(x; \alpha, \beta) = \frac{\beta}{2\alpha \Gamma(1/\beta)} e^{-\frac{|x|^{\beta}}{2\alpha}},$$

where, $x$ is the variable representing the distribution of wavelet coefficients, $\Gamma$ is the "gamma" function, $\alpha$ is the scale parameter and $\beta$ is the shape parameter. Thus, each sub-band is represented by the two coefficients $\langle \alpha, \beta \rangle$, which are defined using the maximum likelihood estimator.
Finally, similarity luminance and texture measures are used to merge regions. Indeed, the estimated distributions have distinct behaviors. Distances between pairs of GGDs of corresponding sub-bands are effectively represented by the Kullback-Leibler Divergence (KLD) [23]. Thus, we define the distance $d_{\text{kl}}$ between two textures as the average of KLD between pairs of GGDs calculated from all considered sub-bands. However, for non-centered distributions, such as the approximation component distributions, Bhattacharyya distance is more effective than KLD [22]. Two regions with different luminance histograms have necessarily different visual appearance. However, two regions may have similar luminance histograms but different textures. Thus, we begin by calculating Bhattacharyya distance $d_{\text{bh}}$ between luminance histograms. If this distance is greater than a predefined threshold, the two regions will not be merged. In the opposite case, texture features are calculated and the average of the two obtained distances is considered as a criterion for fusion. Thus, two regions $R$ and $R'$ are merged only if they are adjacent, belong to the same texture class and their distances $d_{\text{bh}}(R,R')$ and $d_{\text{bh}}(R',R)$, are almost null (Fig. 1.d).

### 4. RESULTS AND EVALUATION

To evaluate objectively the segmentation results, we tested the suggested method on the Berkeley image database [18] composed of 300 481×321 natural grayscale images. For each image, a set of 4 to 7 segmentation results of reference provided by experts are available to assess the segmentation.

<table>
<thead>
<tr>
<th></th>
<th>LDA</th>
<th>SVM</th>
<th>kNN</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>$k=1$</td>
</tr>
<tr>
<td>EWC</td>
<td>0.22</td>
<td>0.22</td>
<td>0.3</td>
</tr>
<tr>
<td>GLCM</td>
<td>0.06</td>
<td>0.16</td>
<td>0.22</td>
</tr>
<tr>
<td>GLRM</td>
<td>0.35</td>
<td>0.38</td>
<td>0.41</td>
</tr>
</tbody>
</table>

**Tab. 1.** Error rates of different tests of regions classification.

![Original image](image1.png) ![Over-segmented image](image2.png) ![Textural nature: coarse (black) vs. fine (white)](image3.png) ![Final segmentation](image4.png)

**Fig. 1.** Different steps of the proposed segmentation method.

<table>
<thead>
<tr>
<th></th>
<th>PRI</th>
<th>BDE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Humans</td>
<td>0.8754</td>
<td>4.994</td>
</tr>
<tr>
<td>[3]</td>
<td>0.7841</td>
<td>9.9497</td>
</tr>
<tr>
<td>[16]</td>
<td>0.7229</td>
<td>9.6038</td>
</tr>
<tr>
<td>[24]</td>
<td>0.7561</td>
<td>9.4211</td>
</tr>
<tr>
<td>Proposed method</td>
<td>0.7512</td>
<td>9.4199</td>
</tr>
</tbody>
</table>

**Tab. 2.** Performance measures of segmentation methods.

<table>
<thead>
<tr>
<th></th>
<th>With regions’ nature identification</th>
<th>Without regions’ nature identification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Image 1</td>
<td>0.7615</td>
<td>0.7492</td>
</tr>
<tr>
<td>Image 2</td>
<td>0.7412</td>
<td>0.7309</td>
</tr>
<tr>
<td>Image 3</td>
<td>0.7505</td>
<td>0.7488</td>
</tr>
<tr>
<td>Image 4</td>
<td>0.7498</td>
<td>0.7456</td>
</tr>
<tr>
<td>Image r</td>
<td>0.7419</td>
<td>0.7219</td>
</tr>
</tbody>
</table>

**Tab. 3.** Contribution of defining textural natures of regions.

We proposed in this work a two-level segmentation method, which adapts treatment to the region textural nature. Evaluation results prove the effectiveness of the proposed adaptive segmentation method for both coarse and fine textures. However, in the case of purely uniform regions, the merging was not effectively carried out and over-segmentation persists in these regions. Thus, an improved segmentation result can be estimated by the classification of regions into three classes: fine-textured, coarse-textured and uniform regions. For this latter class of regions, descriptive vectors and similarity measures must be calculated based only on the luminance histograms.

**7. CONCLUSION**

We proposed in this work a two-level segmentation method, which adapts treatment to the region textural nature. Evaluation results prove the effectiveness of the proposed adaptive segmentation method for both coarse and fine textures. However, in the case of purely uniform regions, the merging was not effectively carried out and over-segmentation persists in these regions. Thus, an improved segmentation result can be estimated by the classification of regions into three classes: fine-textured, coarse-textured and uniform regions. For this latter class of regions, descriptive vectors and similarity measures must be calculated based only on the luminance histograms.
12. REFERENCES


MIRACLE Session
OUT-OF-SAMPLE CALIBRATION APPROACH FOR CLASSIFICATION METHODS
BASED ON SPECTRAL GRAPH THEORY

Belhomme Philippe¹, Toralba Simon¹, Plancoulaine Benoît¹, Oger Myriam², Bor-Angelier Catherine¹,²

¹PATHIMAGE EA 4656, Normandie Université; UNICAEN, CLCC F. Baclesse, Caen, France
²Pathology department, CLCC F. Baclesse, Caen, France

ABSTRACT

Spectral graph theory (SGT) relies on the study of properties of a graph in relationship to eigenvalues and eigenvectors of Markov matrices. SGT is commonly used for dimensionality reduction, machine learning (classification, clustering) but is very CPU consuming. This is a problem for “out-of-sample” applications aiming to compare new unknown complex objects to an already built database. In this paper, we present a simple method allowing to get around this problem by adding to Markov matrices some “spy points” later used to calibrate the unknown data with a knowledge database. The spectral graph theory method studied here relies on diffusion maps. Results obtained on artificial images composed of texture samples and on virtual slides of follicular lymphomas are serve to explain the general approach.

Index Terms— Spectral graph theory, manifold learning, virtual slide images, out-of-sample extension

1. INTRODUCTION

Research in signal and image analysis is going on for many decades now and is directly linked with the exceptional development of computer technologies. But after all these years, it must be admit that there are not so many real working applications in practice, especially in the medicine area where the expert's eye is still more accurate and faster than many automated systems dealing with large amounts of data. However, reliable automated systems could really help pathologists in their daily work as the number of pathological cases increases as far as the early screening campaigns do. To illustrate this statement with a medical case, for example images of histological tissue sections, the complex structures to be observed, the very large staining differences encountered with preparations providing from different laboratories and even from the same one, the image file size being more and more large for they are now acquired on digital scanners at higher resolutions (a typical virtual slide image (VSI) is commonly 50 Gb now), all these finally assimilate image processing to the analysis of masses of more or less correlated non-linear data. In some previous works dedicated to the development of a computer-aided diagnosis system (CADS) based on image retrieval and classification [1,2], we have used a method coming from spectral graph theory, the diffusion maps (DM) [3], to process VSI split in small parts called ‘patches’. The DM algorithm, in which eigenvalues and eigenvectors of a Markov matrix defining a random walk on the data are computed, allows to both cluster non-linear input data thanks to its inner classification properties preserving local neighborhood relationships, but also to reduce the input data dimensionality in a space (3D concerning this paper) where it is therefore possible to compute euclidean distances between the objects to be analyzed [4,5]. To briefly describe the CADS we are developing, the first step consists in building a knowledge database involving many features extracted from a set of well-known images; this is an ‘off-line’ procedure conducted once. These features are represented by vectors of non-linear data acting as a signature. In a second step, signatures are obtained from unknown images and then compared with those in the database; this is an ‘on-line’ procedure that has to be conducted each time a new image is processed. The diffusion maps technique belongs to unsupervised learning algorithms working only for given training points with no straightforward extension for out-of-sample cases. One of our previous work [6] focused on a way to get around this problem and explained how unknown VSI may be classified by considering the diffusion maps as a learning eigenfunction of a data-dependent kernel. The Nyström formula [7] was thus used to estimate the diffusion coordinates of new data. But even if the Nyström formula approach allowed to drastically limit the computational workload, for the step of eigenvalues and eigenvectors determination has a $O(n^3)$ complexity, the final dimensionality reduction (DimR) result is unfortunately constrained by an intrinsic property of DM. Indeed, with the DM algorithm, the column sum of any eigenvector is always zero and by the way it is directly impossible to compare the projections in a 3D space of two different sets of data points; this is what we will call later the “scaling effect”. Moreover, for a given Markov matrix, the absolute values of eigenvector coordinates are independent from the data order (so the order of rows in the matrix) but their sign do. And as the software we are developing mainly deals with parallel computation, the same data set is never processed in the
same order twice; this second effect is later called the “rotation effect”.
In this paper, we propose a simple but efficient approach allowing to further decrease the computational workload of the out-of-sample extension of spectral graph theory methods, while making it easier to compare new data sets thanks to the use of “spy points”. These “spy points” come from a first set considered as a reference, then are used to fit the other sets in the same 3D space by rotation-scaling of their coordinates. To illustrate our approach in a practical way, we use data sets of feature vectors obtained from image patches extracted in large VSI of follicular lymphomas.

2. MATERIALS
VSI come from histological sections of four different follicular lymphomas stained in the same laboratory according to the Hematoxylin-Eosin-Safron protocol. Images have been acquired by a digital scanner (ScanScope CS; Aperio Technologies) at 20X with a resolution of 0.5 μm per pixel and stored in TIFF 6.0 file format (compression 30%). For this study, histological sections are split in squared areas (also called “patches”) of size 100×100 pixels. Each area is then extracted at plain resolution and stored as an uncompressed TIFF image. Tools developed here are written in Python language with the help of specialized modules (PIL: Python Imaging Library, SciPy and matplotlib).

3. METHODS
3.1. Features extraction
From each patch, statistical parameters based on color and texture information are computed and embedded in a feature vector. They are obtained as global measurements from the RGB color components (reduced to 64 values) and from the two first components (H, E) of the color deconvolution specific to Hematoxylin and Eosin staining [8]. From any given component, the computed features include the mean, median, mode, Skewness and Kurtosis values, the 20%-40%-60%-80% quantiles of its cumulated histogram and 13 Haralick parameters of texture in four directions, that is a total of \( F=305 \) features \((61 \times 5)\) per patch. Considering the sparse numerical range of extracted features, the symmetric Kullback-Leibler distance has been retained for its ability to easily manage such values, while remaining fast to implement. The distance between two vectors \( p_1, p_2 \) of length \( F \) is then given by:

\[
D_{KL}(p_1, p_2) = \frac{1}{2} \sum_{j=1}^{F} \left[ p_{1j} \log \left( \frac{p_{1j}}{p_{2j}} \right) + p_{2j} \log \left( \frac{p_{2j}}{p_{1j}} \right) \right]
\]  

3.2. Dimensionality reduction (DimR)
In any classical CADS, one of the key components is a visualization tool showing relationships between supervised images, stored in a knowledge database, and new images that are presented to the system. Typically, these relations may be expressed as a connected graph in a 3D space where one hopes to find distinctive clusters corresponding to histological types or sub-types. It is therefore mandatory to reduce dimensionality from \( F (F=305 \) in our application) to just 3. With feature vectors containing non linear data, it is not appropriate to perform a principal component analysis (PCA). In papers [3,4] authors have shown that methods based on Spectral Connectivity Analysis (SCA) such as diffusion maps, involving eigenvalues and eigenvectors of a normalized graph Laplacian, are well suited to non linear data. Let \( X=\{x_1, x_2, \ldots, x_n\} \) be a set of \( n \) patches that we assimilate to a fully connected graph \( G \), that means a distance function is computed for each pair \( \{x_i, x_j\} \). A \( n \times n \) kernel \( P \) is obtained from a Gaussian function whose coefficients are given by:

\[
p(x_i, x_j) = \exp \left( -\frac{d(x_i, x_j)^2}{c^2} \right)
\]

with

\[
d(x_i, x_j) = \frac{1}{n} \sum_{x_k \in X} w(x_i, x_k)
\]

and

\[
w(x_i, x_j) = \exp \left( -\frac{D_{KL}(x_i, x_j)}{c} \right)
\]

In fact, \( p(x_i, x_j) \) may be considered as the transition kernel of the Markov chain on \( G \). In other words, \( p(x_i, x_j) \) defines the transition probability for going from \( x_i \) to \( x_j \) in one time step. The eigenvectors \( q_k \) of \( P \), ordered by decreasing positive eigenvalues, give the practical observation space axes. It must be noticed that \( q_k \) is never used since linked to the smallest eigenvalue \( \lambda=1 \) (i.e. the data set mean or trivial solution). The 3D projection is then achieved along \( (q_1, q_2, q_3) \). Choosing \( \varepsilon \) in \( w(x_i, x_j) \) is an empirical task which should permit a moderate decrease of the exponential in equation (4); some works [4] use the median value of all \( D_{KL}(x_i, x_j) \) distances whereas other works [5] use the mean distance obtained in the \( k \) nearest neighbors from a subset of \( X \). We have retained the first solution.
3.3. Out-of-sample extension

3.3.1. Nyström formula

SCA techniques share one major characteristic that is to compute the spectrum of a positive definite kernel. It is known that the eigenvalue decomposition of a matrix $P \in \mathbb{R}^{n,n}$ can be computed no faster than $O(n^3)$; this limits SCA techniques to moderately sized problems [9]. Fortunately the Nyström extension, originally applied for finding numerical solutions of integral equations, can be used to compute eigenvectors and eigenvalues of a sub-matrix formed by $m$ columns of $P$ randomly subsampled and then extended to the remaining $n-m$ columns [7]. Given an $n\times n$ P matrix and an integer $m \leq n$. Let call $P^{(m)}$ the matrix formed by $m$ columns of $P$ that is the graph Laplacian of a set $Y \subset X$ with $|Y| = m$. $Y$ is then a training set. The orthonormal matrix of eigenvectors $U^{(m)}$ and their associated eigenvalues in a diagonal matrix $\Lambda^{(m)}$ are classically obtained from $P^{(m)}$ by solving: $P^{(m)} U^{(m)} = \Lambda^{(m)} U^{(m)}$. This step has to be run once and then may be considered as an ‘offline’ procedure. The Nyström formula allows to obtain the approximate eigenvectors of all the set $X$ by:

$$
\hat{u}_i = \sum_{j=1}^{m} \frac{1}{n} P_{i,j} u_j^{(m)}
$$

(5)

where $\Lambda_{i}^{(m)}$ and $u_i^{(m)}$ are the $i$th diagonal entry and $i$th column of $\Lambda^{(m)}$ and $U^{(m)}$ respectively. $P_{i,j}$ is a $n \times m$ sub-matrix of the complete graph obtained from distances $w(x_i, x_j)$. Its computation is an ‘on-line’ procedure having to be conducted for each new test set ($X$). For a 3D visualization, the second to fourth columns are used (the first one being the trivial solution).

3.3.2. ‘Linear Spy Points’ (LSP) approach

The DimPR procedure achieved thanks to the diffusion maps yields to a set of data points where the column sum of each eigenvector is necessarily equal to 0. Therefore, each new test run provides a set of coordinates that cannot be compared to a previous computation. The approach explored in this “linear spy points” section consists in simply run each test with a set of 3 data points further used as a reference. In the original 3D space $\mathcal{E}$, 3 spy points corresponding to 3 patch images are selected and expressed by $A(x_a, y_a, z_a)$, $B(x_b, y_b, z_b)$ and $C(x_c, y_c, z_c)$. Then, for each new test run, these 3 patches are first added in the directory where the other ‘unknown’ patches stand and the DimPR procedure is processed. The 3 spy points become $A', B', C'$ and are now expressed by $A'(x_{a'}, y_{a'}, z_{a'})$, $B'(x_{b'}, y_{b'}, z_{b'})$ and $C'(x_{c'}, y_{c'}, z_{c'})$ in the new space $\mathcal{E}'$. If we assess that going from $A', B', C'$ to $A, B, C$ is just a linear transform, thus any point in $\mathcal{E}'$ will be expressed from its corresponding element in $\mathcal{E}'$ by a linear vectorial expression as in equation (6).

By using $A, B, C$ and $A', B', C'$ coordinates in (6), the nine matrix elements are easily found by resolving the nine independent linear equations, thus providing the transformation matrix $M$ between $\mathcal{E}'$ and $\mathcal{E}$.

$$
\begin{align*}
X' &= a_x x + b_x y + c_x z \\
Y' &= a_y x + b_y y + c_y z \\
Z' &= a_z x + b_z y + c_z z \\
\begin{pmatrix}
X' \\
Y' \\
Z'
\end{pmatrix} &= \begin{pmatrix}
a_x & b_x & c_x \\
a_y & b_y & c_y \\
a_z & b_z & c_z
\end{pmatrix}
\begin{pmatrix}
x \\
y \\
z
\end{pmatrix}
\tag{6}
\end{align*}
$$

4. RESULTS AND DISCUSSION

4.1. Nyström formula

As the DM method preserves the local proximity between data points, DM+Nyström thus allow to compare a new data set of size $n-m$ with a knowledge database of size $m$. Fundamentally, this is not a real calibration process since the $n$ total points are embedded in a 3D space which is different from both the space where the $m$ points and the $n-m$ points were obtained. Moreover, the space properties depend on $m$. Fig. 1 shows the Nyström extension for $n=m=1000$ patches applied on $m=1000$ references, compared with the raw computation on the 2000 patches (patches come from breast cancer images from the study cited in [6]). Fig. 2 shows the same approach with $n=m=1500$ points and $m=500$. In the second case, it may be noticed that the red point cloud keeps quite the same shape than the black point cloud but is located much more far away from it than with $m=1000$.

4.2. “Linear Spy Points” approach

In order to test if a linear transformation matrix may be used with DM, an artificial image $I$ composed of texture samples coming from the Brodatz database [XX] has been created. At the beginning, $I$ is processed as a whole and 3 spy points are randomly selected under a constraint on a minimal euclidean distance between them in the 3D reference space $\mathcal{E}_{ref}$. $I$ is then split in 2 equal parts ($I_1$ and $I_2$) which are processed independently but with the set of spies, so projected in 2 different 3D spaces $\mathcal{E}'$ and $\mathcal{E}''$. Once the linear transformation matrices are computed and applied on all data points, $\mathcal{E}'$ and $\mathcal{E}''$ are supposed to match $\mathcal{E}_{ref}$. To better represent this approach, spaces $\mathcal{E}$, $\mathcal{E}'$ and $\mathcal{E}_{ref}$ with their own orthonormal coordinate system $(q_1, q_2, q_3)$ are assimilated to RGB color cubes in which each point is associated with a false color. A reliable calibration process would be encountered if the two half colormap obtain on $I_1, I_2$ really match the colormap of $I$. 

31
Figure 1: 2000 true eigenvectors coordinates (black dots) vs estimated coordinates obtained from 1000 points (red dots).

Figure 2: 2000 true eigenvectors coordinates (black dots) vs estimated coordinates obtained from 500 points (red dots).

Figure 3: Linear spy point procedure to calibrate 2 data sets successively computed in a dimensionality reduction scheme by the diffusion maps.

Not finished yet for follicular lymphomas (FL)... The paper length needs also to be reduced to add results and images for FL.

5. CONCLUSION

The linear spy point approach exposed in this paper is a first answer to the problem of out-of-sample extension encountered with dimensionality reduction methods such as the diffusion maps. Unfortunately, for non-linear input data, it cannot be a real solution. The next way that will be explored is to work on a non-linear spy point approach taking into account much more spy points whose coordinates come from the Nyström formula.
6. REFERENCES


MULTI-SCALE DIRECTIONAL FILTERING BASED METHOD FOR FOLLICULAR LYMPHOMA GRADING

Alican Bozkurt, A. Enis Cetin

Bilkent University
Department of Electrical and Electronics Engineering
TR-06800, Bilkent, Ankara, Turkey

ABSTRACT
Follicular Lymphoma (FL) is a group of malignancies of lymphocyte origin that arise from lymph nodes, spleen, and bone marrow in the lymphatic system in most cases and is the second most common non-Hodgkins lymphoma. Characteristic of FL is the presence of follicle center B cells consisting of centrocytes and centroblasts. One common way of grade FL images is an expert manually counting the centroblasts in an image, which is time consuming. In this paper we present a novel multi-scale directional filtering scheme, and utilize it to classify FL images into different grades. Instead of counting the centroblasts individually, we classify the texture formed by centroblasts. We apply our multiscale-directional filtering scheme in 2 scales and along 8 orientations; and use mean and standard deviation of each filter output as features. For classification, we use support vector machines with radial basis function kernel. We map the features into two dimensions using linear discriminant analysis prior to classification. When SVM parameters are optimized, this method achieves 100.0% 10-fold cross validation accuracy on a 270-image dataset.

Index Terms— One, two, three, four, five

1. INTRODUCTION
Follicular Lymphoma (FL) is a group of malignancies of lymphocyte origin that arise from lymph nodes, spleen, and bone marrow in the lymphatic system in most cases and is the second most common non-Hodgkins lymphoma [1]. Characteristic of FL is the presence of a follicular or nodular pattern of growth presented by follicle center B cells consisting of centrocytes and centroblasts. World Health Organization’s (WHO) histological grading process of FL depends on the number of centroblasts counted within representative follicles, resulting in three grades with increasing severity [2]:

Grade 1 0-5 centroblasts (CBs) per high-power field (HPF)
Grade 2 6-15 centroblasts per HPF
Grade 3 More than 15 centroblasts per HPF

Therefore, accurate grading of follicular lymphoma images is of course essential to the optimal choice of treatment. One common way of grade FL images is an expert manually counting the centroblasts in an image, which is time consuming. Recently, Suhre proposed 2-level classification tree using sparsity-smoothed Bayesian classifier, and reported very high accuracies [3].

The dataset provided by [3] is also used in this paper. The dataset consists of 90 images for each of 3 grades of Follicular Lymphoma. In Follicular Lymphoma Grading problem, we aim to grade microscope images according to their centroblast counts. Instead of counting the centroblasts individually, we try to classify the texture formed by centroblasts.

2. DIRECTIONAL FILTERING FRAMEWORK
Directional filtering is a new framework developed in this paper. In this framework, we start with a given filter impulse response \( f_0 \) with filter length \( N \) in one-dimension (1D) and we wish to use \( f_0 \) to filter images in various directions. To do so, we propose to create a set of filters obtained by rotating \( f_0 \) along a set of angles parameterized by \( \theta \).

Instead of rotating \( f_0 \) by bilinear (or cubic) interpolation, we use the following method: For a specific angle \( \theta \), we draw a line \( l \) going through origin \( (l: y = \tan \theta x) \) and determine the coefficients of the rotated filter \( f_\theta(i, j) \) proportional to the length of the line segment within each pixel \( (i, j) \), which is denoted by \( |l_{i,j}| \). For odd \( N \), \( f_0(0) \) is exactly the center of rotation, therefore value of \( f_0(0, 0) \) does not change in \( f_\theta(0, 0) \). Therefore we take line segment in origin pixel \( |l_{0,0}| \) as reference (\( |FG| \) in Figure 1(b)). For \( \theta \leq 45^\circ \), \( |l_{0,0}| = \frac{1}{\sqrt{2}} \), assuming each pixel is of unit side. For each pixel in column \( j \) in the grid, we calculate the \( f_\theta(i, j) \) as \( f_\theta(i, j) = f_0(i) \times \frac{|l_{i,j}|}{|l_{0,0}|} \). This approach is also used in computerized tomography [4].

Calculating the line segment \( |l_{i,j}| \) is straightforward. To rotate the filter for \( \theta \leq 45^\circ \) (which corresponds to \( N \leq 1 \), we place \( f_0 \) to the vertical center of a \( N \times N \) grid, where \( C_X(i, j) \) and \( C_Y(i, j) \) are the coordinates of the center of cell with horizontal index \( i = 0, \ldots, N - 1 \), and vertical index \( j = 0, \ldots, N - 1 \). Then we construct a line \( l \) along the de-
sired direction where the bisector of the line is the exact center of the grid (which is also the center of the filter). For every cell of the grid, we calculate the rotated filter coefficients as: \( f_{\theta}(i, j) = f_0(i, j) \times \max(0, 1 - C_x(i, j) + l(C_x(i, j))) \). To rotate the filter for \( \theta \geq 45^\circ \) we first rotate the filter \( 90^\circ - \theta \) then transpose \( f_{90^\circ - \theta} \) to get \( f_{\theta} \). Note that this method of rotation retains the DC response of the original filter, since \( \sum_{i, j} f_{\theta}(i, j) = \sum_k f_0(k) \).

\[
\begin{align*}
\tan(\theta) & \geq \frac{l}{2} \quad (1) \\
\tan(\theta) & \geq \frac{1}{2C_x(\frac{N-1}{2}, 0) + 1} \quad (2) \\
\frac{1}{N} & \geq \frac{1}{N} \quad (3) \\
\theta & \geq \arctan\left(\frac{l}{2}\right) \quad (4)
\end{align*}
\]

Resulting filters form a directional filter bank as shown in the first row of Table 1. These directional filters are used in a multi-resolution framework for feature extraction. For the first scale, directional images can be extracted by convolving the input image with this filter bank. Mean and standard of these directional images are used as the directional feature values of the image (other statistics, or the image itself can also be used). To obtain direction feature values at lower scales, the original image is low-pass filtered and decimated by a factor of two horizontally and vertically and a low-low sub-image is obtained. Since downsampling is a shift variant process, we also introduce a half-sample delay before downsampling. To implement this, we downsample two shifted versions of input image (corresponding to \( (\Delta x, \Delta y) = \{(0, 0), (1, 1)\} \)), pass two downsampled images from our directional filter bank, and fuse the outputs to construct one output image per filter in directional filter bank. Fusing method used in thesis is simply taking square of images, summing them, and taking the square root of the sum.

A variant of this multi-scale filtering framework uses four shifted versions instead of two (corresponding to \( (\Delta x, \Delta y) = \{(0, 0), (1, 1)\} \)). Although this increases the accuracy by average 1%, it also doubles the computational complexity. This speed vs. accuracy trade-off should be evaluated for potential applications.

The lowpass filter used in downsampling \( f_0 \) can be the corresponding lowpass filter of a wavelet filter bank. If \( f_0 \) is chosen as such, or it can be a simple half-band filter. The low-low sub-image can be filtered by directional filters to obtain the second level directional subimages and corresponding feature values. This process can be repeated several times depending on the nature of input images. The filtering flow diagram is shown in Figure 3.

In our experiments we use directional filters in 3 scales, \( \theta = \{0^\circ, \pm26.56^\circ, \pm45^\circ, \pm63.43^\circ, 90^\circ\} \) and lowpass filter is halfband filter \( f_1 = [0.25, 0.5, 0.25] \). For filter bank we use Kingsbury 8\( h \) order q-shift analysis filter [5]: \( f_0 = [-0.0808, 0.1415, -0.5376, 0.1653, 0.0624, 0 - 0.0248] \)
3. FEATURE EXTRACTION AND CLASSIFICATION

Since images in this dataset are of relatively uniform texture, there is no need to segment the images prior to feature extraction. Also, it is not possible to have 2 different grades of lymphoma in an image, so we give one decision per image. We take input image and feed it to feature extraction algorithms directly after converting to grayscale. After feature extraction, we experiment with dimension reduction. Each feature is classified once without any dimension reduction, once after principal component analysis (PCA) [6], and once after linear discriminant analysis (LDA) [7]. For PCA, dimension is reduced keeping the 99.9% of the cumulative energies of eigenvalues. For LDA, since the maximum number of dimensions is bounded by the number of classes, dimension is reduced to 2D for each feature.

By definition, all multi-dimensional directional feature extraction algorithms output features as a filter response for each scale-direction pair. These filter outputs cannot be used directly as features because they are variant to size of input window, scale, small perturbations in input image such as translation or rotation. In order to make the feature more standardized and more robust to these factors, we use mean and standard deviation of filter outputs for each scale-direction pair. If we use filter outputs of a 3-scale and 6-directional feature extraction algorithm directly, we cannot say anything about the size of our feature vector: it depends on the size of input image. Also, even if we assume the input is just $24 \times 24$ pixels, the feature size would be $6 \times 24 \times 24 + 6 \times 12 \times 12 + 6 \times 6 \times 6 = 42552$, which is too large to classify efficiently. If we use mean and standard deviation of filter outputs for the same algorithm, we easily see that the feature size is $2 \times 3 \times 6 = 36$ regardless of the input. Then we classify the extracted features using Support Vector Machines (SVM) with Radial Basis Function (RBF) as kernel function. The accuracy of the system, as mentioned before, is measured by 10-fold cross validation, which is a very standard method for measuring the accuracy of classification in the literature. In order to find optimal accuracy, we perform a parameter search for $C$ and $\gamma$ parameters of SVM using simple heuristics.

4. RESULTS

We compare the proposed features with various multi-scale directional feature extraction algorithms, such as curvelets [8], contourlets [9], steerable pyramids [10], complex wavelets [11], Gabor filters [12], texture filterbanks [13, 14, 15], and Gray-level co-occurrence matrices [16]. We also compare our results with state of the art [3].

Furthermore, we also performed tests to measure the computational complexity of algorithms. These tests are done on a computer with Intel i7-4700MQ CPU and 16 GB memory. Values presented in Table 3 are average times over 10 runs. It is clear that directional filters are the most efficient among tested algorithms.

5. CONCLUSION

A method for grading follicular lymphoma images, based on a novel multi-scale directional feature extraction framework is proposed. In this framework we draw a line $l$ going through origin $(l : y = \tan \theta x)$ for a specific angle $\theta$, and determine the coefficients of the rotated filter $f_{\theta}(i, j)$ proportional to
the length of the line segment within each pixel \((i, j)\). This new multi-scale directional framework is compared with a number of multi-scale directional image representation methods including the complex wavelet transforms, curvelets, contourlets, gray level co-occurrence matrices, Gabor filters, steerable pyramids, and texton filter banks.

In terms of computational efficiency, directional filter banks are the fastest among all tested methods, extracting features from a 512 × 512 image in 8 directions and three scales in 0.032 seconds.

When features extracted with proposed method are reduced to 2D using linear discriminant analysis, a SVM classifier with optimum parameters achieves 100% 10-fold cross-validation accuracy, surpassing other multi-scale directional feature extraction algorithms and state of art.

### Table 2. 10-fold cross-validation accuracies of each grade, for each feature

<table>
<thead>
<tr>
<th>Feature</th>
<th>Dimension Reduction</th>
<th>Grades</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Grade 1</td>
<td>Grade 2</td>
</tr>
<tr>
<td>CWT</td>
<td>LDA</td>
<td>100.00</td>
<td>100.00</td>
</tr>
<tr>
<td></td>
<td>PCA</td>
<td>76.67</td>
<td>76.40</td>
</tr>
<tr>
<td></td>
<td>None</td>
<td>96.67</td>
<td>98.88</td>
</tr>
<tr>
<td>LM</td>
<td>LDA</td>
<td>98.89</td>
<td>100.00</td>
</tr>
<tr>
<td></td>
<td>PCA</td>
<td>80.00</td>
<td>92.13</td>
</tr>
<tr>
<td></td>
<td>None</td>
<td>95.56</td>
<td>93.26</td>
</tr>
<tr>
<td>MR8</td>
<td>LDA</td>
<td>95.56</td>
<td>94.38</td>
</tr>
<tr>
<td></td>
<td>PCA</td>
<td>65.56</td>
<td>73.03</td>
</tr>
<tr>
<td></td>
<td>None</td>
<td>97.78</td>
<td>92.13</td>
</tr>
<tr>
<td>Contourlet</td>
<td>LDA</td>
<td>100.00</td>
<td>8.99</td>
</tr>
<tr>
<td></td>
<td>PCA</td>
<td>82.22</td>
<td>82.02</td>
</tr>
<tr>
<td></td>
<td>None</td>
<td>87.78</td>
<td>88.76</td>
</tr>
<tr>
<td>Curvelet</td>
<td>LDA</td>
<td>95.56</td>
<td>4.49</td>
</tr>
<tr>
<td></td>
<td>PCA</td>
<td>78.89</td>
<td>93.26</td>
</tr>
<tr>
<td></td>
<td>None</td>
<td>84.44</td>
<td>96.63</td>
</tr>
<tr>
<td>Dir. Fil</td>
<td>LDA</td>
<td>100.00</td>
<td>100.00</td>
</tr>
<tr>
<td></td>
<td>PCA</td>
<td>71.11</td>
<td>80.90</td>
</tr>
<tr>
<td></td>
<td>None</td>
<td>97.78</td>
<td>97.75</td>
</tr>
<tr>
<td>Gabor</td>
<td>LDA</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>PCA</td>
<td>83.33</td>
<td>86.52</td>
</tr>
<tr>
<td></td>
<td>None</td>
<td>86.67</td>
<td>89.89</td>
</tr>
<tr>
<td>GLCM</td>
<td>LDA</td>
<td>88.89</td>
<td>88.76</td>
</tr>
<tr>
<td></td>
<td>PCA</td>
<td>22.22</td>
<td>34.83</td>
</tr>
<tr>
<td></td>
<td>None</td>
<td>85.56</td>
<td>86.52</td>
</tr>
<tr>
<td>Pyramid</td>
<td>LDA</td>
<td>98.89</td>
<td>98.88</td>
</tr>
<tr>
<td></td>
<td>PCA</td>
<td>71.11</td>
<td>94.38</td>
</tr>
<tr>
<td></td>
<td>None</td>
<td>96.67</td>
<td>94.38</td>
</tr>
</tbody>
</table>

### Table 3. Time required for each feature to be extracted from a \(N \times N\) image, for \(N = [512, 1024, 2048]\)

<table>
<thead>
<tr>
<th>Feature</th>
<th>Required time per (N \times N) sample (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N=512</td>
</tr>
<tr>
<td>CWT</td>
<td>0.0615</td>
</tr>
<tr>
<td>Curvelet</td>
<td>0.1863</td>
</tr>
<tr>
<td>Contourlet</td>
<td>0.178</td>
</tr>
<tr>
<td>GLCM</td>
<td>0.3643</td>
</tr>
<tr>
<td>Gabor</td>
<td>0.9655</td>
</tr>
<tr>
<td>Pyramid</td>
<td>0.2714</td>
</tr>
<tr>
<td>Dir. Fil</td>
<td>0.0323</td>
</tr>
<tr>
<td>MR8</td>
<td>0.2083</td>
</tr>
<tr>
<td>LM</td>
<td>2.1083</td>
</tr>
<tr>
<td>RFS</td>
<td>1.6273</td>
</tr>
</tbody>
</table>

### 6. REFERENCES


IMAGE ACQUISITION AND DETECTION OF THE IRIS FOR IRIDIology

Petra Perner

Institute of Computer Vision and applied Computer Sciences, IBal
Kohlenstr. 2, 04107 Leipzig, Germany
pperner@ibai-institut.de, www.ibai-institut.de

ABSTRACT
The iris of a human is not only relevant for biometry; it is also relevant for the prediction and diagnosis of human health. One understands by iris diagnosis (Iridology) the investigation and analysis of the colored part of the eye, the iris, to discover factors which play an important role for the prevention and treatment of illnesses. Up-to-date the iris diagnosis is done manually and is concerned with the know problems, objectivities and reproducibility. An automatic system would pave the way for much wider use of the iris diagnosis for the diagnosis of illness-eses and for the purpose of individual health protection. In this paper we describe the state-of-the-art of the Iridology. Different ways of image acquisition and image preprocessing are explained. We describe the image analysis method for the detection of the iris. This method is based on our novel case-based object recognition and case mining method.

Index Terms— Iris diagnosis, biometry, recognition method, image acquisition

1. INTRODUCTION

The iris of a human is not only relevant for biometry; it is also relevant for the prediction and diagnosis of the health of a human. The later is called iris diagnosis.

One understands by iris diagnosis (Iridology) [1] the investigation and analysis of the colored part of the eye, the iris, to discover factors which play an important role for the prevention and treatment of illnesses, but also for the preservation of an optimum health [2-4].

One of the advantages of the iris diagnosis consists in the fact that it is able to provide a lot about the state of the health of a person. An iris picture can point out a health problem. For example, the fact that more than only one single organ is concerned or that the problem also has an emotional or mental component. Thus one can discuss much better with a patient who must decide between different possibilities of treatment much better or initiate preventive measures before the illness comes to the outbreak.

The iris diagnosis has set up in many countries a complementary-medicine discipline [5-8]. Thanks to her special qualities the iris diagnosis is able to cross some borders which have been established in the last decades with the heavyweight on „evidence-based medicine“ in the medical science. The iris diagnosis is an easy diagnostic method that gets by without big apparatus expenditure and the costs linked with it. It gives to general doctors and also other holistically working therapists a secured diagnosis instrument in the hand.

The iris diagnostic is one of few disciplines which pull up the eye for the diagnosis position. Ophthalmologists already know this; they judge the ocular inside around illnesses to ascertain. Besides, they know some illness signs with which the iris diagnosis works and they are of use already.

In addition, there are investigations with the help of the irises to derive the constitutional type namely the basic disposition of the individual as well as his personality picture. This constitution decides on it for which problems and illnesses an individual is especially susceptible. Moreover, the preserved information about the personality type can be pulled up for the composition of teams.

The aim of the project is to develop an automatic iris image acquisition and diagnosis system. The development of algorithms and procedures for the analysis of the structure inside the iris, the color information and the patterns on the irises that can be automatically used together with the expert's knowledge on illness pictures.

The image acquisition and the preprocessing is describe in Section 2. The case-based object recognition method is described in Section 3. Results on the image acquisition, the preprocessing, and the recognition of the iris are given in Section 4. Finally, we give conclusions in Section 5.

2. IMAGE ACQUISITION

The aim of this work was to develop an easy useable image acquisition unit that allows a person to inspect his iris by himself.

To understand the conditions necessary for iris image acquisition, we first started with the normal microscopic setting of the ophthalmologist. This image acquisition unit consists of an ophthalmologist microscope with a special locking of the head, a white lamp and a digital camera.
CANON AS 710. The magnification of the lens is 450 xs. The light is irradiated into the eye with an angle of 45grad. Note, the eye ball is a moving objects therefore it is not possible to position the light reflex point into a certain part object of the eye.

The image has been taken by a human after having found the right focus level and a sharp image. The resulting digital images are shown in Figure 1a-b.

It is a single shot image not a movie. Such an image acquisition unit cannot be used by human by himself.

![Fig. 1a The image cut out of the microscope](image1)

![Fig. 1b The rectangular cut out of the image in Fig. 1a](image2)

The second choice was a handheld microscope with a ring of four white light lamps and a 400x magnification. The microscope was equipped with a gum eye mus-cle in front of the microscope to ensure safety image acquisition for the person, no foreign light irradiation and a defined image acquisition distance to the object. There is still a manual focus. A sequence of images is taken and the best images of this sequence are cut out for further evaluation. Some sample images of three different subjects are shown in Fig. 2a-c.

![Fig. 2a Iris of Subject1](image3)

![Fig. 2b Iris of Subject2](image4)

![Fig. 2c Iris of Subject3](image5)

The iris is never fully centered in the image. Sometimes we get only part of the iris. Sometimes we have the lid in the image and sometimes not. That is because we cannot lock the eye in front of the camera. The light reflection points are a bit annoyingly. However, that are white spots in the image and they can easily be removed. Unfortunately the area under the white spots is not useable for diagnosis anymore. The setting of the light reflection points into the pupil would be much more preferable but since the eye is moving it is almost impossible unless the observer is waiting for the time were the light reflection points are perfectly located inside the pupil.

3. DETECTION OF THE IRIS BY CASE-BASED OBJECT RECOGNITION

We first need to find a reference point in the image. Our reference point is the pupil of the eye. From the center of the pupil we set out a circular model and match this model against the image contours. Where the image points give the best fit with the model is the boundary of the iris located. Based on the color we can judge how much area the iris will cover in the image. The iris is colored while the surrounding is white or skin-type color.

The model can be a general model such as a circle or different types of models taken from different example images such as described in [10]. We choose the later approach and use case-based object recognition [9] [11] for the detection of the iris.

3.1. CASE-BASED OBJECT RECOGNITION

The heart of our case-based object recognition system is a case base of shapes. These shapes are represented as contour chains. Therefore a case is comprised of a set of contour Points \( S_c = \{ x_c, y_c \}_{\delta \leq c \leq \eta} \) where each point has the grey value 1 and a class label for the shape. Based on this information we can transform the shape from the contour point list into a 2-D image matrix, further called case image. The case base is filled up for the actual application by shapes that we learnt based on our novel case acquisition and case mining method [10]. An index over the case base should allow us to find the closest case among the numerous cases in short time. A case image is matched against the image by constructing an image pyramid from the actual image and the case image. This allows us to reduce the computation time while matching.

![Figure 3. Architecture of a Case-Based Object Recognition System](image6)

Beginning with the highest level of the image pyramid the scores are calculated and the areas of interest are marked. The area of interest is the area where an object can be detected. This area is recursively used for further matching by going downward the levels of the image pyramid. Finally the closest match is given to the output. Depending on the actual value of the similarity measure the next level of the
index structure is selected and the process repeats until a final node is reached. The architecture of our case-based object recognition system is shown in Figure 3.

3.2. Case Representation

In general, we can distinguish between three different case representations according to the pixels that are used for matching:

1. Region of Interest (ROI): A region of interest ROI is obtained by taking a cut-out from the original image. All pixels of the obtained image matrix are used as case pixels regardless if they are object or background pixels.

2. Object Case: In the image matrix shown in Figure 2b are only those pixels as case points that lie inside and at the contour of the object. In this case the shape and the inner structure of the object are taken into consideration.

3. Contour Case: Only pixels that lie on the contour of an object are taken as case points. Thus only the shape of the object of interest is matched.

The kind of representation used for the cases depends on the special image quality the matcher should detect. Our goal is to recognize the fungi spores. To use an object case would not be sufficient for our application since the appearance of the structure inside the objects is very diverse and because of that it would result in a case base where for each case is stored an object. The only representation that gives us a more generalized view to the objects is the shape. Therefore we use a contour case as case representation.

Note that an object might appear in an image with different size and under a different rotation angle and on various locations in an image. But it is still the same object. It makes no sense to store all these identical but different sized and rotated objects in the case base. Rather there should be stored a unit object with the origin coordinates x0 and y0 that can be translated, resized and rotated during the matching process. Therefore the case pixels $p_k = (t_k, u_k)^T$ and the direction vectors $m_k = (v, w)^T$ have to be transformed with a matrix A to:

$$p'_k = A \cdot p_k$$
$$m'_k = A \cdot m_k$$

If $\phi$ denotes the angle of rotation and $r$ the scaling factor the matrix may look like the following:

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} = \begin{pmatrix} r \cos \phi & -r \sin \phi \\ r \sin \phi & r \cos \phi \end{pmatrix}$$

Subheadings should appear in lower case (initial word capitalized) in boldface. They should start at the left margin on a separate line.

3.3. Image Representation

Since we are looking for the contour of the object which is the boundary between the background and the object and which is usually an area of high grey level chance we are representing the image by the edges. The edges can be represented by the gradient of the pixels. In order to determine the gradient, first the direction vector of a pixel at the position is calculated from the grey level matrix. The direction vector indicates the change of the grey value in vertical and horizontal direction respectively. The length of this vector is equal to the gradient and it is commonly determined from the direction vector through the following formula:

$$\|\nabla f(x,y)\| = \sqrt{(\Delta x)^2 + (\Delta y)^2}$$

(3)

Due to the discreteness of the grey level matrix which represents the grey value function only in some well-chosen points, the direction vectors cannot be calculated by the known analytic derivation formula. Therefore many operators were developed that allow us to determine the direction vectors from the grey level matrix. We used the Sobel operator. The corresponding edge image is obtained by applying such an operator to the grey level image. After that the pixels represent the gradient instead of the grey level value. Besides that the direction vectors for each pixel are stored. This representation is calculated for the case and the actual image before the matching.

3.4. Similarity Measure based on the Dot Product

As we have pointed out above the calculation of the Hausdorff distance is more costly than the calculation of the cross correlation. While we have to search for correspondences between case and image pixels in case of using the Hausdorff distance, we evaluate the image pixels that coincidence with the case pixels by using the cross correlation. On the other hand we are interested in matching oriented edge pixels which Olson and Huttenlocher [15] described for the Hausdorff distance. Therefore we propose a similarity measure based on the cross correlation and by using the direction vectors of an image. This approach requires the calculation of the dot product between each direction vector of the case $m_k = (v_k, w_k)^T$ and the corresponding image vector $i_k = (d_k, e_k)^T$:

$$s_i = \frac{1}{n} \sum_{i=1}^{n} m_k \cdot i_k = \frac{1}{n} \sum_{i=1}^{n} \left( m_k \cdot i_k \right) = \frac{1}{n} \sum_{i=1}^{n} \left( v_k \cdot d_i + w_k \cdot e_i \right)$$

41
with \( k = 1, \ldots, n \) case pixels.

The similarity measure of Equation (4) is influenced by the length of the vector. That means that \( s \| \) is influenced by the contrast in the image and the case. In order to remove the contrast, the direction vectors are normalized to the length one by dividing them through their gradient:

\[
x_2 = \frac{1}{n} \sum_{i=1}^{n} \frac{m_{ik} \cdot \hat{i}_{ik}}{\| m_{ik} \| \| \hat{i}_{ik} \|}
\]

In this respect the similarity measure differs from the normalized cross correlation (NCC). The NCC normalizes each pixel value by the expected mean of all values of the considered pixels. Therefore the normalized cross correlation is sensitive to nonlinear illumination changes while our method is not because it takes only into account the angle between two corresponding direction vectors. The values of \( x_2 \) can range from -1 to 1. If \( x_2 \) is equal to one then all vectors in the case and the corresponding image vectors have the same direction. If \( x_2 \) is equal to -1 then all the image vectors have exactly opposite directions as the case vectors. That means that only the contrast between the case and the image is changed.

The above described global contrast changes can be excluded by computing the absolute value of \( s_2 \):

\[
x_3 = \frac{1}{n} \sum_{i=1}^{n} \frac{m_{ik} \cdot \hat{i}_{ik}}{\| m_{ik} \| \| \hat{i}_{ik} \|}
\]

However in case half of the vectors have the same contrast and the other half have the opposite contrast than the similarity based on \( s_3 \) is zero. That might not be preferable for cases where objects are touching. To avoid this we calculate the similarity based on \( s_4 \):

\[
x_4 = \frac{1}{n} \sum_{i=1}^{n} \frac{m_{ik} \cdot \hat{i}_{ik}}{\| m_{ik} \| \| \hat{i}_{ik} \|}
\]

4. Results

The original images (see Fig. 4a-c) are transformed into a grey level image. The thresholded image used to find the pupil and the center of mass inside the pupil is shown in Fig. 5 a-c for three subjects. Around the center of mass is set the model and then object detection is started. The edge filtered image by Sobel-phase operator is shown in Fig. 6 a-c for the three subjects. The resulting image after applying the case-based object matcher is shown in Fig. 7 a-c.

Twenty subjects participated in this study. From each of the subject were taken the iris with the handheld microscope. Four different models were inserted into the case base of the case-based matcher ranging from circular to ellipse-like model. These ellipse-like models are flattened at the bottom and the top as how it appears on the normal eye. Each of the images was preprocessed in the same way as the three images described above.

The iris could be detected by our method in all of the twenty cases. However, due to occlusion not the full iris could be seen in the image and part of the detected object needs to get removed afterwards. Since it is mostly skin and hair that occlude the iris this removal can be easily done by the color information.
5. CONCLUSION

In this paper we have presented our work on image acquisition, preprocessing and iris recognition for Iridology. We have used a handheld microscope with a ring of white lamps and equipped with a gum eye muscle in front of the microscope to acquire the iris. From the image sequence is taken the image that shows most of the iris and is sharp enough for further analysis. The iris is detected with our case-based object recognition methods using different models from circular to ellipse-like models. We were able to recognize the iris of our entire subjects with good quality. Occluded areas could have been taken out based on the color information and they are not used for further evaluation. Further work will be done for image interpretation according to the knowledge of Iridology and further improvement of image acquisition.

References


DETECTION OF CENTROBLASTS IN H&E STAINED IMAGES OF FOLLICULAR LYMPHOMA

Emmanouil Michail1, Evgenios N. Kornaropoulos1, Kosmas Dimitropoulos1, Triantafyllia Koletsas2, Ioannis Kostopoulos2, Nikos Grammalidis1

1Information Technologies Institute, Centre for Research and Technology Hellas, Thessaloniki, Greece, 2Pathology Department, Medical School, Aristotle Univ. of Thessaloniki, Thessaloniki, Greece. michem@iti.gr, ekornarop@iti.gr, dimitrop@iti.gr, koletsa@med.auth.gr, kostop@med.auth.gr, ngramm@iti.gr,

ABSTRACT

This paper presents a complete framework for automatic detection of malignant cells in microscopic images acquired from tissue biopsies of follicular lymphoma. After preprocessing to remove noise and suppress small details, images are segmented by using intensity thresholding, in order to detect the cell nuclei. Subsequently, touching cells are being separated using Expectation Maximization algorithm. Candidate centroblasts are then selected for classification by using size, shape and intensity histogram criteria. Finally, candidates are classified by using a Linear Discriminant Analysis classifier. The application of the methodology in a generated dataset of microscopic images, stained with Hematoxilin and Eosin, showed promising results by detecting in average 82.58% of the annotated malignant cells.

Index Terms— Follicular lymphoma detection, H&E stained images, centroblasts, cell segmentation, touching-cell splitting

1. INTRODUCTION

Follicular lymphoma (FL) is the second most common lymphoma diagnosed in the United States and Western Europe. It accounts for about 20% of all non-Hodgkin lymphomas [1] and mainly affects lymph nodes. When the affected lymph nodes are seen under the microscope, they show rounded structures called "follicles", which explains the term 'follicular'. The neoplastic cells consist of a mixture of centrocytes which are small- to medium-sized cells and centroblasts (CBs) which are large cells. The World Health Organization Classification has adopted grading from 1 to 3 based on the number of CBs counted per high power field (HPF) defined as 0.159 mm²: Grade I with 0-5 CBs/HPF, Grade II with 6-15 CBs/HPF and Grade III with more than 15 CBs/HPF [1].

CB count is performed manually by the pathologist using an optical microscope and Hematoxilin and Eosin (H&E) stained tissue sections. An average CB number is calculated over ten random HPFs. Manual histological grading of FL is a time consuming process and requires considerable effort and extensive training. Furthermore, since this method uses only ten HPFs for CB count, results for specimens with high tumor heterogeneity are vulnerable to sampling bias. This may lead to inappropriate clinical decisions on timing and type of therapy [2]. Hence, there is a need for a computer assisted method which will improve reproducibility and reliability of the grading process and will reduce the time needed for diagnosis.

Computer-aided diagnosis (CAD) has been reported to be beneficial in classifying tissue subtypes associated with various grades of FL. The main steps of automatic FL grading are usually the following. The HPF image is segmented into its basic cytological components in order to extract the cell nuclei. Segmentation is usually performed using algorithms like k-means [2, 3], Expectation Maximization [4], Otsu thresholding, [5], Graph Cuts [6] etc. In many cases, segmentation algorithms tend to merge together nuclei that are too close to each other. This is usually referred to as “touching cells”. Several algorithms have been proposed for touching-cell splitting, like watershed segmentation [7], radial-symmetry interest points [8], h-minima [9], active contours [10], concave points [11], ellipse/curve fitting [12], and Graph Cuts [13]. The identified cells are subsequently classified into CB and non-CB cells by extracting morphological and topological features from the cell regions [4], texture features [2, 5], as well as graph-based features [3]. Principal Component Analysis (PCA) is often employed to identify the most discriminative features.

This paper presents a complete framework for automatic CB detection in H&E stained images acquired from tissue biopsies of FL. This framework was developed to address the special characteristics of the images used in this paper. Specifically, the images were obtained from 1 to 1.5 μm thick tissue sections. The advantage of the small thickness of tissue sections is the detailed depiction of the nuclei (especially the large ones), as seen in Figure 1. Thus, contrary to microscopic images used in previous studies, nucleoli in these images are more distinguishable. However, the disadvantage is that cell segmentation and nuclei
detection is becoming difficult, since often their interior has the same color and texture as their exterior.

Figure 1: Details from two different HPF images, used for testing the methodology. CBs are marked with a circle.

2. METHODOLOGY

The complete image analysis scheme applied in this study is illustrated in Figure 2 and consists of five main steps: image pre-processing, image segmentation, touching-cell splitting, selection of candidate CBs and classification. These steps are described in the following subsections.

2.1 Pre-processing

The algorithm uses as input HPF images of FL stained with H&E. In order to remove noise from the image and suppress small details, input images are first converted to grayscale and filtered using a Gaussian filter with a 3x3 kernel. Additionally, in order to facilitate the detection of nuclei, differences between nuclear membrane and background are enhanced, by applying histogram equalization to the filtered image.

2.2 Image segmentation

There are five major cytological components in the FL tissue: nuclei, cytoplasm, extra-cellular material, red blood cells (RBCs) and background regions [2]. Nuclei and cytoplasm regions are usually dyed with hues of blue and purple. However, in cases of large cells, nuclei also contain white components, which hinder their detection. Extra-cellular material is dyed with hues of pink and red blood cells (RBCs) are dyed with hues of red. In addition to these components, there are also white background regions that do not correspond to any tissue component.

For the identification and elimination of RBCs, an RBC mask is generated using the following threshold

\[ I_{\text{red}}/(I_{\text{red}}+I_{\text{blue}}+I_{\text{green}}) > T_{\text{RBC}} \]

where \( I_{\text{red}}, I_{\text{blue}}, I_{\text{green}} \) are pixel intensity values corresponding to red, blue and green channel respectively. The threshold is empirically set to 0.37.

After the elimination of RBC pixels, Otsu thresholding [14] is applied to the remaining pixels in the grayscale image in order to segment nuclei (dark) from extra-cellular material and background regions (bright).

As a post-processing step, connected component labelling is used to identify individual objects and an area threshold (of 10 pixels) is used to remove very small objects.

Due to the transparency of large cells, their interior appears hollow after Otsu thresholding. Furthermore, in some cases, the perimeter of the cells remains open after segmentation (open cells) and a simple hole-filling operation is not able to sufficiently fill the inner area of the cell. Figure 3 presents a characteristic case of an open cell. In order to address this issue, an additional post-processing procedure consisting of three steps is applied to each object in the image, which is illustrated in Figure 4. Specifically, each object is isolated and subjected to dilation with a diamond-shaped structuring element of radius \( r=1 \). Subsequently, a hole-filling operation is applied to the object and, finally, the object is subjected to erosion with the same structuring element. After this procedure, the resulting object replaces the original one in the image.

Figure 3: A detail of an HPF image containing a centroblast: a) initial image, b) grayscale image, c) image after applying Gaussian filtering, d) after histogram equalization, e) after Otsu thresholding.

1 In the rest of the paper nuclei are also referred to as “cells”
Figure 4: Post-processing procedure for open cells: a) the cell is extracted from the image, b) image dilation, c) hole-filling, d) image erosion.

2.3 Touching-cell splitting

In order to address the issue of touching-cells after segmentation, a cell-splitting algorithm is proposed based on Gaussian mixture modeling.

Initially, connected component labeling is applied and all cells larger than a threshold $T_{CB}$ are used as candidate touching cells. $T_{CB}$ is set to the minimum size of CBs as computed from an annotated database, consisting of CBs and non-CBs (as described in experimental results).

Expectation Maximization (EM) algorithm [15] is used to estimate a) the order of the mixture by using the minimum description length (MDL) criterion [16] and b) the parameters of the Gaussian mixture. MDL works by attempting to find the model order which minimizes the number of bits that would be required to code both the input data samples and parameters of the Gaussian mixture. Data samples consist of the pixels coordinates of each candidate touching cell; thus only spatial information is used. We observed that, when all the pixels of the cell are provided to the algorithm, MDL tends to produce a large number of clusters. Thus, subsampling is applied to the pixels according to their distance from the perimeter. An empirical threshold is set for this distance corresponding to 65% of the maximum distance. Only pixels with distance larger than this threshold are used as data samples. Despite subsampling, in some cases MDL still produces ubiquitous numbers of clusters, with their centers being too close to each other (e.g. 1 pixel). In order to address this problem, clusters that are too close to each other are being merged and the centroid of their centers is used as the center of the final cluster. The threshold for the merging is empirically set to 80% of the maximum distance from the perimeter. Figure 5 demonstrates the application of the cell-splitting algorithm on an object consisting of two touching cells.

2.4 Selection of candidate CBs

At this step candidate CB cells are being selected, based on their size, shape and intensity histogram. Initially, cells with area smaller than $T_{CB}$ are excluded from further processing steps. Regarding the shape, we used the fact that nuclei of CB cells are usually round or oval [17]; thus cells with irregular or elongated shape are rejected. For this reason, the best fitting ellipse is estimated using the Orthogonal Distance Regression (ODR) algorithm [18]. Two criteria are used regarding the shape: a) the aspect ratio (major to minor axis ratio) and ellipse residual (average geometric distance of the pixels in the perimeter from the ellipse). Two thresholds were defined based on the annotated CB training set. The thresholds for aspect ratio and ellipse residual were set to 2.8 and 1.8 respectively. Cells with values larger than these thresholds are discarded. In order to exclude some small dark non-CB cells, the mean value of the grayscale histogram was computed over the CB training set and a threshold was set to 135.7. Cells with mean histogram value smaller than this threshold are excluded. All remaining cells are provided to the classification step as candidate CBs.

Figure 5 Cell splitting on touching cells: a) initial image, b) image after segmentation, c) the Euclidean distance of each pixel from the perimeter, d) the most distant pixels from the perimeter, e) the centers of connected cells, as computed by the MDL, f) splitted cells according to the result of EM.

2.5 Classification

For the classification between CBs and non-CBs, the annotated database of CBs and non-CBs was used as a training set. An $mn \times n$ matrix $A$ was defined having as elements the intensity values of all images of the training set, where $n$ is the total number of training images (CBs and non-CBs) and $m$ is the number of pixels in each image (in our experiments: $71 \times 71$).

$$A = \begin{bmatrix} p_{11} & \cdots & p_{ni} \\ \vdots & \ddots & \vdots \\ p_{1m} & \cdots & p_{nm} \end{bmatrix}$$

Then, in order to remove the redundancy from the image set, Singular Value Decomposition (SVD) is used. According to SVD, $A$ can be written as:

$$A = U \Sigma V^T$$

where $U$ is an $mn \times mn$ matrix and $V$ is a $n \times n$ matrix, representing the left and right eigenvectors of $A$ respectively. Moreover, $\Sigma$ corresponds to the eigenvalues of $A$. The left eigenvectors are an orthogonal basis for the column space of $A$, i.e. the “image space of CBs and non-
CBs". The eigenvector with the highest eigenvalue points to the highest variance among the images. We assume that the discriminative features of CBs/non-CBs will be revealed on the directions pointing to the highest variance among the images. Therefore, eigenvectors are ordered in a sequence of descending eigenvalues. The projection of training images onto a subspace which spans by only the first few eigenvectors will well characterize the cells. In our case 47 eigenvectors were used, which was the optimal number as indicated by the training results of the classifier.

Each candidate CB in the testing image is subjected to classification according to the following procedure: The centroid of the cell is computed and a 71x71 region of the initial image is kept around this centroid. The projected training and testing images are used as input to a Linear Discriminant Analysis (LDA) classifier, in order to classify them into one of two classes (CBs and non-CBs).

In order to compute the optimum number of eigenvectors, that would reveal the most discriminative features of CB and non-CB images, we trained the classifier based on “Hold-out K-folds” cross-validation approach [19]. For this reason, the images of CBs and non-CBs were randomly divided K times into training (80%) and validation (20%) sets. The training set was further processed in order to obtain the optimal number of eigenvectors. Specifically, after examining every number of eigenvectors from 1 to 50, the optimal number of eigenvectors was selected to be the smallest one through which the best classification results were calculated. After computing the optimal number of eigenvectors at each of the K iterations, classifier’s accuracy was validated by using the validation set, and the final optimal number of eigenvectors was regarded as the one derived by the iteration that produced the best classification results. Empirically, K was set to 10.

### 3. EXPERIMENTAL RESULTS

The methodology described in previous section was tested on three 40× microscopic HPF images derived from tissue biopsies of grade II FL, stained with H&E. Images were acquired at the Pathology Department of Medical School of Aristotle University of Thessaloniki, Greece. Tissue sections were sliced at a thickness of 1 to 1.5. They were scanned using Nikon DN100 digital network camera and were inspected by two medical experts, in order to identify the number of CBs in each image. The average number of CBs in each image was 10.

In addition, a training set containing cropped images of CBs and non-CBs was generated for the classification process. Specifically, nine HPF images of FL, stained with H&E, were scanned by using the same procedure as for the testing images. Subsequently, they were examined by medical experts in order to mark CBs on them. By using these markings, a set of cropped images of CB cells was created. Each cropped image contains the CB cell at its center and is of size 71x71 pixels. Similarly, a second set of images of size 71x71 pixels containing only non-CBs was created. In total, a training set of 70 images of CBs and 110 images of non-CBs was used.

The overall efficiency of the algorithm was assessed by comparing CBs that were annotated by doctors with the detected CBs. 82.58% of the annotated CBs were successfully detected on average in the three images. The disadvantage of the methodology is that it produces a large number of false positives. Specifically, the average number of false positives in the three images was 50. Analytical results are shown in Table I.

<table>
<thead>
<tr>
<th>TABLE I. IMAGE ANALYSIS RESULTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Image 1</td>
</tr>
<tr>
<td>Annotated CBs</td>
</tr>
<tr>
<td>Detected CBs</td>
</tr>
<tr>
<td>True positives</td>
</tr>
<tr>
<td>False negatives</td>
</tr>
<tr>
<td>False positives</td>
</tr>
<tr>
<td>Correct detection rate</td>
</tr>
<tr>
<td><em>Number of TPs divided by the number of annotated CBs.</em></td>
</tr>
</tbody>
</table>

The large number of false positives might be partially explained by the resemblance of certain types of large cells, like endothelial and dendritic cells to CBs. As already mentioned, the transparency of large cells poses a problem on the detection of nuclei, but also provides a better description of the nuclear area. In this paper the effort was focused mainly on the accurate extraction of cells. Future work should focus on the development of a texture descriptor that will detect differences between the different kinds of large cells.

### 4. CONCLUSIONS AND FUTURE WORK

A complete methodology was proposed in this paper for detection of CBs in H&E stained microscopic images of FL. The methodology addresses the special characteristics of the images used and specifically, the transparency of the nuclei due to the small thickness of tissue sections. Intensity thresholding has been used for the segmentation of images into their cytological components. Additionally, Expectation Maximization algorithm is being used for the separation of touching cells. Candidate CBs are selected by using size, shape and intensity histogram criteria. Finally, candidates are classified into CBs and non-CBs by using a Linear Discriminant Analysis classifier. An average number of 82.58% of the annotated CBs was detected in three HPF images. However, the algorithm also produces a large number of false positives and future work should focus on their elimination by investigating textural differences between CBs and large non-CB cells.
5. REFERENCES


Regular Session 2
REAL-TIME EMBEDDED SYSTEM FOR ROAD-CROSSING ASSISTANCE

Adi Perry and Nahum Kiryati
School of Electrical Engineering
Tel Aviv University
Israel

ABSTRACT
We present a low-cost embedded system to detect and alert pedestrians with visual impairments of approaching vehicles in streets with no traffic control. The system can be mounted on existing street infrastructure resulting with simple and fast integration. The complex incoming video signal is transformed to a one-dimensional signal, that is forwarded to a decision module. We achieve high probability-of-detection and a low false alarm rate, while providing sufficiently early warning to the pedestrian. We are developing a prototype that uses low-cost and low-power hardware that can obtain energy from a compact solar panel. The suggested system is a step towards safer road crossing for all pedestrians, especially those who are visually impaired.

Index Terms— Computer Vision, Optic Flow, Road Crossing, Blind, Visually Impaired, Embedded, Real-Time, Low-Power, Resource-Limited

1. INTRODUCTION
For visually impaired pedestrians, crossing a street in the absence of traffic control is a real challenge. Because drivers often do not reliably yield to pedestrians, even those who are clearly visual impaired (like those using a white cane), pedestrians need to cross in traffic gaps [1]. To identify traffic gaps, visually impaired pedestrians rely on hearing. The common strategy is “cross when quiet” [2], but due to reduced vehicle noise, it is becoming less reliable.

US data [3] indicates that the average crossing speed is 4 feet per second. A standard urban one-way single-lane street with two shoulders is 18 feet wide and takes 4.5 seconds to cross. A standard two-way street or one-way with two lanes is 28 feet wide, taking 7 seconds to cross. Thus, the length of the gap in traffic must be greater than these figures, and for safe crossing it is reasonable to require about 10 seconds of gap in traffic.

Pun et al [4] reviewed the field of assistive devices, especially those that use image and video processing to convert visual data into an alternate modality, such as auditory or haptic (or a combination of both) which can be delivered to the blind person. Several systems help locating and identifying points of interest [5] and road crossings [6, 7]. Other devices provide auditory or other indications regarding traffic light status [8, 9]. “Smart canes”, employing various types of sensors, have been suggested [10, 11, 12], but are intended primarily for short-range obstacle detection and navigation. All these systems do not address the fundamental difficulty of crossing a road in the absence of traffic control.

An ideal solution to the road-crossing problem must reliably analyze traffic gaps, should be sufficiently robust to operate in all weather conditions, day and night, and has to be easy and cheap to install, operate and maintain. This work is intended to be a first step towards meeting this challenge.

In this paper, we present a low-cost system to detect and alert pedestrians of vehicles approaching a crosswalk. When fully developed, it will be possible to mount the system on existing street infrastructure, such as traffic signs or illumination poles. The system adapts to the scene, thus minimizing installation and maintenance effort. It can detect approaching vehicles about 10 seconds before they reach the road crossing, thus indicating traffic gaps that are sufficient for safe crossing.

2. HARDWARE PLATFORM & INSTALLATION
The system includes a camera capturing video that is processed by compact hardware. The camera is aimed towards the approaching traffic, as illustrated in Fig. 1.

The hardware platform can include a compact solar cell and a battery, to allow autonomous operation without reliance on the electric power grid. This can enhance the applicabil-

Fig. 1: System installation
ity of the system, and reduce its installation and maintenance costs. A block diagram is shown in Fig. 2.

Indication of incoming traffic, or lack thereof, can be delivered to the user as an audio or tactile signal, or via a local networking interface (such as Bluetooth or WLAN) to a smartphone or a dedicated receiver.

We have developed the system on a standard laptop PC, with a Logitech QuickCam 9000 WebCam. Targeting a low-cost solution, we have also implemented the algorithm on BeagleBoard xM hardware, which is an open-source platform consisting of TI’s OMAP 3530 (ARM processor + DSP) and suits rapid prototyping. Nowadays, since Android-capable hardware is inexpensive and popular, we are also porting the system to Android OS.

3. ALGORITHM

We examined various approaches for efficient detection and evaluation of approaching traffic on a resource-limited platform. A generic approach that comes to mind is detecting any new object appearing in the field of view, tracking it, analyzing its motion and estimating its Time to Contact (TTC) as the basis for issuing an alert. Lee [13] observed that in a simple but typical case, the TTC can be estimated based on the incoming object’s expansion rate, measured in the image plane. A newer approach for TTC estimation [14] treats the approaching vehicle as a moving plane.

In our application, the field of view (FOV) of the camera should be sufficiently wide to capture nearby vehicles (possibly leaving a parking spot). However, most approaching vehicles first appear as tiny spots near the vanishing point (VP)\(^1\). Given the limited pixel-count of reasonably-priced video cameras, the wide FOV implies that the spatial resolution near the vanishing point is quite low. These constraints imply that the main issue is early detection of the incoming vehicle, meeting the advance warning time requirement. Therefore, in most cases precise TTC estimation is irrelevant. Also note that the approximations on which TTC algorithms rely do not hold at the most interesting moment, when the vehicle is seen as a tiny, far-away spot.

We convert the space-time video processing problem to a 1-D signal analysis problem, by computing a scalar motion measure, referred to as Activity, reflecting the entire relevant motion in the scene. Moving objects towards the camera induce pulses in the Activity signal, such that a significant rising Activity slope suggests an approaching vehicle. To accomplish early detection of a true rising Activity slope with few false alarms, the noise level in the Activity signal should be as small as possible.

3.1. Activity Estimation

Brightness patterns in the image move as the objects that give rise to them move in the scene, leading to Optical Flow. We estimate the optical flow using a computationally-efficient coarse to fine version of the Lucas-Kanade [15] algorithm. The estimation errors typical to the Lucas-Kanade method can be tolerated in our application; as will be seen, the Activity signal, derived from the optical flow, is an integral measure in which these errors are spatially averaged over parts of the image domain.

Parts of the optical flow field are associated with riskposing approaching vehicles. Other parts might reflect distancing traffic on another lane, pedestrians crossing the road, and movements due to wind or camera vibrations. The proposed activity measure \(A(t)\) is obtained by projecting the optical flow field \(\vec{u}(x, y, t)\) onto a projection map. The projection map \(\vec{m}(x, y)\) is a vector field supported on image regions corresponding to lanes carrying traffic towards the camera, each vector representing the local direction of approaching traffic. Formally,

\[
A(t) = \sum_{x, y} \vec{m}(x, y) \cdot \vec{u}(x, y, t) \quad (1)
\]

The scalar, time-dependent Activity is fast to compute, and reflects the entire risk-inducing motion in the scene.

Assuming a one-way road, the projection map can be automatically generated by temporal averaging of the optical flow over a training period, see Fig. 3. The averaging process cancels the randomly oriented contributions that are due to vibrations, wind and similar phenomena, while highlighting the consistent, dominant, risk-inducing traffic motion directions in the road area alone. Slight adaptation of this procedure is necessary for dealing with two-way roads, where consistent distancing traffic is also expected within the visual field of the camera.

The magnitude of the projection-map vectors can be modified to highlight certain traffic elements. In particular, early detection of approaching vehicles can be facilitated by emphasizing projection map vectors in the vicinity of the vanishing point. An example is shown in Fig. 4. Note that the vanishing point can be automatically detected by back-tracking the projection-map vectors until their source is reached.

---

\(^{1}\)We use the term vanishing point in a loose sense, including the case of a curved road.
3.2 Detecting Approaching Vehicles

In certain applications, the raw Activity signal can be delivered to the pedestrian in analog form, leaving the actual decision regarding road crossing safety in the human domain. However, in most cases we wish to provide the pedestrian with a binary signal, suggesting either that traffic is approaching (‘TRAFFIC’ state) or that a sufficient traffic gap occurs at that moment (‘GAP’ state).

Despite the substantial SNR in the Activity signal near Activity peaks, early detection of an approaching vehicle, at the early rising stage of the corresponding activity pulse, when the SNR is poor, is not easy. Note that the pulse shape and magnitude are generally not known in advance, as they depend on the particular vehicle characteristics, as well as on the specific scene structure and viewing conditions.

Robust detection cannot be accomplished by a simple thresholding operation, hence the Activity signal should be examined within a sliding temporal window. Even when all the sliding window data is used to classify the current Activity sample, and regardless of the specific classifier employed, direct binary classification to TRAFFIC or GAP states is not sufficiently stable. We therefore admit one or more additional intermediate states, reflecting temporary hypotheses of limited certainty. The output of the classifier is forwarded to a state machine with a binary output - TRAFFIC or GAP. The state machine ensures consistency and stability of the decision forwarded to the pedestrian.

For the initial classification of the Activity sample within the sliding window, we considered various classifiers, based on Gaussian Mixture Models, Matched Filters Bank, Low Order Polynomial Fit Coefficients Analysis and Correlation / Spectrum Analysis. We discuss the latter, as it leads to a simple classifier that yields good results.

Viewing the activity $A(t)$ as a random signal, its autocorrelation is $R_A(\tau) = \mathbb{E}(A(t)A(t+\tau))$, where $\mathbb{E}$ denotes expected value. Suppose that $R_A(\tau)$ is estimated based on $A(t)$ data within a finite temporal window $T$, i.e.

$$\hat{R}_A(\tau) \approx \frac{1}{|T|} \int_T A(t)A(t+\tau).$$

Assume that within the temporal window $T$, there is either a gap in traffic, such that $A(t) = n(t)$, where $n(t)$ is a random noise process, or a single vehicle appears at $t = t_o$, so $A(t) = S(t-t_o) + n(t)$, where $S(t)$ is the pulse shape corresponding to the specific vehicle and viewing conditions. It can be seen that the estimated autocorrelation function $\hat{R}_A(\tau)$ is much wider in the latter case, when a vehicle appears, than in traffic gap conditions, where $A(t)$ within the temporal window is due to noise alone. We suggest to classify the current activity sample based on the width of the autocorrelation function estimated within a temporal window ending at the current sample. A double threshold can be used in to obtain an intermediate level, reflecting uncertainty, to be resolved by the subsequent finite state machine. Similar analysis can be conducted in the frequency domain, by thresholding the spectral bandwidth of the activity signal within the temporal window.

4. RESULTS

Fig. 5c shows snapshots from a typical day-time video sequence obtained by the system camera in a single-lane one-way street. Fig. 5a is the Activity signal corresponding to the same experimental session. Fig. 5b zooms on the time interval in which the snap-shots shown in Fig. 5c were taken.

The coloring of the graphs in Figs. 5a and 5b present the indication provided to the pedestrian, red meaning TRAFFIC and blue corresponding to GAP. As can be seen, TRAFFIC is declared 8-10 seconds before the vehicle reaches the camera, i.e., before the peak of the Activity signal.

These results were obtained with a standard USB camera, characterized by a wide field of view. With a narrow field of view lens, the advance warning time is expected to rise.
5. DISCUSSION

We presented a low-cost embedded system to detect and alert pedestrians of vehicles approaching a road-crossing, where traffic control is absent. Our system automatically adapts to the scene, and can be mounted on existing street infrastructure, suggesting fast and easy installation.

We tested our system online and offline using a standard PC at several locations, during both day and night time. We achieve warning-times approaching 10 seconds at typical day conditions with a wide field of view lens, slightly less at night scenarios.

A single instance of the system is capable of detecting traffic approaching from a given direction, ignoring distancing traffic. Full support of two-way traffic requires two instances of the system, mounted in opposite directions, possibly with a unified decision and pedestrian interface module.

Time to Contact (TTC) estimation, based on Horn et al [14], was also tested. As in our approach, it processes entire frames and yields a scalar signal, that could be considered as an alternative to the proposed Activity signal. The results were noisy and unstable compared to the results obtained using the Activity measure.

In addition to the PC implementation, we ported the system to BeagleBoard xM hardware running Linux, and we are also considering an Android version using low-cost smartphone-like hardware.

6. REFERENCES


Gabor Wavelet Transform Based Facial Expression Recognition Using PCA And LBP

Muzammil Abdulrahman, Tajuddeen R. Gwadabe, Fahad J. Abdu, Alaa Eleyan
Electrical & Electronic Engineering Department, Mevlana University, Konya, Turkey
muzammil.abdulrahman.tr@ieee.org, {tgwadabe,fjibrin}@yahoo.com, aeleyan@mevlana.edu.tr

ABSTRACT

This paper proposed a facial expression recognition approach based on Gabor wavelet transform. Gabor wavelet filter is first used as pre-processing stage for extraction of the feature vector representation. Dimensionality of the feature vector is reduced using Principal Component Analysis and Local binary pattern (LBP) Algorithms. Experiments were carried out of The Japanese female facial expression (JAFFE) database. In all experiments conducted on JAFFE database, results obtained reveal that GW+LBP has outperformed other approaches in this paper with Average recognition rate of 90% under the same experimental setting.

Index Terms— Facial Expression Recognition, Gabor Wavelet Transform, Principal Component Analysis, Local Binary Patterns.

1. INTRODUCTION

The increase need in researches in the area of digital processing and pattern recognition give rise to the development of algorithms for different applications such as automated access control, human-computer interaction, and surveillance [1]. For automated access control, the commonly used method is face recognition but the faces might exhibit different Facial Expressions (FE). Face recognition and FE analysis has attracted the interest of researchers in the field of computer vision.

Although human being can recognize different FE exhibited by face very easily [2], it is not so for computers. In a survey on automatic Facial Expression Recognition (FER) [3], Scientists conducted experiments in order to recognize the FE from unoccluded facial images taken under controlled laboratory conditions. But at times, the human subject may be talking, thus changing his facial features or his face may be partially occluded. Rosenblum et al. [4] applied a technique of networks where the complexity of recognition FE was grouped into three layers of decomposition. In [5], FER in the presence of occlusion was investigated. The approach is based on a localized representation of FE features and on the fusion of classifier outputs. Abboud and Davoine [6] suggested a bilinear factorization expression classifier for facial recognition. Ekman and Friesen [7], proposed well known FE representation and coding system, called facial action coding system (FACS). Their goal was to describe all the visually distinct local facial muscle a movement using 46 actions units (AUs), a FE corresponds to the combination of a set of AUs. Lyons et al [8] suggested a technique for classifying facial images automatically based on labelled elastic graph matching and 2D Gabor Wavelet (GW) representation. Researches in [9, 10, 11] demonstrated that GW representation is a good feature extraction technique because of its biological relevance and computational properties. W.K. Kong et al [12] used GW as a feature extraction technique for palm print recognition, A.B. Sharizai et al [13] enhances finger print images by the use of GW transform.

The rest of the paper is organized as follows. Section 2 and 3 briefly give an overview of Gabor wavelet filter and Dimensionality reduction (i.e. LBP, PCA) respectively. Section 4 describes the proposed method and section 5 discusses the simulation results. Finally the paper is concluded in section 6.

2. GABOR WAVELETS

A GW filter is an essential tool used to extract local features both in spatial and frequency domain which can be applied on images to extract features aligned at particular angles (orientations). The GWs filter captures significant visual features such as spatial localization, orientation selectivity, frequency selectivity, and quadrature phase relationship [14]. The GW kernel in spatial domain is a complex exponential modulated by a gaussian function [15].

The GWs kernel can be defined by the following equation:

$$
\psi(x, y, \sigma, \theta) = \frac{1}{2\pi\sigma^2} e^{-\frac{(x^2+y^2)}{2\sigma^2}} e^{i\theta x}
$$

(1)
where \((x,y)\) denote the pixel position in the spatial domain, \(\omega\) is the central frequency of a sinusoidal plane wave, \(\theta\) is the orientation of the Gabor filter and \(\sigma\) is the standard deviation along \(x\) and \(y\) directions. The parameters \(X\) and \(Y\) can be defined by the following equations:

\[
X' = X \cos \theta + Y \sin \theta, \quad Y' = -X \sin \theta + Y \cos \theta
\]  

(2)

Gabor filter bank of various frequencies and orientations has been frequently used to extract features of face image. However in previous work review [14-17] mainly a forty Gabor filter bank at five scales, \(\nu=\{0,1,\ldots,4\}\) and eight orientations, \(\mu=\{0,1,\ldots,7\}\) with \(\sigma=2\pi\), \(\omega=2\sqrt{2}\) were used. Fig. 1 shows the magnitude of the Gabor at four scales and the Real Gabor filter bank with 4 different scales and 6 different orientations.

![Figure 1](image1.png)

Figure 1 (a) Magnitude of the Gabor filter bank at four scales (b) Real part of the Gabor kernels at four scales and six orientations.

The GW feature representation \(\phi_{m,n}(x,y)\) is obtained by convolving the Gabor filter bank \(\psi(x,y,\omega,\theta)\) with input image.

\[
\phi_{m,n}(x,y) = I(x,y) * \psi(x,y,\omega,\theta)
\]  

(3)

In practice usually the magnitude of the convolution output Gabor feature vector representation is used for FER, because they do vary slowly with the displacement while the phases are very sensitive under small displacement [18].

3. **DIMENTIONALITY REDUCTION**

A. **Local Binary Patterns**

LBP operator takes a local neighborhood around each pixel, thresholds the pixels of the neighborhood at the value of the central pixel and uses the resulting binary valued image patch as a local image descriptor. It was originally defined for \(3 \times 3\) neighborhoods, giving 8 bit codes based on the 8 pixels around the central one, the operator labels the pixels of an image by thresholding a \(3 \times 3\) neighborhood of each pixel with the centre value and considering the results as a binary number, and the 256-bin histogram of the LBP labels computed over a region is used as a texture descriptor. Fig. 2 gives an example of the basic LBP operator.

![Figure 2](image2.png)

Figure 2 Example of LBP operators

The limitation of the basic LBP operator is that its small \(3 \times 3\) neighborhood cannot capture the dominant features with large scale structures. As a result, to deal with the texture at different scales, the operator was later extended to use neighborhoods of different sizes [19]. One of such is the uniform patterns: an LBP is ‘uniform’ if it contains at most one 0-1 and one 1-0 transition when viewed as a circular bit string. For instance, 00000000, 001110000 and 11100001 are uniform patterns. It is observed that uniform patterns account for nearly 90% of all patterns in the \((8, 1)\) neighborhood and for about 70% in the \((16, 2)\) neighborhood in texture images, where the notation \((P,R)\) denotes a neighborhood of \(P\) equally spaced sampling points on a circle of radius of \(R\) that form a circularly symmetric neighbor set [4].

B. **Principal Component Analysis**

PCA [20] can be defined as an orthogonal linear transformation that transforms data to a new coordinate system such that the highest variance by any projection of the data comes to sit on the first coordinate (called the first principal component), the second highest variance on the second coordinate, and so on. Thus PCA finds orthogonal basis for data, sorts dimensions in order of importance and discard low significance dimensions. The main purpose of PCA is to reduce the dimensionality of the raw data while retaining as much as possible the variation present in the data set [17]. PCA can also be thought as a way of recognizing patterns in data, and expressing the data in such a way as to emphasize their similarities and differences. Since pattern in data may be difficult to find especially in data of high dimension. PCA is used in many pattern recognition applications such FER [21] and face recognition [22].

4. **PROPOSED METHOD**

In this section we explain the method used in this paper. It can be summarized using the diagram shown in Fig. 3. The FE database used is the Japanese Female Facial Expression (JAFEE) Database, the images where categorized into training and testing. All expressions in the database where first pre-processed using GW Transform (GWT) with 5 scales and 8 orientations and no down sampling were done.
After the pre-processing each image within the database has a corresponding 40 images and each image is 256×256. Due to this large number of images, dimensionality reduction is used. Two methods; PCA and LBP are used here. A uniform LBP is used in which each image is divided into a region of 64×64, a radius of 1 used with 8 samples within each radius. K-nearest neighbor is used for similarity measure and in this case we used Euclidean distance (L₂).

![Figure 3 block diagram of the proposed method](image)

The database contains 213 images of 10 subjects with 7 FEs namely; Neutral, Anger, Disgust, Fear, Happy, Sad and Surprise. Fig. 4 shows example of the images from the database. 137 images containing all the expressions were used for training while 76 images were used for testing.

![Figure 4 Examples of images from the JAFFE database.](image)

### 5. SIMULATION RESULTS

The simulation of the proposed work is implemented in MATLAB and the JAFFE database. The JAFFE data base contains 213 images posed by 10 female with 7 facial expressions. They were taken from Japanese female model each having a resolution of 256×256 pixels. The mages in the database are greyscale images in tiff format. We categorized the images into 137(64%) images containing all the expressions used for training data set while 76(36%) used as a testing data set. In order to demonstrate and compare the performance of the proposed algorithms, we first obtain the PCA, LBP, Gabor, Gabor+PCA, Gabor+LBP features and used the K-nearest neighbour with Euclidean distance (L₂) for similarities measure to classify the face images. The results of the confusion matrix are given in Tables, 1, 2, 3, 4 and 5 for PCA, LBP, Gabor, Gabor+PCA, and Gabor+LBP, respectively. From the result shown below, it reveals that the performance of Gabor+LBP has better performance with averaged recognition rate 90% and promise to be the best result compared to the rest, though all the remaining Algorithms appear to demonstrate a good performance as well when compared to previous work reviewed.

### 6. CONCLUSION

Gabor wavelets were used as a pre-processing stage followed by dimensionality reducing using PCA/LBP for facial expression recognition in this paper. Experimental evaluations the proposed approach were conducted on JAFFE database. The results obtained showed that pre-processing with Gabor wavelets improves the performance of directly applying both PCA and LBP.

<table>
<thead>
<tr>
<th>Table 1 Confusion Matrix for FER using PCA</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCA</td>
</tr>
<tr>
<td>-----</td>
</tr>
<tr>
<td>Neutral</td>
</tr>
<tr>
<td>Anger</td>
</tr>
<tr>
<td>Disgust</td>
</tr>
<tr>
<td>Fear</td>
</tr>
<tr>
<td>Happy</td>
</tr>
<tr>
<td>Sad</td>
</tr>
<tr>
<td>Surprise</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 2 Confusion Matrix for FER using LBP</th>
</tr>
</thead>
<tbody>
<tr>
<td>LBP</td>
</tr>
<tr>
<td>-----</td>
</tr>
<tr>
<td>Neutral</td>
</tr>
<tr>
<td>Anger</td>
</tr>
<tr>
<td>Disgust</td>
</tr>
<tr>
<td>Fear</td>
</tr>
<tr>
<td>Happy</td>
</tr>
<tr>
<td>Sad</td>
</tr>
<tr>
<td>Surprise</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 3 Confusion Matrix for FER using Gabor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gabor</td>
</tr>
<tr>
<td>-----</td>
</tr>
<tr>
<td>Neutral</td>
</tr>
<tr>
<td>Anger</td>
</tr>
<tr>
<td>Disgust</td>
</tr>
<tr>
<td>Fear</td>
</tr>
<tr>
<td>Happy</td>
</tr>
<tr>
<td>Sad</td>
</tr>
<tr>
<td>Surprise</td>
</tr>
</tbody>
</table>
Table 4 Confusion Matrix for FER using Gabor+PCA

<table>
<thead>
<tr>
<th></th>
<th>Neutral</th>
<th>Anger</th>
<th>Disgust</th>
<th>Fear</th>
<th>Happy</th>
<th>Sad</th>
<th>Surprise</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neutral</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Anger</td>
<td>18</td>
<td>82</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Disgust</td>
<td>0</td>
<td>0</td>
<td>67</td>
<td>0</td>
<td>0</td>
<td>25</td>
<td>8</td>
</tr>
<tr>
<td>Fear</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>58</td>
<td>17</td>
<td>25</td>
<td>0</td>
</tr>
<tr>
<td>Happy</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>70</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>Sad</td>
<td>9</td>
<td>18</td>
<td>0</td>
<td>9</td>
<td>63</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Surprise</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>90</td>
</tr>
</tbody>
</table>

Table 5 Confusion Matrix for FER using Gabor+LBP

<table>
<thead>
<tr>
<th></th>
<th>Neutral</th>
<th>Anger</th>
<th>Disgust</th>
<th>Fear</th>
<th>Happy</th>
<th>Sad</th>
<th>Surprise</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neutral</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Anger</td>
<td>0</td>
<td>91</td>
<td>0</td>
<td>0</td>
<td>9</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Disgust</td>
<td>0</td>
<td>0</td>
<td>92</td>
<td>0</td>
<td>8</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Fear</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>75</td>
<td>8</td>
<td>17</td>
<td>0</td>
</tr>
<tr>
<td>Happy</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>100</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Sad</td>
<td>0</td>
<td>0</td>
<td>18</td>
<td>0</td>
<td>82</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Surprise</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>90</td>
</tr>
</tbody>
</table>

7. REFERENCES


A DIRECT APPROACH FOR HUMAN DETECTION WITH CATADIOPTRIC OMNIDIRECTIONAL CAMERAS

Ibrahim Cinaroglu, Yalin Bastanlar

Computer Engineering Department, Izmir Institute of Technology, Izmir, Turkey
{ibrahimcinaroglu,yalinbastianlar}@iyte.edu.tr

ABSTRACT
This paper presents an omnidirectional vision based solution to detect human beings. We first go through the conventional sliding window approaches for human detection. Then, we describe how the feature extraction step of the conventional approaches should be modified for a theoretically correct and effective use in omnidirectional cameras. In this way we perform human detection directly on the omnidirectional images without converting them to panoramic or perspective image. Our experiments, both with synthetic and real images show that the proposed approach produces successful results.

Index Terms— Omnidirectional cameras, object detection, pedestrian detection, human detection

1. INTRODUCTION
Detecting people with cameras is an important task for many research and application areas such as visual surveillance, ambient intelligence and pedestrian safety. Last decade has witnessed significant advances in human detection both in terms of effectiveness and processing time.

Quite a variety of approaches have been proposed for pedestrian detection and in general for human detection. A major group in these studies uses the sliding window approach in which the detection task is performed via a moving and gradually growing search window. A significant performance improvement was obtained with this approach by employing HOG (Histogram of Oriented Gradients) features. Inspired by SIFT (Scale Invariant Feature Transform) [1], Dalal and Triggs [2] proposed to use HOG for the feature extraction step and they used SVM (Support Vector Machines) for the classification step. Later on, this technique was enhanced with part based models. For instance, Felzenswalb et al. [3] proposed a method using parts of the object which are spring-like connected to each other and can move independently. Another notable enhancement was using pyramid HOG features and Intersection Kernel SVM proposed by Maji et al. [4].

Edge based features [5] and ‘shapelets’ [6] are examples of other features which were used for human detection. More recently, it was shown that using combinations of features outperforms the approaches that use a single type of feature [7]. For a detailed summary and comparison of methods we refer readers to [8], where an extensive evaluation of the above mentioned and many other pedestrian detection algorithms exists.

Omnidirectional cameras provide 360° horizontal field of view in a single image (vertical field of view varies). If a convex mirror is placed in front of a conventional camera for this purpose, then the imaging system is called a catadioptric omnidirectional camera (Fig. 1). With its enlarged view advantage, fewer omnidirectional cameras may substitute many perspective cameras. However, so far omnidirectional cameras have not been widely used in object detection research area and also in traffic applications like pedestrian and vehicle detection.

In a study on object recognition with omnidirectional cameras [9], a mobile robot is given the images of several objects in the environment and it is asked to recognize these objects. Actually, the omnidirectional image is warped into a cylindrical panoramic image before matching with the images of the objects. SIFT matching is employed without any modification for omnidirectional cameras. In [10], authors use Haar features to perform face detection with catadioptric omnidirectional cameras. Instead of modifying the feature extraction step, they convert the omnidirectional images into panoramic images and directly use the conventional (perspective) camera technique. In a similar manner, panoramic images are used in [11] for human detection.

A human tracking method for omnidirectional cameras is proposed in [12]. As a part of the proposed algorithm, HOG features are computed. However, a rectangular rotating and sliding window is used with no mathematical modification for the omnidirectional camera.

In this paper, we propose a direct approach to tackle human detection on catadioptric omnidirectional images. That is, we do not convert the omnidirectional images to panoramic or perspective images. To our knowledge, the proposed method is the first one to detect humans directly on omnidirectional images. In Section 2, we explain why our approach is theoretically correct. We adopt HOG+SVM [2] approach for human detection and as explained in Section 3, we modify the HOG feature extraction step for catadioptric omnidirectional cameras. With experiments, given in Section 4, we demonstrate that the adaptation of HOG features improves the performance significantly.

Fig. 1. (a) A mirror apparatus is placed in front of a conventional camera to obtain a catadioptric omnidirectional camera. (b) An example image obtained by such a camera.
2. PROCESSING OF OMNIDIRECTIONAL IMAGES

Due to their nonlinear imaging geometry, working with omnidirectional cameras requires geometric transformations. At first sight, converting an omnidirectional image to a panoramic or several perspective images may seem to be a practical solution. However, it has two major drawbacks: The conversion can be computationally expensive for large frames especially when an omnidirectional image is converted to numerous perspective images to properly fit sliding windows. More importantly, the interpolation required by the image warping introduces artifacts that affect the detection performance.

Among a small number of omnidirectional object detection studies (cf. Section 1), none of them developed a method peculiar to omnidirectional cameras. On the other hand, last decade studies (cf. Section 1), none of them developed a method peculiar that affect the detection performance.

We describe these approaches and summarize their properties.

Step of SIFT and avoid warping omnidirectional images. Below, we witnessed some effort on computing SIFT features in to omnidirectional cameras. On the other hand, last decade studies (cf. Section 1), none of them developed a method peculiar that affect the detection performance.

3. THE PROPOSED HOG COMPUTATION

To detect the standing people in omnidirectional images, we rotate the rectangular sliding window around the image center. In addition, to achieve a mathematically correct detection method, we modify the image gradients. The operations that we perform can be divided into two steps:

1. Modification of gradient magnitudes using Reimannian metric
2. Conversion of gradient orientations to form an omnidirectional (non-rectangular) sliding window.
where $g^{ij}$ is the inverse of $g_{ij}$.

A similar reasoning is used in [16] and [19] to obtain the Laplace-Beltrami operator, which is the second order differential operator defined on $\mathcal{M}$ and used for scale space representation for SIFT. In this paper, we are working on the first derivatives. Let us briefly go over the para-catadioptric case.

Consider the unitary sphere $S^2$ with radius = 1 (Fig. 3a). A point on $S^2$ is represented in Cartesian and polar coordinates as

$$ (X, Y, Z) = (\sin \theta \sin \varphi, \sin \theta \cos \varphi, \cos \theta) \quad (2) $$

The Euclidean line element in Cartesian coordinates, $dl$, can be expressed in polar coordinates as

$$ dl^2 = dX^2 + dY^2 + dZ^2 = d\theta^2 + \sin^2 \theta \, d\varphi^2 \quad (3) $$

The stereographic projection of the sphere model sends a point on the sphere $(\theta, \varphi)$ to a point in polar coordinates $(R, \varphi')$ in the image plane (plane $\mathbb{R}^2$), for which $\varphi'$ remains the same and $\theta = 2 \tan^{-1}(R/2)$ in a para-catadioptric system (Fig. 3b).

Using the identities, $R^2 = x^2 + y^2$, $\varphi = \tan^{-1}(y/x)$ the line element reads

$$ dl^2 = \frac{16}{(4x^2y^2)} (dx^2 + dy^2) \quad (4) $$

giving the Riemannian inverse metric

$$ g^{ij} = \frac{(4x^2+y^2)^2}{16} \quad (5) $$

We refer the reader to [16] and [18] for a detailed derivation of catadioptric Reimannian metric. With this metric, we can compute the differential operators on the sphere using the pixels in the omnidirectional images. In particular, norm of the gradient reads

$$ |\nabla_x I|^2 = \frac{(4x^2+y^2)^2}{16} |\nabla_\varphi I|^2 \quad (6) $$

We see that the para-catadioptric gradients are just the scaled versions of the gradients in Euclidean domain. Therefore, we simply multiply our gradients with metric $g^{ij}$.

At the center of the omnidirectional image, $(x, y) = (0, 0)$, Reimannian and Euclidean gradients are the same. At an image location when $x^2 + y^2 = 2$, which corresponds to a 3D point at the same horizontal level with the sphere center (mirror focal point), the Reimannian metric is equal to 4. Therefore the gradients are magnified as we move from the center to the periphery of the omnidirectional image. This metric is extended to all central catadioptric systems by Puig et al. [19].

3.2. Conversion of gradients for omnidirectional sliding window

After the image gradients are obtained with Reimannian metric, we convert the gradient orientations to form an omnidirectional (non-rectangular) sliding window. A rectangular object in a perspective image is warped in the omnidirectional image, therefore the gradients in the sliding window should be computed as if a perspective camera is looking from the same viewpoint.

The reader should note that we train our model for human detection using INRIA perspective image dataset as described in [2], i.e. we do not train an omnidirectional HOG model. Since the shape of the non-rectangular sliding window varies according to the location in the omnidirectional image, it is not plausible to train many omnidirectional HOG models. The modifications we made for HOG computation in omnidirectional sliding window enables us to compare it with the perspective camera HOG model. Fig. 4a shows a half of a synthetic para-catadioptric omnidirectional image (400x400 pixels) where the walls of a room are covered with rectangular black and white tiles. Conventional HOG result of the marked region (128x196 pixels) in this image is given in Fig. 4b where gradient orientations are in accordance with the image. However, since these are vertical and horizontal edges in real world, we need to obtain vertical and horizontal gradients. Fig. 4d shows converted gradients for the region marked in Fig. 4c, which is an example of the proposed HOG computation.

Since the Cartesian coordinates in the detection window (Fig. 4d) corresponds to a nonlinear distribution of pixels in the image (Fig. 4c), we employ bilinear interpolation with backward mapping both for gradient orientations and gradient magnitudes.

![Fig. 4. Description of how the gradients are modified for an omnidirectional sliding window. Result in (b) is the regular HOG computed for the region marked with dashed lines in (a). Modified HOG computation gives the result in (d) for the region marked in (c). Vertical and horizontal edges in real world produce vertical and horizontal gradients in the modified version.](image313x435_to_532x591)

4. EXPERIMENTS

4.1. Evaluation of the modified HOG using SVM scores

Let us first compare the results of the proposed HOG computation and the regular HOG computation on the omnidirectional images. Since the computed HOG features are given to an SVM trained with person image dataset, we aim to obtain higher SVM scores with the proposed omnidirectional HOG computation.

We artificially created 210 omnidirectional images containing humans. While creating this set, we followed an approach similar to [15], where images in INRIA person dataset are projected to omnidirectional images using certain projection angle and distance parameters. Fig. 5 shows an example omnidirectional image, where the regular HOG window (rectangular, 128x64 pixels) and the proposed omnidirectional HOG window (non-rectangular) are shown. The HOG features computed with the two window types are compared with their resultant SVM scores. Since the locations of projections in these images are known, no search is needed for this experiment. However, vertical position of the window affects the result. For both approaches, we chose the position that gives the highest mean SVM score. Table 1 summarizes the result of the comparison, where we see that the mean score for the proposed approach is higher than that of regular HOG window.
Fig. 5. Depiction of the regular HOG window (green rectangle) and the proposed window (red doughnut slice) on an omnidirectional image artificially created by projecting a perspective image from INRIA person dataset.

Table 1. Comparison of the regular and proposed HOG window by their SVM scores

<table>
<thead>
<tr>
<th></th>
<th>Mean SVM score</th>
<th>Minimum SVM score</th>
<th>Maximum SVM score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regular HOG window</td>
<td>1.69</td>
<td>-1.01</td>
<td>3.21</td>
</tr>
<tr>
<td>Proposed HOG window</td>
<td>1.93</td>
<td>-0.42</td>
<td>3.64</td>
</tr>
</tbody>
</table>

Fig. 6. Human detection results on an omnidirectional image with SVM scores (upper left corners) greater than 1. (a) Proposed sliding windows. (b) Regular (rectangular) sliding and rotating windows. (c) Regular sliding windows on panoramic image.

4.2. Experiments with real images

In this subsection, we present the results for a set of images taken with our catadioptric omnidirectional camera. We compared the proposed HOG computation not only with the regular HOG window, but also with the approach that first converts the omnidirectional image to a panoramic image and then performs HOG computation. Although it was explained in Section 2 that working on panoramic images is not a theoretically correct approach, we wanted to test its performance. Fig. 6 shows the results for one of the images in the set. SVM scores greater than 1, after non-maximum suppression, superimposed on the images with the proposed HOG window, the regular HOG window on omnidirectional image and HOG after panoramic conversion. For the humans in the scene, the average SVM scores for the proposed HOG, the regular HOG and HOG on panoramic image approaches are 2.94, 2.11 and 2.41 respectively.

To evaluate the overall performance of these three approaches, we plot precision-recall curves for the whole dataset which consists of 20 real images taken in different scenes including indoor and outdoor environments (Fig. 7). The aim of these curves is to show the two performance metrics together: Precision (#True positives / #Predicted positives) and Recall (#True positives / #Actual positives). The larger the area under the curve, the better the performance of the algorithm. As the threshold increases, all approaches reach Precision=1. One can observe that the performance of the proposed HOG computation is better than the others up to a threshold value of 4.0. At higher threshold values our method loses its advantage. However, since the recall comes below 0.5 for those values, it is not plausible to use them in practical systems.

A detection window is considered to be a True-positive if it overlaps an annotation by 50%, where the overlap is computed as

$$O = \frac{\text{area}(\text{detection window } \cap \text{ annotation})}{\text{area}(\text{detection window } \cup \text{ annotation})}$$

For a fair comparison, the annotations are separately prepared for the mentioned three methods. Annotations of the proposed HOG approach (e.g. Fig. 6a) are doughnut slices, annotations of the regular HOG approach are rectangles rotating around the omnidirectional image center. Finally, annotations of the HOG on panoramic image approach are upright rectangles.

Fig. 7. Precision-Recall curves to compare the proposed HOG computation, the regular HOG and HOG after panoramic conversion. The data points in this curve correspond to the varying threshold values for the SVM score, which change from 0 to 5.

5. CONCLUSION

We aimed to perform human detection directly on the omnidirectional images. As a base, we took the HOG+SVM approach which is one of the popular human detection methods. After describing how the feature extraction step of the conventional method should be modified, we performed experiments to compare the proposed method with the regular HOG computation in omnidirectional and in panoramic images. Results of the experiments indicate a performance increase for the proposed approach.

In the near future, we are planning to prepare a larger set of real omnidirectional images, and perform tests using that set.
6. REFERENCES


Regular Session 3
Abstract—Our goal is to explore the techniques of 2D/3D facial recognition in the presence of expression and also the optimization of SRV3D by selection of the best features for optimal descriptor. The step of extracting relevant information from the face is carried out by the method of linear discriminant analysis improved Fisher (EFM). These characteristic parameters are then submitted to the classifier known for its effectiveness on SVM (Support Vector Machine). In this work we use a system optimization 3D face recognition based on PSO "Particle Swarm Optimization" or "OEP". This algorithm allows us to achieve the best candidates to improve the success rate and the system performance. The PSO is a computational model based on the idea of cooperative behavior is inspired by the social behavior of birds, fish, bees or ant colonies. This is a very simple method based on the calculation of mathematical equations known as speed, position, inertia, distance, error measurement and the fitness function (equal error rate TEE). Computing the velocity and position calculations are fundamental in this method. All our work is done in 3D and on the BDD Casia3D V4.

Keywords— Optimization; PSO (Particle Swarm Organisation); EFM (Enhanced Fisher Model); SVM (Support Vector Machine); FRS3D (Face Recognition System 3D)

I. INTRODUCTION

Facial recognition is a biometric technique widely used in various areas such as security and access control, the forensic medicine and the police control. It is to determine if the image of the face of a given person corresponds to one of the images of faces stored in a database.

Facial recognition is a kind of identification and authentication, which primarily uses the function global-face. However, the rate of accuracy of recognition is not yet high enough. This research aims to develop a method to increase the efficiency of the recognition by using the function of optimization based on the PSO.

In the last decade, there are a variety of effective methods in the facial recognition as the analysis of sub-space, matching elastic graphs, analysis local characteristic, neural networks, forming brightness plan of the curve and so on. However, the face recognition is rather complex and difficult to describe because the influence of factors: the angle, the lighting, the expression, scale, rotation, mantle, hairdressing. At the present time, the majority of research work is always focused on the conditional recognition limited.

In previous research, PCA is often adopted. But the algorithm for matching characteristic based on the gray value is able to be less resistant to disturbances, so that the same image to an individual under different angles is often regarded as another. Our approach is a combination of features and especially the optimization of the SRV3D to ensure the recognition and authentication of the face 3D. We are focusing on the search for robust functionality and precise, without any intervention of the user. The system is generally divided into three parts: the data acquisition, feature extraction and authentication or recognition. The time of operation of the camera is a fraction of a second, while the rest of the time is devoted to the software to generate a triangle mesh of the face, which takes a few seconds. The triangle mesh is the means of representation the more common to represent the 3D data from systems of scan. The phase of extraction of characteristics is based on the classification by surface of forms using the bends, find the symmetries, the critical points and the profile curves, the nose and other biometric parties concerned of the face. Finally, the authentication module / recognition made comparisons of the side of entry by report to the database.

Our method can handle different poses of faces with different facial expressions which may include hair on the face. The entry can contain in more of the sweep face the area of the upper part of the body with clothes. We are increasing the discrimination through a system of extraction of characteristics.

II. RELATED WORK

Yongzhong Lu [1] proposed in 2007 an effective hybrid ADP-PSO strategy for optimization and its application to Face Recognition. In order to distinguish faces of various angles during recognition, an algorithm of the combination of approximate dynamic programming (ADP) which is called action dependent heuristic dynamic programming (ADHD) and particle swarm optimization (PSO) is presented and used, that is to say, ADP is applied for dynamically changing the values of the PSO parameters. During the process of face recognition, the discrete cosine transformation (DCT) is first introduced to reduce negative effects. Then K-L transformation can be used to compress images and decrease data dimensions. According to principal component analysis (PCA), main parts of vectors are extracted for data representation. Finally, radial basis function (RBF) neural network is enrolled to recognize various faces. The training of RBF neural network is exploited by ADP-PSO. In terms of ORL Face Database, the experimental result gives a clear view of its highly accurate efficiency. Lanzarini Laura and al [2] interested in 2010 to Face Recognition using SIFT and Binary PSO descriptors. In this work, a strategy for face recognition based on SIFT descriptors of the images involved is presented. In order to reduce the number of false positives and computation time, a selection of the most representative feature descriptors is carried out by applying a variation of the binary PSO method. The results obtained show that the strategy proposed is better than the direct application of SIFT descriptors.
A new method for face recognition based on pca optimize strategy [3] was study by Jian Zhang and al in 2010. They focused on the problem of selecting optimum discrimination eigenvectors of PCA and improving the recognition accuracy. A new method for face recognition based on PCA optimize strategy is presented, in which the PSO algorithm is embedded, which select the recognition accuracy as the fitness value of particle swarm, to find out the optimum discrimination eigenvectors of PCA and obtain the optimal recognition accuracy simultaneously. We validate the effectiveness of this method with the ORL database and the Yale database. The experimental results indicate that the method can obtain the optimum discrimination eigenvector of PCA and a major improvement on recognition accuracy compared with the eigenvector selection approach based on the energy accumulative contribution rate.

Adil Abdulwahhab Ghidan and al [4] present a novel feature selection algorithm based on particle swarm optimization (PSO). The algorithm is applied to coefficients extracted by two feature extraction techniques: the discrete wavelet transform (DMWT). The proposed PSO-based feature selection algorithm is utilized to search the feature space for the optimal feature subset where features are carefully selected according to a well defined discrimination criterion. Evolution is driven by a fitness function defined in terms of maximizing the class separation (scatter index). The classifier performance and the length of selected feature vector are considered for performance evaluation using the ORL face database. Experimental results show that the PSO-based feature selection algorithm was found to generate excellent recognition results with the minimal set of selected features.

The authors R. Raghavendra and al [5] in 2011 worked on particle swarm optimization based fusion of near infrared and visible images for improved face verification. They present two novel image fusion schemes for combining visible and near infrared face images (NIR), aiming at improving the verification performance. Sub-band decomposition is first performed on the visible and NIR images separately. In both cases, they employ particle swarm optimization (PSO) to find an optimal strategy for performing fusion of the visible and NIR sub-band coefficients. In the first scheme, PSO is used to calculate the optimum weights of a weighted linear combination of the coefficients. In the second scheme, PSO is used to select an optimal subset of features from visible and near infrared face images. To evaluate and compare the efficacy of the proposed schemes, they have performed extensive verification experiments on the IRVI database. This database was acquired in their laboratory using a new sensor that is capable of acquiring visible and near infrared face images simultaneously thereby avoiding the need for image calibration. The experiments show the strong superiority of the first scheme compared to NIR and score fusion performance, which already showed a good stability to illumination variations. Aneesh M U and al [6] have worked in 2012 on the optimal feature selection based on image pre-processing using Accelerated Binary Particle Swarm Optimization for enhanced Face Recognition. Feature Selection is an optimization problem in any Face Recognition technology. Authors propose a novel method of Binary Particle Swarm Optimization called Accelerated Binary Particle Swarm Optimization (ABPSO) by intelligent acceleration of particles. Together with Image Pre-processing techniques such as Resolution Conversion, Histogram Equalization and Edge Detection, ABPSO is used for feature selection to obtain significantly reduced feature subset and improved recognition rate. The performance of ABPSO is established by computing the recognition rate and the number of selected features on ORL database and Cropped Yale B database.

III. PROPOSED FR BASED ON PSO

The block diagram of the proposed FR system is shown in Fig. 1. The two stages involved are: Training stage and Recognition stage.

![Fig. 1. Block Diagram of the proposed Face Recognition system](image)

IV. IMAGE PRE-PROCESSING, FEATURES EXTRACTION AND BPSO

A. Image pre-processing

Among the various pre-processing techniques, the three techniques of relevance are Bi-Cubic interpolation, Histogram equalization and Edge detection using Laplacian of Gaussian (LoG). Image interpolation provides a technique of producing high-resolution image from its low-resolution counterpart [7]. Interpolation basically, is the process of estimating intermediate values of a continuous event from discrete samples [8]. It is a type of approximating function whose value must coincide with the sample data at the interpolation nodes or sample points. Bi-Cubic interpolation is a resolution conversion method preserving finer details of images with increased sharpness, better than bilinear algorithm. Whenever an image is resampled, there will be a loss of focus within the image, but bi-cubic interpolation, among various methods, provides maximum sharpness. Histogram equalization is a nonlinear process aimed to highlight brightness in a way particularly suited to human visual analysis [9]. This tries to transform the distribution of pixel intensity values in the image into a uniform distribution and subsequently improves the image’s global contrast. In order to preserve local features, despite the influence of lighting, useful for recognition, LoG [10] is applied after histogram equalization. LoG is applied for
Cropped Yale B database to obtain efficient edge extraction and detection of different resolution edges.

B. Feature extraction

Feature extraction is an essential section in pattern recognition. It’s a key for improving the recognition accuracy to get the effective discrimination features. The PCA (Principal Component Analysis) and the FLDA (Fisher Linear Discrimination Analysis) are both the widely-used features extraction methods in the field of face recognition. The FLDA extracts the optimal discrimination features by maximizing the between-class scatter matrix and minimizing the within-class scatter matrix simultaneously. However, the FLDA is confronted with the problems of small samples size (SSS) and the singular of within-class scatter matrix in face recognition. Kirby et al. first used the Karhunen-Loeve Transform to represent face images. Turk et al. proposed the Eigenface method applied Karhunen-Loeve transform to face recognition, which is also called PCA. Nevertheless, the PCA only extracts the optimal representation features with a view to reconstruct face images rather than the optimal discrimination features for face recognition, so the direct PCA method could not get the best recognition accuracy. Belhumeur et al. proposed the Fisherfaces method [11], which first used the PCA to reduce the dimension of face images and then used the FLDA to recognize faces. Fisherfaces improves the recognition accuracy and avoids the singular of the within-class scatter matrix of the FLDA, thus it is widely utilized for the face recognition. There are still some linear dependence feature vectors after dimensionality reduction with PCA. The experiment showed that the recognition accuracy could not be improved by using all eigenvectors of PCA. It’s a hard issue how to select the optimal discrimination eigenvectors for face recognition from PCA. C.J.Liu et al. put forward a eigenvector selection method which is called cumulative contribution rate of energy (CCRE) and it gets a better recognition accuracy. They think that there is more interference information in the eigenvectors corresponded to the small eigenvalues, so it is a natural selection to keep the eigenvectors corresponded to the large eigenvalues. In fact, the eigenvectors of large eigenvalues reflect the profile feature and the eigenvectors of small eigenvalues depict the minutiae of the face images, thus the CCRE tends to leave out a part of minutiae discrimination information in theory.

Inspired by the PSO (Particle Swarm Optimization), in this research, a new face recognition method based on PCA+EFM optimize strategy (Optimum-PCA+EFM) is presented. Using the PSO and accuracy of regression ideology to pick up the optimal discrimination eigenvectors of EFM and get the optimal recognition accuracy simultaneously.

C. Particle Swarm Optimization

PSO is a stochastic, population based optimization technique aiming at finding a solution to an optimization problem in a search space. The PSO algorithm was first described by Kennedy and Eberhart in 1995 [12]. PSO is an intelligent and efficient optimization algorithm, which simulates the bird predator behavior and gets the optimal solution. PSO initializes some particles in the solution space firstly. Each particle denotes a potential optimal solution of the global optimization problem. The feature of particle is represented by position, velocity and fitness value. The fitness value is calculated by the fitness function, and the value denotes the merits of the particle. The particles move in the solution space, replace their position and velocity by tracking the individual extreme value and the global extreme value, then calculate the fitness value and replace the position of the individual extreme value and the global extreme value. Let RF denotes the solution space, PSO is an intelligent optimization algorithm inspired by bird flocking and fish-schooling.

The swarm of particles corresponds to a population of simple agents, called particles. Each particle is considered as a solution of the problem, where it has a position (the solution vector) and a speed. In addition, each particle has a memory allowing him to remember his best performance (in position and value) and of the best performance achieved by the particles “nearby” (informants): each particle has in effect of a group of informants, historically called his neighborhood.

A swarm of particles, which are of potential solutions to the problem of optimization, "overflew" the search space, the search for the global optimum. The movement of a particle is influenced by the following three components:

1. A component of inertia: the particle tends to follow its current direction of travel;
2. A cognitive component: the particle tends to move toward the best site by which it is already past;
3. A social component: the particle tends to rely on the experience of its congeners and, thus, to move toward the best site already achieved by its neighbors.

The strategy of displacement of a particle is shown in (Fig 2)

In the basic PSO, potential solutions called particles fly through the problem space by following the current optimum solutions. Binary PSO was formulated to have a discrete version of PSO and the algorithm has been developed. The only difference compared to the continuous PSO is that each new positional value is set to 0 or 1 by applying a sigmoid transformation and a probabilistic rule. The Binary PSO requires the position to be coded as a string of 1’s and 0’s. The particle velocity function is used as the probability function for the position update. A potential solution is represented as a particle having positional coordinates $x_i^t = [x_{i1}^t, x_{i2}^t, ..., x_{id}^t]$ in a D-dimensional space where the subscript $i$ denotes the particle number and superscript $t$ represents the iteration number. Each $i$th particle maintains a record of the position of
In the binary support vector machine (BSVM), an SVM performs classification tasks by constructing optimal separate hyperplanes (OSHs). An OSH maximizes the margin between the two nearest data point belonging to two separate classes. Suppose that we have $D$ is a dataset, which given $n$ as the amount of the labeled training samples, $x_i$ are the training samples while $y_i$ are the labels or targets in $p$-dimensional real vector as:

$$D = \{(x_i, y_i) | x_i \in \mathbb{R}^p, y_i \in \{-1, +1\} \}$$

Subject to $w x_i + b = 0$ (5). The outcome in a linearly separable problem corresponding to determination function can be expressed as $f(x) = (w \cdot x) + b$. The distance from the hyperplane to a point $x$ is given by function as:

$$\text{margin} = \frac{f(x)}{||w||} = \frac{2}{||w||}$$

Therefore, the hyperplane which optimally separates the data is the one that minimizes the following expression as:

$$\min_{||w||^2} \sum_{i=1}^{n} a_i (y_i(w \cdot x_i) + b - 1)$$

The Lagrange multiplier is used to solve the optimization problem can be expression as:

$$L(w, b, a) = \frac{1}{2} ||w||^2 - \sum_{i=1}^{n} a_i (y_i(w \cdot x_i) + b - 1)$$

The Lagrange function has a minimize value for $w$ and $b$. (9). The Lagrange function has a minimize value for $w$ and $b$. Classical Lagrange duality enables the primal problem (9) to be transformed into its dual problem, which is easier to solve. The dual problem is defined as follows:

2) Classification by similarity measure

When one wants to compare two vectors of characteristics reduced from the module data reduction of a biometric system, you can either perform a similarity measure (likeness), either a measurement of distance (divergence). Although the Euclidean distance is optimal in the theory, the various experiences have found that the Euclidean distance is surpassed by other distances. One of them is the normalized correlation which is defined by:

$$S(A, B) = \frac{A^T B}{||A|| ||B||}$$

This function simply calculates the cosine of the angle between the two vectors characteristics $A$ and $B$. A high value of normalized correlation corresponds to a good similarity between the two vectors.

V. EXPERIMENTATION AND RESULTS

To validate our work we have tested our approach on the frontal images from our database CASIA 3D V4. The database is built by a digitizer 3D without contact Minolta VIVID 910, working on the quick mode. This database...
contains 123 subjects, each subject having 37 or 38 images with the individual variations of poses, expressions, and illuminations, the combined variations of expressions under illumination and install under expressions. It contains complex variations which are difficult to any algorithm. In our work we are studying the variations of illuminations (Fig 4), expressions (Fig5) and the combined variations of expressions under illumination, therefore we used 15 images for each subject. The database of 1845 images is divided into two subsets, the set of gallery and the test set. The whole of gallery contains an image for each subject (under the condition of front view, office lighting, and the neutral term).

![Fig. 4. Illumination variations of the database 3D CASIA [CASIA]](image)

![Fig. 5. Expression variations of the database 3D CASIA [CASIA]](image)

### TABLE I. OUR PROTOCOL

<table>
<thead>
<tr>
<th>Together</th>
<th>Customer</th>
<th>Impostor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Learning</td>
<td>500 Images (1, 4, 8, 9, 10)</td>
<td>0 Images</td>
</tr>
<tr>
<td>Evaluation</td>
<td>500 Images (2, 6, 7, 14, 15)</td>
<td>195 Images (1:15)</td>
</tr>
<tr>
<td>Test</td>
<td>400 Images (3, 5, 11, 12, 13)</td>
<td>150 Images (1:15)</td>
</tr>
</tbody>
</table>

### A. System for recognition of 3D faces

Our job is to make a system optimization SRV 3D by using the maps of depth which is very beneficial and helped to improve the performance of the SRV 3D. What is the goal of all work on the identification of persons in biometrics.

### B. Implementation of the approach

This program illustrate by (Fig6) is presented in 3 main modules:

1) **Module 1: Detection and preprocessing:** It is a fundamental phase of a system of recognition of face. It is composed of two basic blocks:
   - The cutting operation;
   - The preprocessing based on the generation of maps of depth 2.5D on the other hand.

A face is described by a cloud of 3D points captured by a laser scanner 3D. Each point cloud consisting of thousands of points in 3D space these discrete points describe the surface of the face. In our database of face 3D CASIA each point is described with spatial coordinates 3D and the coordinates RGB color correspondents. In this section, we describe how the original data in 3D are preprocessed.

The 3D data are converted into an image of depth (see Fig 7 (a)). In most images, the nose is the nearest part of the face to the scanner 3D, that is to say, it has the highest value in depth between all points of the face. By using a window size of 3x3 which calculates the sum of the values of depth in its corresponding pixels, the nose is detected as the coordinates of the center pixel of the window which returns the maximum value (see Fig 7 (b)). After having detected the nose, all the images in the database are cut by a rectangular window of size fixed centered around the center of nose (see Fig 7 (c)).

![Fig. 7. Preprocessing of image depth: (a) the image of depth; (b) detection of end of nose; (c) sliced Image; (d) after removing the noise and the filling of holes.](image)
2) Module 2: Features Extraction: who performs the extraction of features obtained by calculating the ACP of the covariance matrix of the image. The aim is to extract the characteristics of the face that can make both different from those of other people and robust to variations of the person itself. This is the information necessary for the face of a person does not look like that of another person and at the same time that it resembles itself in other conditions of acquisition. At the beginning of work on the recognition of face, it has been estimated that a representation of the face should go by the use of the mouth, the eyes, nose, their relative positions and their geometry. But this procedure has shown its limits. It must be then a more thorough face analysis to find other characteristics. In some methods, we do not used elsewhere that the eyes detection to normalize the face and then made a comprehensive study of the face (type algorithm ACP, LDA, EFM etc.) are calculated and a characteristics facial vector transformed by the ACP and EFM component the vector of each image of face. And then we select the best characteristics parameters by the PSO algorithm.

3) Module 3: classification based SVM or similarity measure.

4) Module 4: optimization by swarm of particles: The learning algorithm and optimization of the EER is shown on the Fig 9.

Fig. 2. Flowchart of PSO learning.

C. Results of our approach of optimization

In the table II we present the results obtained by varying the number of iterations Ni by setting weight w = 0.95; max speed = 6; number of particles Np = 20; constants acceleration c1 = 0.7 and c2 = 1.2. We vary the number of iterations until the minimum TEE.

<table>
<thead>
<tr>
<th>Ni</th>
<th>Evaluation</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>TEE</td>
<td>FAR</td>
</tr>
<tr>
<td>20</td>
<td>8.5</td>
<td>6.49</td>
</tr>
<tr>
<td>40</td>
<td>8.51</td>
<td>6.62</td>
</tr>
<tr>
<td>70</td>
<td>7.71</td>
<td>6.33</td>
</tr>
<tr>
<td>100</td>
<td>8.13</td>
<td>6.55</td>
</tr>
</tbody>
</table>

For the number of parameters selected characteristics Ncs/100 equal to 51, we get a SR = 86.27% a EER = 7.71% for a number of iteration or equal to 70. For this, we are trying to secure the nor to 70 by varying the number of particles Np. The resulting from these experiences is represented in the table III. In the table III we present the results obtained by varying the number of particles in securing weight w = 0.95; number of particles Np = 30; number of iteration max Ni = 70; constants acceleration c1 = 0.7 and C2 = 1. We vary the number of iterations until the minimum EER.

- Success Rate SR = (100- (FAR+FRR))
- Equal rates of error EER on FAR=FRR

<table>
<thead>
<tr>
<th>Np</th>
<th>Evaluation</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>EER</td>
<td>FAR</td>
</tr>
<tr>
<td>20</td>
<td>7.71</td>
<td>6.33</td>
</tr>
<tr>
<td>10</td>
<td>7.72</td>
<td>6.54</td>
</tr>
<tr>
<td>15</td>
<td>7.34</td>
<td>6.26</td>
</tr>
<tr>
<td>25</td>
<td>7.66</td>
<td>6.59</td>
</tr>
<tr>
<td>30</td>
<td>7.28</td>
<td>5.97</td>
</tr>
<tr>
<td>40</td>
<td>7.6</td>
<td>6</td>
</tr>
</tbody>
</table>

We are seeing a slight improvement of performance SRV3D with a EER = 7.28% and a SR = 86.42%. This motivates us to extend our investigations by varying the speed of particles to study its effect on the optimization of our SRV3D see table IV. In the table III we present the results obtained by varying the number of particles in securing weight w = 0.95; number of particles Np = 30; number of iteration max Ni = 70; constants acceleration c1 = 0.7 and C2 = 1.2. We do vary the max speed up to obtaining the minimum EER.

<table>
<thead>
<tr>
<th>Vmax</th>
<th>Evaluation</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>EER</td>
<td>FAR</td>
</tr>
<tr>
<td>6</td>
<td>7.28</td>
<td>5.97</td>
</tr>
<tr>
<td>4</td>
<td>8.12</td>
<td>6.65</td>
</tr>
<tr>
<td>8</td>
<td>8.23</td>
<td>7.18</td>
</tr>
</tbody>
</table>

According to the resulting obtained we can that there is stability in the values of the rates obtained from or stability of the SRV3D. Finally in the last experiment we have used a classification by SVM classifier this efficient and rapid in the table 5 we present the bulk of resulting obtained in this case as well as a comparative study of different SRV3D: PCA+EFM+Distance without PSO ; PCA+EFM+Distance with PSO ; PCA+EFM+SVM with PSO.

<table>
<thead>
<tr>
<th>Np</th>
<th>Evaluation</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>EER</td>
<td>FAR</td>
</tr>
<tr>
<td>20</td>
<td>85.30</td>
<td>8.2</td>
</tr>
<tr>
<td>40</td>
<td>85.17</td>
<td>8.2</td>
</tr>
<tr>
<td>70</td>
<td>86.22</td>
<td>7.4</td>
</tr>
<tr>
<td>100</td>
<td>85.44</td>
<td>8</td>
</tr>
</tbody>
</table>

TABLE V. COMPARATIVE STUDY OF DIFFERENT SRV3D
Finally we can confirm that the PSO algorithm has played its role of optimization especially by associating it with the classifier SVM-capable. This result is encouraging and satisfactory in the area of face recognition 3D especially on a basis if complex to identify such the BDD CASIA 3D or faces are Asians under various variants.

**VI. CONCLUSION**

In this work, we have defined the recognition of faces 3D, and then we have presented a state of the art systems for recognition of faces in which we cited the methods used the most and the most promising to assess this type of system. We can say that the PSO optimisation method might be a good candidate. We have presented the various stages of the design of our approach system optimization of face recognition 3D. We have tried to examine all the variants associated with our algorithms in order to optimize the maximum our recognition system. The results are encouraging and we pushed to retain the PSO as a tool for optimization in biometric systems. We can say that we have been able to optimize the SRV3D by the EFM-PSO-SVM which is proved by the table V. We find a clear improvement of the performance of the SRV3D in the case of the Association of the SVM to the PSO with the passage of the rate EER = 13.03% and a SR = 81.95% (classification by distance metric) to EER = 1.9% and TR = 90.40% (classification SVM). This result is encouraging especially on a BDD complex and difficult to identify and also in the case of SRV3D or the rate of recognition are yet to improve. Several tests have been carried out on this basis to study the parameters of the relevant system and highlight the optimal values. These tests have shown the effectiveness of the use of the filtering method compared to the direct method. This is explained by the fact that the method of filtering uses only the useful information (discriminant information). The results of these tests have shown the contribution of PSO to the performance, however, it was found that the superiority of such a technique on another strongly depends on characteristics of the database selected. We confirm that this method remains a good candidate for addressing the faces to 3D improvements remains to carried out mainly at the level of the preprocessing phase and also and especially the phase of classification. Because our use to the metric measurement remain a choice guided by its simplicity. That said other standards, such as Manhattan or geodesic measurements remain possible, and always in the objective of enhancing those performance parameters of the biometric system. Overall, the results are satisfactory, but the rate of false rejection remains average. This is surely due to the preprocessing step that remains to improve as well as the choice of a best classifier.

<table>
<thead>
<tr>
<th>Method</th>
<th>EER</th>
<th>SR</th>
<th>TR</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCA+EFM+Dis without PSO</td>
<td>13.03</td>
<td>9.04</td>
<td>9</td>
</tr>
<tr>
<td>PCA+EFM+Dis with PSO</td>
<td>7.28</td>
<td>5.97</td>
<td>7.6</td>
</tr>
<tr>
<td>PCA+EFM+SVM with PSO</td>
<td>1.9</td>
<td>6.39</td>
<td>3.2</td>
</tr>
</tbody>
</table>

**VII. REFERENCES**


EXTENSION OF GBVS TO 3D MEDIA

Zbigniew Zdziarski  Rozenn Dahyot
zdziarze@tcd.ie  Rozenn.Dahyot@tcd.ie
School of Computer Science and Statistics
Trinity College Dublin, Ireland

ABSTRACT
Visual saliency has been studied extensively in the past decades through perceptual studies using eye tracking technologies and 2D displays. Visual saliency algorithms have been successfully developed to mimick the human ability to quickly spot informative local areas in images. This paper proposes to investigate the extension of visual saliency algorithms to media displayed in 3D. We show first that the Graph-Based Visual Saliency (GBVS) algorithm outperforms all the other common 2D algorithms as well as their 3D extensions. This paper then extends GBVS to 3D and shows that these new 3D GBVS based algorithms outperform other past algorithms.

Index Terms—Visual saliency, 3D media

1. INTRODUCTION
Humans optimise the analysis of a visual scene by quickly sifting through irrelevant information. People’s attention shifts from one important region to another and it is during this time that the brain builds its representation of the surrounding world [1]. This optimising behavioural action is the subject of research in visual saliency (VS) and it is studied extensively with the use of eye trackers monitoring people’s gaze as they look at specific images and videos. In the past three decades, several visual saliency algorithms (VSAs) mimicking how people look at 2D images have been proposed [2, 3, 4] and have already proven to be valuable in several applications such as compression [5], content-based image retrieval [6], quality assessment [7] and retargetting [8].

Most of VS research has been focussing on the visual perception of images and videos displayed on 2D screens. More recently, however, it has been shown that humans look differently at images displayed in 3D compared to 2D [9, 10] and subsequently, a few algorithms for 3D saliency (3D-VSAs) have been proposed [11, 12]. The lack of ground truth has limited the comparison of VSAs in the past and recently Wang et al. [12] published an eye-tracking database with stereoscopic 3D images to act as ground truth to test VSAs. Using this database, they have proposed a comparison of several 2D-VSAs and 3D-VSAs [12]. We first propose to extend their study to include the Graph-Based Visual Saliency algorithm (GBVS) [2] and we show that GBVS performs better than other 2D-VSAs (Section 4). We then propose several extensions to GBVS to make it more effective for analysing 3D stereoscopic images (Section 3). We show that our new algorithms (3D GBVS) outperforms other state of the art 3D-VSAs [12] (section 4). We start next with a review on VSAs and the metrics used to assess their performances.

2. REVIEW
Visual saliency can be broken down into two groups: bottom-up and top-down [13] [14]. Bottom-up saliency is the preattentive, involuntary phase before any prior knowledge related to the task at hand or personal factors come into play that can affect attention guidance. Top-down saliency is highly dependent on semantic information and the current task being performed. For example, if we need to locate a red dragon in a box of toys, we would naturally scan the scene first for red objects and from these possibilities locate the desired toy.

This paper focuses on bottom-up saliency algorithms that are more generic (i.e. not dependent on the nature of the task) and faster to compute than top-down ones. Section 2.1 presents algorithms that do not incorporate any knowledge about the 3D nature of the scene depicted in the images (e.g. no depth or disparity is used to compute saliency). These algorithms are called here 2D-VSAs and Section 2.2 presents several extensions to these methods that include 3D information about the scenes (hence noted 3D-VSAs). Metrics have been introduced to measure the performance of VSAs using VS as ground truth captured with eye tracking technologies. Section 2.3 presents a few such metrics used in our experiments (Section 4).

2.1. 2D Visual saliency algorithms
Itti et al. [15] proposed an algorithm (noted 2D Itti) that first calculates features (i.e. colour, orientation and intensity) at multiple scales in an image. An activation map (an initial saliency map) is then calculated in parallel for each of these
channels. Next, the activation maps are combined into a master saliency map. Finally, a ranking of salient regions is computed through the use of a Winner-Takes-All (WTA) network.

Hou et al.'s approach [16] (noted 2D Hou) analyses the log-spectrum of an image based on luminance only and extracts the spectral residual in the spectral domain. A saliency map is then created from this information. The algorithm is assessed in an object detection task and is shown to outperform Itti's method [16].

Bruce et al's algorithm [11] (noted 2D Bruce) is derived from efficient coding and information theory and is also shown to outperform Itti's algorithm. It is based on the premise that localised saliency computation serves to maximise information sampled from one's environment. Content of interest, hence, are areas where there is the most amount of 'surprise' and self-information.

Proposed by Harel et al. [2], the GBVS algorithm is a 2D VSA that uses a Markovian approach to calculate its saliency maps. The first step in this algorithm is to break up the input image into the following feature channels: colour, intensity and orientation. Salient regions are then located in each of these channels by computing:

\[ d(i, j)(|p, q|) = \log \frac{M(i, j)}{M(p, q)} \]  

where \( M(i, j) \) is the value of the pixel \((i, j)\) in the feature map \( M \) (i.e. in the feature channel of colour, intensity or orientation). Following this, a fully connected graph \( G_A \) is created by connecting every node with all other nodes in each \( M \). The edge of each node connection from \((i, j)\) to \((p, q)\) is assigned the following weight:

\[ w((i, j), (p, q)) = d((i, j), |(p, q)|) \times F(i - p, j - p) \]

where

\[ F(a, b) = \exp \left( -\frac{a^2 + b^2}{2\sigma^2} \right) \]

and \( \sigma \) is a free parameter set to approximatively one tenth to one fifth of the image width. All the weights, \( w \), are, hence, proportional to their dissimilarity and distance in \( M \). A Markov chain is then defined over \( G_A \) by treating nodes as states and edge weights as transition probabilities. The equilibrium distribution calculated for each \( G_A \) reflects the time a random walker would spend at each node. Higher values are obtained for nodes with higher dissimilarities with their neighbouring nodes. This is because it is more likely to transition into subgraphs with lower similarity measures. This algorithm has been stated as being biologically plausible, meaning that its method for calculating saliency is based on psychophysical and physiological evidence [2].

2.2. 3D Visual saliency algorithms

Stereoscopic images displayed on 3D screens allow us to immediately perceive depth information [17]. Incorporating depth or disparity information in the calculation of saliency is therefore a natural extension to VSAs for automatically analysing 3D content. Two strategies have been proposed to use depth information.

Chamaret et al. [18] have proposed to multiply the saliency map computed with any 2D VSAs by the inverse of the depth map. Regions appearing closer to the viewer are then made more salient. Indeed, close areas have been shown to be viewed more often than regions further away [10]. These regions, hence, should be deemed more salient. Chamaret et al. then used this saliency calculation to refine a single region-of-interest that was selected earlier through a nearest-neighbour filtering and thresholding approach.

Wang et al. [12] explore two common ways to include depth information. The first is the depth-weighting model that weights 2D saliency computations with a corresponding depth map. The second is the depth-saliency model that creates a depth saliency map (DSM) by first looking for features in the depth map and then linearly pooling this with 2D VSA computations. Wang et al.'s DSM calculations included a Difference of Gaussian (DoG) filtering step and then correlation of the contrast map with the degree of depth saliency through the use of results obtained from a psychophysical experiment of theirs. Through quantitative experiments, Wang et al. [12] show that the best algorithm for incorporating depth information was to add 2D VSAs saliency map with DSM. They also confirm quantitatively that 2D Bruce outperforms 2D Hou (and that 2D Hou outperforms 2D Itti) and this ordering remains true in 3D when adding the DSM. These algorithms proposed by Wang et al. are noted 3D Hou, 3D Itti and 3D Bruce. Adding DSM provided the best results but was the most computationally demanding. Alternatively, Chamaret’s method to include depth information performed poorly but was the most computationally efficient.

2.3. Metrics for VSAs

Several metrics have been introduced to measure the performance of VSAs. In our experiments, we used the Pearson Linear Correlation Coefficient (PLCC) [19, 20] and Kullback-Leibler divergence (KLD) [21, 19], which were also used in Wang et al.’s study [12]. These measures compare the saliency maps obtained from saliency algorithms with fixation density maps ( maps created from eye-tracking experiments).

The PLCC measures the linear correlation between the saliency and fixation maps \( H \) and \( P \):

\[ \text{PLCC}(H, P) = \frac{\text{Cov}(H, P)}{\sigma_H \sigma_P} \]

where \( \text{Cov}(H, P) \) is the covariance and \( \sigma_H \) and \( \sigma_P \) denote the standard deviations of \( H \) and \( P \) respectively.

The KLD calculates the dissimilarity between normalised saliency and fixation maps (to be understood as two probabil-
ity density functions (PDFs):
$$\text{KLD}(H, P) = \sum_x h_x \ln \left( \frac{h_x}{p_x} \right)$$  
(5)

where \(h_x\) and \(p_x\) denote the values of the normalised maps of H and P respectively at pixel location \(x\).

### 3. GBVS EXTENSIONS TO 3D MEDIA

The GBVS algorithm has never been extended to 3D and we propose to incorporate depth information in three ways: using Chamaret et al.’s approach [18] (noted 3D GBVS (Chamaret)), using the DSM as proposed by Wang et al. [12] (noted 3D GBVS (Wang)), and using our own approach (noted 3D GBVS (our approach)).

This last method (3D GBVS (our approach)) involves a three-step process: selecting a low or high scaling factor corresponding to depth values, restriction of the depth-range that saliency values are affected by this scaling factor and then the subsequent scaling of these saliency values. The first step entails calculating the median and mean values of the depth map. If the median is smaller than the mean, a higher scaling factor is chosen (\(SF\) in eqs 7 and 8) and vice versa otherwise. The idea behind this is to detect images that have unique objects in the foreground and hence will stand out on their own in 3D. A smaller median value compared to the mean would indicate the presence of unique objects and should only be used on areas with high-enough saliency and should not be ‘wasted’ elsewhere. If salient regions are only present in the foreground, the competition for scaling should only take place there.

### 4. EXPERIMENTAL RESULTS

We used the eye-tracking database supplied by Wang et al. [12] as ground truth. This database contains 18 stereoscopic images, eye-tracking data obtained from 35 human subjects, corresponding depth and disparity maps and eye fixation maps. This database serves as a ground truth and is noted 3D VS because the stereoscopic images were displayed on a 3D screen (as opposed to 2D VS ground truth that collects eye tracking data with the image displayed on a 2D screen).

#### 4.1. Evaluation of 2D VSAs against 3D VS

In Table 1, we compared the 2D GBVS algorithm on this dataset with Itti’s, Hou’s and Bruce’s algorithms using the PLCC (defined in Eq. (4)) and KLD (Eq. (5)) metrics. Note that the 2D GBVS algorithm was not analysed by Wang et al. [12] but they did compute the PLCC and KLD for 2D Itti, 2D Hou and 2D Bruce. Our numerical results slightly differ from theirs due to rounding errors and different image scaling algorithms. Our results in Table 1 confirm Wang et al.’s assessment for 2D Itti, 2D Hou and 2D Bruce. Our contribution in Table 1 is to show that 2D GBVS algorithm outperforms all other 2D VSAs by far.

<table>
<thead>
<tr>
<th></th>
<th>PLCC</th>
<th>KLD</th>
<th>Yr of publication</th>
</tr>
</thead>
<tbody>
<tr>
<td>2D Itti</td>
<td>0.154</td>
<td>2.781</td>
<td>1998 [15]</td>
</tr>
<tr>
<td>2D Hou</td>
<td>0.299</td>
<td>0.877</td>
<td>2007 [16]</td>
</tr>
<tr>
<td>2D Bruce</td>
<td>0.346</td>
<td>0.704</td>
<td>2009 [11]</td>
</tr>
<tr>
<td>2D GBVS</td>
<td>0.589</td>
<td>0.314</td>
<td></td>
</tr>
<tr>
<td>UTPL [12]</td>
<td>0.897</td>
<td>0.127</td>
<td></td>
</tr>
</tbody>
</table>

**Table 1**: Performances of 2D VSAs. Higher PLCC and lower KLD values indicate better performances.

The Upper Theoretical Performance Limit (UTPL) [22] computed by Wang et al. [12] is also reported for both PLCC and KLD. The UTPL is commonly used as a benchmark for 2D visual saliency models, and 2D GBVS is halfway in performance between this theoretical limit and the best 2D VSA previously reported (2D Bruce). Figure 1 (c)-(f) shows some saliency maps for these 2D VSAs.
4.2. Evaluation of 3D VSAs against 3D VS

Our 3D VSAs are also evaluated on the same dataset. Wang et al. found their proposed 2D+DSM approach to be the best way to incorporate depth in saliency calculations, and this method is referred to as (Wang) in Table 2. We implemented their approach for all the four 2D VSAs as well as Chamaret’s and our own proposed method with the GBVS algorithm. Table 2 confirms the results for Itti (Wang), Bruce (Wang) and Hou (Wang) as reported by Wang et al. [12].

Our contribution in Table 2 is to show that the three 3D VSAs we proposed outperform all 3D VSAs proposed by Wang et al [12]. Surprisingly, Chamaret’s method in the GBVS fared better than Wang’s algorithm (that is also the most computationally demanding method). Note however that 3D GBVS (Chamaret) and 3D GBVS (Wang) are outperformed by 2D GBVS (cf. Table 1). Our simple and fast proposed method for incorporating depth obtained the best results. Figure 1 (f)-(i) shows example saliency map results for 2D GBVS, 3D GBVS (Wang), 3D GBVS (Chamaret) and 3D GBVS (Our method). An improvement on 2D-GBVS can be seen in 3D GBVS (Chamaret) and 3D GBVS (Our method) - both are closer to the ground truth.

<table>
<thead>
<tr>
<th>3D Saliency Algorithms</th>
<th>PLCC</th>
<th>KLD</th>
<th>Yr of pub.</th>
</tr>
</thead>
<tbody>
<tr>
<td>3D GBVS (Chamaret)</td>
<td>0.573</td>
<td>0.379</td>
<td></td>
</tr>
<tr>
<td>3D GBVS (Our method)</td>
<td>0.606*</td>
<td>0.306*</td>
<td></td>
</tr>
<tr>
<td>3D GBVS (Wang)</td>
<td>0.561</td>
<td>0.484</td>
<td></td>
</tr>
<tr>
<td>3D Itti (Wang)</td>
<td>0.364</td>
<td>0.627</td>
<td>2013 [12]</td>
</tr>
<tr>
<td>3D Bruce (Wang)</td>
<td>0.419</td>
<td>0.657</td>
<td>2013 [12]</td>
</tr>
<tr>
<td>3D Hou (Wang)</td>
<td>0.436</td>
<td>0.558</td>
<td>2013 [12]</td>
</tr>
</tbody>
</table>

Table 2: Performances of 3D-VSAs. Higher PLCC and lower KLD values indicate better performance. * indicates significant difference from the 2D model (paired t-test, p<0.1 [12]).

5. CONCLUSION

In this paper we demonstrated that the GBVS algorithm is greatly superior in predicting fixations in 3D compared to other state of the art algorithms. We also showed that our simple and fast extension to Chamaret’s method of depth incorporation outperforms all other depth incorporation methods analysed by Wang et al. Future work will investigate the application of our 3D VSAs in areas such as content-based image retrieval, compression and image segmentation.
6. REFERENCES


ABSTRACT

The increasing use of digital media in daily life has resulted in a need for novel multimedia data analysis for forensic purposes. Processing multimedia documents for police investigations or as pieces of evidence in courts of law entails that their interpretation must be reliable, trustworthy, and efficient in terms of human time and other required resources. Accomplishing this task will help greatly to speed up investigations and judicial processes, making them more effective.

In this paper, we present our views on processing techniques for varied multimedia sources. In particular, we focus on speech, video and images, handwriting, and text documents. The new techniques we propose will form a toolkit to be used in law force investigation or in a court of law, to extract information and evidence from different multimedia sources. This toolkit will support the user in a way that will lead to standardisation and (semi-) automation in the analysis of forensic data.

Index Terms—Multimedia, Forensic Data Analysis, Image Enhancement, Feature Extraction, Case-Based Reasoning

1. INTRODUCTION

The increasing amount of digital documents used in daily life has resulted in the need for novel multimedia data analysis techniques to enable their use in forensics. This means that their origin and integrity must be confirmed, and the data interpretation must be robust and reproducible. Furthermore, it is necessary to extract knowledge from large as well as sparse sample sets. Reliability and trustworthiness are basic requirements for all the data used as legal evidence. The processing techniques used to obtain such data must ensure these requirements and must also be efficient in terms of human time and other required resources. Developing techniques fulfilling these requirements will help greatly to speed up investigation and judicial processes, and to make them more effective.

Multimedia content comprises images, videos, text and audio. In addition, social media (e.g. Facebook, Twitter) [5] provide multimedia data in new formats. Such data allow the investigator to identify and compare objects, events or even human beings based on data properties like biometric features or more symbolic features that can point to coincidences and anomalies.

Our consortium is proposing a program named NPRF [1], which will develop new multimedia processing techniques to be used in forensics. In particular we will focus on speech, video and images, handwriting, and text documents. The new techniques will form a toolkit to be used in law force investigation or in a court of law, to extract information and evidence from different multimedia sources. This toolkit will support a law force user in a way that will lead to standardisation and (semi-) automation in the analysis of forensic data. It will address different standard tasks during forensic preparation for a court of law. As a consequence, our project will help to standardise law court investigations. Moreover, the developed methods will allow the storage of analysed cases in a database so that they can easily be retrieved, generalised, and comparatively used for new cases, by means of specifically designed techniques.

In Section 2, we describe the motivation and objectives of this work. The overall system architecture is given in Section 3. The data used for our study and the methods that should be developed are described in Sections 4 and 5, respectively. Finally, we give conclusions in Section 6.

2. MOTIVATION AND OBJECTIVES

The goal of NPRF is to provide new methods and a toolkit for automatic forensic multimedia data analysis. The data modalities considered are images and videos, text, handwriting, speech and audio signals, and social media data. The integration of methods for different media in one toolkit allows cross-analysis and event detection by interlinking different data modalities. The proposed methods will face standard forensic tasks, such as identification of events, persons or groups, and device recognition. We will work out new standard tasks, thus helping standardisation in forensic data analysis.

2.1. Signal Enhancement

In image applications, the data represent a degraded version of an original scene. Degradation may appear in different forms and must be removed before the images are used for classification or decision-making. Audio signals are also normally affected by degradations. Forensic applications need images, videos and audio of improved
quality, as well as free of forgeries or other artifacts. In other words, both the quality and the integrity of these data when used as proof of evidence must be clearly assessed. Techniques for detecting artifacts in images, videos and audio files are of paramount importance to this project. To trust the information extracted from these data, it is necessary to make sure that they have not been manipulated.

2.2. Classification and Contextual Interpretation

Most classification methods use a-priori information about the class distribution and concepts. Although we are dealing with mass data, such information can seldom be assumed to be correct. As opposed to controlled-environment applications, we deal with an open-world scenario where constraints cannot normally be enforced and, furthermore, checking their validity is difficult. Consequently, classical statistical methods are not applicable, and novel methods that can incrementally learn and generalise over the data are necessary. Multimedia data are diverse in representation and content. Exploiting them jointly requires innovative data analysis techniques and data representations. The analysis of each media is already an enormous challenge. This is why the combination of all media has been hardly considered in the past, but it will pay off, producing novel methods and techniques that will improve significantly the performance of many forensic and security oriented applications.

2.3. Novelty Detection and Case-Based Reasoning

Novelty detection is another important task in NPRF, since in forensic applications concepts like victims or events are not very well defined. By allowing new tasks to be identified, the project will provide automatic methods to improve standardisation in the analysis of forensic data. We will also develop learning methods to include new data into existing cases and to summarise new and old cases into more general cases. This approach will provide a solution to handle terabytes of data. To this end, a novel case-based reasoning (CBR) system will be developed, based on probabilistic and similarity-based methods, to keep the cases in a case base, so that they can easily be retrieved and utilised in new situations. A special taxonomy to evaluate new similarity measures will be developed and implemented in the CBR system. It will provide explanation capabilities for similarity, thus helping a forensic data analyst to identify the right reasoning method for a particular problem. Probabilistic Reasoning in classification and novelty detection will also be used as a complementary/alternative approach to CBR.

2.4. System Test

The developed toolkit will be tested on real data, provided by the end-user partners. This test phase will provide a very important feedback, which will be used to improve the system.

3. THE CONCEPT OF A FLEXIBLE SYSTEM FOR FORENSIC DATA ANALYSIS

The idea is to come up with an automatic system that covers all the aspects of multimodal and multimedia data analysis, from signal enhancement and pre-processing, to feature extraction, analysis, and interpretation. Image enhancement techniques eliminate degradations that might appear because of a known or unknown blur function, a low-resolution capture or because of possibly severe compression artifacts. An enhancement procedure in these cases consists of non-blind or blind deconvolution (to reduce the effects of known or unknown blur), or superresolution, by which a combination of several low resolution images is used to obtain a high resolution one. To trust the information extracted from images and videos, techniques will be developed to assess that the image/video has been recorded by a camera and no artifact has been added.

Feature extraction consists of selecting a sufficient set of low- and high-level features to complement the existing standards for image, video and audio data, aiming at enabling novel and robust classification and recognition methods. These should allow for a modelling of the standard tasks in forensic data analysis, but should also be flexible enough to cover the needs of newly appearing tasks. Twitter was actively used by rivaling gang members to plan their assaults. Twitter data is hard to analyse because the text fragments are very short, multiple persons can be involved in a conversation about various topics and the data is rapidly changing. Novel methods are necessary to monitor Twitter in real time and identify potential threats, including individuals and communities who are planning illegal activities. We plan to build a dynamic model on Twitter text to forecast the upcoming significant events and emotions of the crowd associated with these events. While there can be many events with strong presence in Social Media, some events could be associated to stronger negative emotions. These events are candidates for further investigation for potential criminal nature or significant social consequences.

The huge amount of CCTV systems has increased the importance of video and image evidence in forensic labs. An automatic system should be able to detect people faces, vehicles, license plates, guns, dresses and all other objects that can link a person to an event. An important focus of police work is the identification of people for which a decision of a public prosecutor's office or a judge applies or an arrest warrant is issued. Within this scope, image and video technology can easily be used in monitored places and facilities. One of the aims of NPRF is to develop methods and procedures for an automatic identification of one or several target people in mobile video recordings based on passport photos or other available pictures.
A significant portion of data collected by Law Enforcement Agencies consists of speech and audio files. They form an important part of legal cases. Speech recognition systems (such as dictation systems) are now available in many languages. However, continuous spontaneous speech recognition is still an unsolved problem. The proposed project will develop new methods for the recognition of continuous spontaneous speech and other audio signals, with the help of compressive learning and case based reasoning methods.

While the off-the-shelf optical character recognition systems are very successful for printed documents, recognition of words in unconstrained settings or “in the wild” still is an open problem and recognition of handwritten text continues to be a challenge. We propose to develop novel Handwriting Recognition Methods for unconstrained settings.

The information extracted from the different media should be used to build the evidence for a case. To select and relate the pieces of information that can actually be used in a law court, we must be able to classify, categorise, identify, recognise, fuse and interpret these data. In a word, we must provide them with semantics. Based on the consortium expertise, we will mainly rely on Case-Based Reasoning (CBR) and will also apply probabilistic modelling and inference when necessary. CBR explicitly uses past cases from the domain expert’s successful or failing experiences, and is very useful in applications where generalised knowledge is lacking. Therefore, CBR can be seen as a method for problem solving as well as a method to capture new experience and make it immediately available for problem solving. It can be seen as a learning and knowledge discovery approach, since it can use new experiences to capture some general knowledge such as case classes, prototypes and some higher-level concept. All these issues make CBR very useful for the analysis of forensic data. The system is able to capture new cases and store new and old cases in a summarised way so that they can easily be retrieved and used for reasoning. The reasoning methods to be used are based on similarity, which makes it very useful to detect and identify similar and identical cases with no generalised knowledge available. Different similarity measures will be developed that can deal with the different modalities of data and their case representation. A taxonomy of similarity will be developed to help forensic data analysts to apply these reasoning methods to their problems by explaining the relation, usefulness, and application of the different similarity measures. Together with CBR, we will use Probabilistic Reasoning: as security and forensic applications involve a high level of uncertainty, Probability theory can be applied. We will use the language of graphical models to specify our models in a concise and intuitive way and perform inference on them. A system overview is given in Figure 1.

4. FORENSIC MULTIMEDIA DATA

Different types of forensic multimedia data will be used provided by End-Users:

Passive millimetre-wave (PMMW) images and video are used for security screening as many materials, including clothing, are transparent to millimetre-waves. The imagers that use this technology are therefore installed at security checkpoints to screen people for hidden weapons (including powders, liquids and gels) and contraband. They are characterised by a low-resolution compared to visible images, due to the wavelength used.

The current software automatically detects objects within the spatial and thermal resolution of the system and draws a red box around them. These are then represented at the
approximate locations on a generic silhouette to preserve the subject’s privacy.

However, object classification to automatically distinguish between a threat and a non-threat object is not currently performed. Within this project, a new system will be developed to make a classification based on the shape and size of the objects detected in the raw mm-wave image. This would reduce the number of false alarms.

Anonymous Data from Text will be collected. This data is freely available on the Web. We propose to perform initial experiments on anonymised data to validate the feasibility of our approach. After authorisation of the responsible superiors of the cybercrime unit is obtained, we will use the developed system for real life investigations. Twitter data will be provided by end users.

A telecom provider will prepare a speech database obtained under various conditions and under various speech coders and encoders to test the new algorithms.

Video and Image databases with case scenarios will be provided by police forces.

Handwriting documents will be collected through the involvement of graduate and undergraduate students. We also plan to use the benchmark data set IAM Database for Off-line Cursive Handwritten Text http://www.iam.unibe.ch/~zimmerma/iamdb/iamdb.html. This database contains forms of unconstrained western handwritten text. It includes 27,000 isolated words (400 pages).

5. THE METHODS USED TO SOLVE THE TASK

The huge amount of digital data to be analysed in forensic practice to identify and collect evidence requires human experts to spend a lot of time to do their job. Ambitiously, NPRF wants to help analysts to perform their tasks with nearly any kind of digital media.

Our scientific and technological objectives are firstly subdivided by type of media: still/motion images, audio/speech signals, and text documents. Secondly, we will derive and extract from the media files the meaningful features, to be organised in an ontology, thus supporting recognition and reasoning. NPRF has a third wide objective: to use statistics, semantics, and machine learning to build a case-base for storing experience so that past cases can be summarised, categorised and generalised, thus allowing CBR and probabilistic inference to be applied, to study new cases by exploiting the stored experience.

To guide our technological choices towards the overall goals, we need a clear picture of the current legal framework regulating the process of gathering, processing, analysing and integrating multimedia data for security and judicial purposes at European (EU Directives and Regulations) and national (EU wide) level. To this end, the legal experts in the Consortium will initially i) investigate on legal issues related to the use of data for forensic purposes by the analysis of case studies and EU and national judgements, ii) provide a framework of standards, quality indicators and approaches to preserve and assess the validity of digital evidence for forensic purposes, iii) help the other partners in sorting out potential ethical issues.

5.1 Media processing

The following methods [3] for media processing will be developed:

- Methods and algorithms for super-resolution of specified details in image streams.
- Methods and algorithms for artifact reduction.
- Methods and algorithms to assess data integrity.
- Improved-quality images or video sequences, to support classification or recognition (e.g., biometry, and handwritten character/word recognition).
- Automatic video processing tools supporting the expert in selecting the portions of videos that might contain the interesting facts. As people may appear in non-frontal poses, we will exploit the presence of multiple cameras in a given area to track a person and make the identification when the face is in frontal position.
- Signal processing techniques to allow distinctive features to be extracted from audio/speech signals, using both temporal and frequency domain information.

5.2 Feature extraction and organization

For feature extraction and organization [4] we will develop the following methods:

- Methods to extract features to detect, describe and relate the multimedia content relevant to forensic activities. Aiming at recognition in the wild, focus will be given to the definition and verification of features that enable detection and recognition in unconstrained conditions and environments. This means that feature invariance to different conditions, as well as robustness to noise, are fundamental issues that will be tackled. We will focus on biometric parameters that characterise individuals in terms of appearance, behaviour, voice and handwritings, so as to enable the process of detection and recognition. More in detail:
  - Face characteristics, distinctive individual marks, morphometric measurements, gait analysis.
  - Distinctive audio features for speaker identification and recognition.
  - Characteristic textual features, with particular emphasis on texts written in non-conventional languages (e.g., transcriptions of speech recognisers, and the language of social media) [6]. The languages considered include Dutch, Italian, Hungarian, English, and Bulgarian.
• Methods based on similarity, CBR, and compressive reasoning and sensing for speech and audio recognition using both temporal and frequency domain information.
• Similarity-based (possibly word-based) methods for automatic recognition of handwritten text and writer identification. As an alternative to character-based systems, word-based systems will be developed.
• An ontological model [4] to organise and categorise all the features collected, based on (or extending) existing metadata standards for image, video, audio, and textual data.
• A feature ontology, integrated with a library of algorithms for feature extraction, enabling:
  • A standardised feature definition and computation,
  • The cataloguing of features, and
  • The modelling of the multimedia data analysis domain.
  • A toolbox for feature extraction.

5.3 Case-Based Reasoning

For classification and interpretation by case-based reasoning [2], the following methods will be developed:
• Methods and techniques for supporting human experts in analysing social media textual data.
• A hierarchical case-base built and organised through different and event-specific features, from which complex cases can easily be retrieved and applied to new situations.
• Learning methods to include new data into the existing cases and summarise new and old cases into more general cases applicable to a wider range of tasks for further law purposes.
• A CBR system consisting of novel probabilistic and similarity-based meta-learning methods, with a special taxonomy for similarity determination. This will offer explanation capabilities and help a forensic analyst to identify the right reasoning method for any particular problem. The lifetime of this system will be considered through special case-base maintenance methods, a modular system architecture, and software engineering rules.
• A CBR-based novelty detection method, considering the incremental nature of the data, to handle novel cases for immediate reasoning or new model build-up. This will also allow novel situations and tasks to be identified, thus providing an automatic method to improve standardisation in forensic data analysis.

6. CONCLUSIONS

We have described our view to novel intelligent forensic multimedia data analysis. This includes besides the technical aspects the legal aspects that are going along with brisant data analysis of security data. The concept of such a system has been given and the advantages have been described. The data used in our group are described. Furthermore the techniques we are studying to solve the analysis of the particular multimedia data sources have been described. The techniques to solve the overall media integration and intelligent data analysis by CBR have been explained. We are constantly studying the particular task and hope to bring this approach into a larger project so that we can develop the full system.

ACKNOWLEDGMENT

Composition of the NPRF consortium:
GEIE ERCIM, Europe; Institut für Bildverarbeitung und angewandte Informatik, Germany; Consiglio Nazionale delle Ricerche, Italy; Universidad de Granada, Spain; Bilkent Universitesi, Turkey; Università degli Studi di Cagliari, Italy; Telecommunication Systems Institute, Greece; Azienda per la Mobilità del Comune di Roma, Italy; ImageInterpret, Germany; Centre d'Excellence en Technologies de l'Information et de la Communication, France; Alfa Imaging, Spain; Data and Control Systems, Greece; ibai solutions Prof. Petra Perner, Germany; Turk Telekomunikasyon, Turkey; Regiopolitie Amsterdam-Amstelland, The Netherlands; Thüringer Polizei, Germany.

REFERENCES

[1] New processing and analysis techniques for forensic data nprf, project propopsal, IBaI Report, November 2012
KEY INDICATORS FOR MONITORING OF AUDIOVISUAL QUALITY

M. Leszczuk(1), M. Hanusiak(1), I. Blanco(1), A. Dziech(1), J. Derkacz(1), E. Wyckens(2), S. Borer(3)

(1) AGH University of Science and Technology, Kraków, Poland
(2) Orange Labs, Rennes, France
(3) SwissQual, Zuchwil, Switzerland

ABSTRACT

Automating multimedia quality checking is currently based on finding major audio/video artifacts. The “Monitoring Of Audio-Visual quality by key Indicators (MOAVI) subgroup of the Video Quality Experts Group is an open collaborative project for developing No-Reference models for monitoring quality. MOAVI is a complementary, industry-driven alternative to Quality of Experience used to automatically measure quality by using simple indicators of perceived degradation. The goal is to develop a set of key indicators describing quality in general, and to select subsets for each potential application. Therefore, the MOAVI project concentrates on models based on key indicators contrary to models predicting overall quality.

Keywords: Video, quality, metrics, artifacts, VQEG

1. INTRODUCTION

In Video-on-Demand (VoD) uses, once content is created, it should be checked against Quality of Experience (QoE), before sending to delivery servers. The check should happen quickly, in order to avoid unnecessary broadcasting delays. If the content is delivered without checking, the risk of broadcasting video sequences with artifacts is increasing.

For live uses, the challenge is even larger because the audio-visual monitoring should raise the alarm before observing the number of customers calling the call center. This is especially important in the case of the big events when the globe browser is needed by the customers.

The QoE issue is also addressing the telecommunications regulatory bodies wanting to analyze the quality between broadcasters and Internet Service Providers (ISP). Both quality supervision staff and customers want high and reliable QoE. When the disturbance happened, the artifacts should be detected in real-time and all the time.

Applicability of Full-Reference (FR) QoE models for this purpose is limited, as they require the un-encrypted, reference audio-visual signals without any artifacts [1]. Resynchronization models, which normally take place to produce reliable Mean Opinion Score (MOS), are under-performing when the processed video is encoded with resolution or frame rate significantly lower than the reference. Moreover, the most of re-synchronization models cannot support some disturbances, if happening close to video sequence boundaries. Due to processing time between video head-end and the end user terminal, it also becomes really difficult to re-synchronize the video for FR model. Finally, the many major industries even fail to provide master contents for checking purpose.

No-Reference (NR) QoE models don’t cover all the problems as well. The models, like the reported in [2], address some quality measurement techniques, using objective parametric models. Still, these models have problems in predicting overall audio-visual QoE. These models are usually limited to very basic artifacts. Yet, in realistic situation, the artifacts are far more sophisticated. Therefore, for example, current NR models predicting the MOS cannot provide any alarm on lip sync, ghost effects, epileptic frames, etc. [1]. Furthermore, parametric models can provide a false positive alarm, if the video databases for training the models, were not made with an emphasis to the realistic architecture of the broadcasters. However, generally, the broadcasters don’t provide conception details of their audio-visual chain. And, finally, when the content is encrypted it obviously becomes impossible to use this kind of measurement approach.

Another issue to be considered is computational complexity. QoE models are limited by CPU performance per video channel. Consequently, the video channel cannot be analyzed for long time, and numerous artifacts can be missed.

Therefore a complementary, industry-driven alternative, used to measure the quality automatically by using simple perceived indicators, can be proposed. The essence of the proposed concept is to use a simple modeling (that is potentially more accurate, and therefore, industrially useful) binary indicators (thresholds) associated with the occurrence of perceptible degradation of the audio/video. The “Monitoring Of Audio-Visual quality by key Indicators” (MOAVI) [3] subgroup of the Video Quality Experts Group (VQEG)
[4], a project for developing NR models for monitoring service quality, started to develop such a set of key indicators1.

The paper is organized as follows. Section 2 describes related work. Section 3 presents MOA VI’s key indicators. Section 4 describes the recent results. Section 5 concludes the paper and drafts the future work.

2. STATE-OF-THE-ART BACKGROUND

This section presents limitations of current FR, Reduced-Reference (RR) and NR metrics for standardized models. Most of the models in the ITU-T Recommendations were validated using one of the following hypotheses: frame freezes up to 2 seconds, no degradation at the beginning or at the end of the video sequence are allowed, no skipped frames, the video reference should be clean (no spatial or temporal distortions), minimum delay is supported between video reference and video (sometimes with constant delay), the up or down-scaling operations are not always taken into account. Most models are based on measuring conventional blurriness, blockiness and jerkiness artifacts for producing predictive MOS.

The most of algorithms producing the predicted MOS scores is a mix between blur, block, and jerkiness metrics. The weighting between each indicator could be a simple mathematical function. If one of these indicators is not correct, the global predictive score is completely wrong. The other indicators are usually not taken into account in predicting MOS.

The ITU-T has worked a long time ago on MOA VI-like distortions [5]. Whereas the history regarding Recommendations is shown in Table 1, metrics based on video signal only are shown in Table 2.

<table>
<thead>
<tr>
<th>Model Type</th>
<th>Format</th>
<th>Rec.</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>RR</td>
<td>SD</td>
<td>J.144 [6]</td>
<td>2004</td>
</tr>
<tr>
<td>Bitstream</td>
<td>VGA–HD</td>
<td>In progress</td>
<td>Exp. 2013</td>
</tr>
<tr>
<td>Hybrid</td>
<td>VGA–HD</td>
<td>In progress</td>
<td>Exp. 2013</td>
</tr>
</tbody>
</table>

Table 1. The history regarding ITU-T Recommendations.

Closely related is a current standardization activity at ITU-T SG12 on models for multimedia and IPTV based on bit-stream information. SG12 is now working on models for IPTV. Q.14/12 is responsible for these projects, which are provisionally called P.NAMS (non-intrusive parametric model for assessment of performance of multimedia streaming) and P.NBAMS (non-intrusive bit-stream model for assessment of performance of multimedia streaming).

P.NAMS utilizes only packet-header information (e.g., from IP through MPEG2-TS), while P.NBAMS is allowed to use the payload information (i.e., coded bit-stream) [13]. However, this work has been focused on the overall quality (in MOS units), while MOA VI is focusing on Key Performance Indicators (KPI).

<table>
<thead>
<tr>
<th>Type of ITU-T Model</th>
<th>FR</th>
<th>RR</th>
<th>NR</th>
</tr>
</thead>
<tbody>
<tr>
<td>HDTV</td>
<td>J.341 [10]</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>SDTV</td>
<td>J.144 [6]</td>
<td>n/a</td>
<td>n/a</td>
</tr>
</tbody>
</table>

Table 2. Synthesis of FR, RR and NR MOS models.

The MOA VI project is meant to explore human behaviour on longer period and propose adapted model with enhanced SSCQE methods.

Most of the recommended models are based on a global quality evaluation of the video sequences as in P.NAMS and P.NBAMS projects. The predictive score is correlated to subjective score obtain with global evaluation methodologies (SAMVIQ, DSCQS, ACR, etc.). Generally, the duration of video sequences is limited to 10 s or 15 s in order to avoid a forgiveness effect (the observer is not enabling to score properly the video after 30 s and can give more weight to artefacts occurring at the end of the sequence). When one model is deploying for monitoring of video services, the global scores are provided for fixed temporal windows and without any acknowledgement of the previous scores.

3. KEY INDICATORS FOR AUTOMATED QUALITY CHECKING

Automating quality checking is currently based on finding major video and audio artifacts. The processing is performed on the video signal and/or the bit-stream. Quality checking can be conducted before, during, and/or after the encoding process. Figure 1 shows the concept of monitoring of audio-visual quality. However, in MOA VI, no MOS is provided.

MOA VI key artifact indicators are classified into four directories based on their origins: capturing, processing, transmission, and displaying.
3.1. Capturing

Artifacts introduced during video recording. Images and video are captured using cameras that comprise of an optical system and a sensor with processing circuitry. Artifacts based on capture will affect both analogue and digital systems as this is at the front end of the image acquisition. Reflected light from the object or scene forms an image on the sensor [14].

3.2. Processing

Processing is required to meet constraints such as bandwidth limitations imposed by the medium and to provide immunity against medium noise. There are many coding techniques for removing the redundancies in images and video. Coding can introduce artifacts such as reduced spatial and temporal resolution, blocking, blurring, and they are the common and dominant undesirable visible effects [14].

3.3. Transmission

When data is transmitted through a medium, some of the data may be lost, distorted or may result in multiple data due to reflections. When data arrives through many paths in addition to the direct path, the distortion is known as multi-path distortion and affects both analogue and digital communications [14].

3.4. Displaying

As the technology developed, different displaying systems offered different subjective quality with the same resolution. With nowadays displaying screens, the difference is reduced to the minimum between OLED, LCD, SED technologies. Much more visible artifacts include: block missing, stripe noise, aspect ratio error, photosensitive epilepsy flashing effect (ITU-R BT.1702 [15]), lip sync, blackout, and framing (pillar-boxing/letter-boxing).

4. RESULTS ON ARTIFACTS AND METRICS

This section describes the considered artifacts and the results obtained, including the developed MOAVI Proof-of-Concept (PoC). The nature and characteristics of the artifacts taken into consideration for the quality evaluation are described, and a description of the metrics (including the principles of operation) for each distortion is included. In order to look for feedback and evaluation of the algorithm designs, psycho-physical experiments have been conducted. The tests show good performance of the algorithms. The developed metrics, further information and some examples of usage are available publicly [16].

The artifacts are divided into 3 sections: video, audio-video, and audio. As work on the project is extendable, the list of artifacts is not closed.

4.1. Video Artifacts

Blocking is the most visible image and video degradation of all artifacts. The effect is caused by all block-based coding techniques. The metric of blocking the artifact is calculated for pixels at boundaries of $8 \times 8$ blocks [17]. Its value depends on two factors: magnitude of color difference at the block’s boundary, and picture contrast near boundaries [18]. The accuracy (the probability of correct classification) against ground truth is: 98.48%.

Blurring shows as reduced sharpness of edges and spatial detail. In compressed video, it results from a loss of high frequency information during coding. Measurement of this artifact is based on the cosine of the angle between perpendiculars to planes in adjacent pixels which is a good characteristic of picture smoothness [17]. The accuracy (the probability of correct classification) against ground truth is: 80.52%.

Ringing artifacts are visible for all compression techniques, especially when the image or video is transformed into a frequency domain. Ringing is a spurious reconstruction of pixel values. It is more evident along high contrast edges, especially if the edges are in areas with a generally smooth texture [18]. The noisiness metric estimates the noise level by a local variance of flat areas.

Exposure time distortions are visible as an imbalance in brightness (frames that are too dark or too bright). They are caused by an incorrect exposure time, or recording video without a lighting device. Mean brightness of the darkest and brightest parts of the image is calculated in order to detect the distortion. The results of the metrics mentioned above were mapped to the MOS. The thresholds were referred to the MOS scale, determining the score below which each distortion is noticeable [19].

Block missing occurs when some of the data packets forming the video signal are lost during one of the transmission stages. The metric relies on finding every single square shaped block with uniform color properties, and differences between the characteristics of that block and its adjacent blocks. The metric also tests whether the pixels in that block have values that fall within a narrow range. The number of blocks found indicates the visibility of the distortion.
Blackout shows as the picture disappearing; a black screen. It appears when all packets of data are lost, or as a result of incorrect video recording [20].

Freezing – stilted and jerky motion, frequently found during high motion as part of IPTV sequences, is seen as time-discrete “snapshots” of the original continuous scene strung together as a disjointed sequence. The detection of freezing is based on differences between consecutive frames. The visibility of both freezing and blackout artifacts, mainly depends on their duration. An initial threshold, based on the first tests developed, was set at 80 milliseconds for both artifacts, which stands for two consecutive frames in a 25 fps video. This is the shortest time during which viewers can perceive the presence of this artifact [21].

Pillar-boxing/Letter-boxing effect occurs when content is generated in a certain aspect ratio, but presented on a different format screen. Initial tests show that both letter-box and pillar-box effects are, for some viewers, unnoticeable even for large sizes (to one-tenth of an image) [20].

Interlace distortion is the result of special video compression where each frame is a connection between two frames in the original video. The interlace detection algorithm calculates the difference between consecutive pixels in columns [21].

Contrast is also stated as an important artifact for image/video quality assessment [22, 23, 24]. Contrast is the difference in luminance and/or color that makes an object (or its representation in an image or display) distinguishable.

4.2. Audio-Video Artifact

Lip sync artifact is the lack of synchronization or timing between the audio signal and the video signal. This means that the observer perceives that the speaking motion of the displayed person is not synchronized with the voice that he can hear. The lip sync artifact is the result of a different time of processing and transmission of sound and image, caused by a delay of one of the signals or in the synchronization signal. It becomes more noticeable during conversations, when the lips move while no voice is heard. The MOAVI metric for Lip Sync consists on the analysis of both the audio and the video information. A Voice Activity Detection (VAD) algorithm has been combined with some Speaker Recognition techniques to extract the precise moments when there is some voice in the video. The video will be analyzed to determine if there is lip activity and to identify the speaker. After that, some statistical tools will combine both sources of information to determine if an audio-visual content is synchronized or not. The accuracy (the probability of correct classification) against ground truth is: 95.8%.

Similar to the lip sync artifact, the synchronization of the sound and video is very important in certain problems such as a soccer game. If the commentary (audio) is perceived before the image sequence, it becomes disturbing for the audience to watch the game. An activity recognition of moving objects (player) may be developed to cope with such artifact.

4.3. Audio Artifacts

Clipping – the original audio signal can be clipped in certain situations during the recording due to environmental noise or recording equipment. The metric for this artifact is based on the detection of consecutive high levels of signal. As some conversions of amplitude may be introduced during the processing, transmission and displaying stages, clipped audio can be masked at low signal levels, although it remains annoying. To make the clipping detectable, the detection algorithm checks whether the audio file analyzed contains some of the common properties of both kinds of clipping, such as consecutive audio frames with low variation of signal value.

Mute – signal losses are one of the most common degradation in audio streaming at low bit rates. The detection algorithm is based on two different thresholds, once the signal has frequency components only in the hearing range (between 20 Hz and 20 kHz). These thresholds are: the minimum level of signal noticeable by the human ear, and the duration of the shortest silent interval perceptible as a differentiated mute. The value of these thresholds must be adaptive to the nature of the sound, as experimental test have demonstrated.

4.4. Proof-of-Concept

A MOAVI PoC has been developed. Figure 2 demonstrates the performance of the PoC. The key indicators are located in the upper-left side. When an artifact is visible its name is highlighted.

Fig. 2. Performance of MOAVI PoC
5. CONCLUSIONS AND NEXT STEPS

Having the algorithm design stage completed, the next aim of the research is to optimize the correlation between psycho-physical experiments with subjects (ground truth data) and the theoretically measured quantitative thresholds. Therefore, refining the thresholds is one of the critical points in MOAVI. Some final thresholds should be considered to successfully determine whether an artifact is noticeable or not, even if the designed algorithm detects that it is present. These thresholds should be strongly related to ground truth data. This ground truth should be obtained via subjective experiments following the ITU Recommendations on this matter.

There are plans to develop psycho-physical experiments in the near future. They are an approach to enhance the future strength of the methods and their correlation with the real subjective video quality. As the result, the thresholds are going to be contributed to the research community (by means of a published scientific paper).

6. ACKNOWLEDGMENTS

The research leading to these results has received funding from the Polish National Centre for Research and Development under Grant Agreement №ROB02201/ID/22/3 (TAPAS) and from the European Community’s Seventh Framework Program (FP7/2007-2013) under grant agreement №218086.

7. REFERENCES


Content-based Video Fingerprinting Method for Fast Key Generation and Retrieval

Shiyang Xu
School of Computer Engineering
Nanyang Technological University, Singapur
xu_shiyang@hotmail.com

Maria Trocan
Institut Superieur d’Electronique de Paris
28 rue NdC, Paris 75006, France
maria.trocan@isep.fr

Abstract—Video fingerprinting represents a content-based identification method wherein feature extraction is prominently used in order to verify the authenticity of a video stream. Most of works focus primary on the robustness of the fingerprint key only. However, with the advancement of technology where video content is often uploaded/downloaded at an amazing speed, there is high need of fast key generation and retrieval. In this paper, we propose a hybrid fingerprinting technique that allows both fast key generation and retrieval and proves good robustness for several types of video content alterations.

I. INTRODUCTION

Recently, rapid development of technology has led to an increased amount of video duplications which can be easily downloaded/uploaded across widespread media channels (e.g. Facebook, forums, YouTube, etc.). These video duplications, which are usually obtained in an illegal manner, often infringe the copyright of the video creator. As such, different methods of content-based video detection (CBVD) have been developed to address this issue. One such method is video fingerprinting. The key idea of this method is to extract important features (colour, edges, etc.) also known as fingerprints or signatures from the video; these extracted "fingerprints" / keys are then compared with a reference set to determine if it is an authorised or unauthorised copy of the reference video.

A good fingerprinting method should ensure certain traits namely robustness, pair-wise independent, speed and efficiency [2], [3], [4]. Two techniques have been widely used, namely dense and sparse fingerprinting methods [5]. In a dense scheme, every single keyframe will be stored in the database, while in sparse scheme, only selected keyframes that represent the salient content will be stored. Dense schemes provide more robustness since each key from the database is used for matching the “test” keys, but it is slower as there are more keys to search. Sparse scheme provide better search efficiency, as only keyframes will be used for matching, but it tended to be less robust [6], [7].

With the increasing piracy issues [8], numerous research works have been done in the CBVD field. Techniques that are commonly used include colour histogram, temporal measure and ordinal signature, [6], [9]. Most of these works are more focused on the robustness of the fingerprinting rather than other properties. Moreover, some of the above methods use dense evaluation to further improve robustness, but this also led to low efficiency and poor searching speed. Speed was never considered a priority in the past due to poorer video technology, fewer video modification software and limited video streaming websites. Thus, the concerns of piracy were mostly related to illegal downloading and pirated VCD. However, with the eminent popularity of YouTube and Facebook, video is streamed continuously [1]. Therefore, fast key matching [10] becomes a necessity to ensure pirated copy can be early detected.

In this work, we propose a hybrid method that ensure fast key generation and maintains the robustness of the fingerprint. The proposed method is based on both spatial and temporal analysis of the video. The proposed technique has numerous advantages. Firstly, sparse approach is used; therefore key generation and retrieval will be fast. Secondly, a wavelet approach is used for edge detection which reduces noise significantly and more real edges will be kept. Furthermore, wavelet transform is a local analysis, which increases flexibility in the level of details to be detected. Besides, by transforming a keyframe to various subbands, a segment of these subbands is generated as key, thus reducing the size of the key by a factor of 4. Lastly, cropping away unimportant information such as borders, logo, etc. will provide a smaller region of interest, improving thus the robustness.

This paper is organised as follows: Section 2 describes the proposed method. Experimental results obtained will be presented in Section 3. Finally, conclusions and future work are drawn in Section 4.

II. VIDEO FINGERPRINT GENERATION METHOD

This technique generates an identifier for a video sequence, which is robust to the most common video transformations, i.e. quality, resolution and framerate variations. Our proposed signature technique is divided into four steps:

• framerate resampling, which resamples the input video clips to a fixed framerate $F$ fps (15fps in our test framework),
• resolution resampling, in order to reduce the computational complexity (QVGA, 4/3 or 16/9 sample aspect ratio, depending on video input),
• three-step feature extraction and fingerprint generation. After the pre-processing steps, i.e. the framerate and resolution have been changed, the video is parsed and cut into
scenes. A scene is defined as a subsequence of the video, on which the fingerprint generation method is applied. The video stream fingerprint will be thus obtained by concatenating the keys (or sub-keys) assigned to each detected and processed scene in the original sequence. Figure 1 shows a video sequence separated into scene-cuts. Note here that fade-in and fade-out are not considered scene-cuts in our particular matter, only the abrupt scene changes will be processed. The scene-cut detection is realized using the following technique: for each frame, the variance of the intensity is calculated over the luminance component and is afterward compared to the next frame. If the value exceeds a given threshold $T$ ($T = 50\%$), a sudden-scene change is, therefore, detected (more refined methods can be used at this step for finer scene-cut detection).

Once the separation of the video stream into scenes has been done, a three-step spatio-temporal feature extraction will be applied in same manner on four keyframes inside each scene, keyframes found at the following temporal positions: the first one is defined on the scene transition and the last three at 25\%, 50\% and 75\% along the temporal axis of the scene. All the treatments explained in the following are applied to each scene of the video, in exactly the same way.

### A. Three-step feature extraction

Fast key matching is one of the important requirements for our video fingerprinting method, so the feature extraction as well as key matching is done into three steps. The first step generates a robust key and acts as a trigger for positive/negative matches, and the last two develop the decision made in the first step, by refining the generated key.

The edges contain important visual information. The variation of edges along temporal dimension is a specific feature to each video sequence. Strong edges of videos degenerated from the same source are almost the same. Thus our aim is to develop an appropriate edge response function and a measurement function to evaluate the temporal stability of edges from neighbor frames.

However, traditional edge detectors are not suitable for video signature, for some drawbacks: At first, noise easily inspires unwanted edges. Secondly, other attacks, like enhancement in color channel or spatial blurring, eliminate some edges. At last, the criteria of certain strength threshold to define edges can’t be applied to online media streaming, especially of different resolution level and compression quality. To solve the problem we propose a novel method to evaluate the stability of strong edges in the wavelet domain.

The first feature of our fingerprint is a spatial one and involves subband wavelet decomposition. Firstly, each frame is converted to a gray image. After wavelet transforming, each gray image is divided into four wavelet subbands: LL, LH, HL, and HH. LH and HL contain abundant horizontal edges and vertical edges respectively, HH containing the diagonal ones. The partial-spatial key (first step) is obtained to comparing the energy of these high-pass subbands resulting after one-level wavelet-transform.

The energies associated to the horizontal, vertical and diagonal subbands are computed, such that a classification of the detail orientation is obtained. Fig. 2(a) shows an example of such a classification, where the vertical subband has the most powerful energy (65\%), the horizontal subband is second (30\%) and finally, the diagonal one is the less energetic (5\%). We label the most energetic subband with 1 and less energetic with 3, and keep the statistics into a VHD mask. Over each detail subband, the energy is calculated as $E_i = \sum c^j_i$, where $c^j_i$ represents the wavelet coefficients in the $i^{th}$ subband. This classification results in a 6 bits-long spatial key for each keyframe in a subsequence, thus 24 bits feature representation for a scene.

Wavelet decomposition exhibits clearly the high frequencies at the first level of decomposition. It is, therefore, possible to locate the contours (whereas discriminating the single and isolated points and the textures), and track those contours in time, along the picture of the scene. The four selected frames are, therefore, needed to perform this tracking. For the 4 frames, we select a $C_m \times C_m$ block containing the strongest contours in the high-frequency subband having the first rank according to the classification criterion developed in the first step (Fig. 2(a)). This block will be translated into the spatial domain, obtaining thus a segment of $(2C_m)^2$ pels, and will be tracked into the next successive frame (considering only the 4 selected frames). The tracking is done by performing motion estimation and keeping the direction of this block.

We define $P$ ($P=2$ = 0) directions for motion vectors and the (0,0) state for no motion (in a first step we will consider $P=8$, as in Fig.2(b)). A basic motion estimation is realized for the $2C_m \times 2C_m$ blocks between any two consecutive frames among the selected 4 ones. The resulted motion directions are thus quantized following the defined $P+1$ directions. Note that
the direction $D$ for each block is given by $D = \arctan \frac{dy}{dx}$, this feature being represented with $4 \log_2 P + 1$ bits. In our test framework, $C_m = 32$ pixels.

The last refinement of our fingerprint is given by two statistical features, skewness and kurtosis, computed on each keyframe within a scene. In order to reinforce the robustness of our proposed technique, we are considering 10% cropping on each side of the keyframes.

The moments calculation is performed on the wavelet coefficients of each spatial detail subband resulted from the first step. The skewness criterion shows the distribution asymmetry; in the same way, the kurtosis reflects the flatness of the distribution. If we denote by $c_{ij}$ the wavelet coefficients in each subband and $E(c) = \frac{1}{N} \sum c_j$ is the expected value (mean) of the coefficients, the moments are obtained as:

$$M_{\text{skewness}} = \frac{[E(c - \bar{c})^3]}{[E(c - \bar{c})^2]^3},$$

$$M_{\text{kurtosis}} = \frac{[E(c - \bar{c})^4]}{[E(c - \bar{c})^2]^2}.$$  

For both moments, we are interested in keeping the sign for each one per spatial subband. We use thus 2 bits per spatial subband to classify the distribution (positive, negative or zero) in each direction (V, H and D). In total, we need 12 bits per keyframe, thus 48 bits per scene, for this last step of fingerprint refinement. The total size of the fingerprint for a video sequence is thus obtained by multiplying the needed bits for features representation at each step with the total number of scenes in the sequence, i.e.:

$$\text{Size fingerprint} = \#\text{scenes} \times (24 + 4 \log_2 (P + 1) + 48).$$  

III. EXPERIMENTAL RESULTS

A. Dataset and simulation method

In this section, simulations are carried out with a database of 100 hours randomly selected videos that account to around 1580000 scenes. Algorithm was coded in C++ and tested on an Intel core i7 architecture with a clock speed of 1.9 GHz. We considered the use of frame resolutions from 360p to 720pHD for our videos. Moreover, a non-GUI video editor such as AviSynth is used [11]. Different kind of editing and processing function (e.g. rotation, crop, tweak, etc) may be inserted on a video stream through AviSynth before the file is being read. Thereby different attacks of the video can be simulated to test the robustness of our proposed technique.

B. Framework setup

In order to do the matching, all video sequences were processed as described in Section 2, using AviSynth as implementation platform. For signature matching, similarity of properties is measured using hamming distance between two fingerprints. To accelerate query performance, cluster-based sequential search is used [12]. With the use of predetermined heuristics, each fingerprint is assigned to a cluster in the database denoted as $Q$. Query fingerprint with value closest to the cluster will then do comparison in it, hence reducing the number of matching by a factor of $\frac{1}{Q}$. However, if the match was not close to the predetermined threshold, a subsequent cluster will be examined, until an accurate match is found or all clusters have been examined.

To verify the robustness of the scheme, different attack types were mounted independently on the video to generate the queries. These alterations: contrast and brightness variation, rotation, text insertion, noise, will be used as part of our experiments and considered as video attacks. To ensure the accuracy of the results, the scenes are generated with detailed variations. Figures 3, 4, 5 show the variations of different attacks ranging from 0% to 30%.

The performance of our proposed technique is measured by its attack detection accuracy. Let us denote a true positive as $P$, true negative as $N$, and the total number of elements in each feature $F$ is denoted by $E$. The scores are accumulated for each feature and subsequently for each keyframe $K$, and the total score of the full set of keyframes added up to form the scene score. Each scene in the cluster will be compared against one another and the highest score is deemed as the closest match. Therefore, the scene score can be calculated as:

$$\text{Score}_{\text{Scene}} = \frac{1}{100KF} \sum_{j=0}^{K} \left( \sum_{i=0}^{F} \frac{P}{E} \times 100 \right) \times 100$$

For each fingerprint, the average matching percentage is computed; if the percentage is less than a predetermined
threshold, the comparison between reference video and query video returns a “no match”. Accuracy can be calculated as: \( \text{match} \rightarrow \text{accuracy} = (\sum_{z=0}^{T} \text{Score}_{\text{Scene}}), \) where \( T \) represents the total number of scenes within the query video.

C. Obtained results

<table>
<thead>
<tr>
<th>Type of attack</th>
<th>Description</th>
<th>Accuracy (in %)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contrast</td>
<td>10% increase</td>
<td>98.21%</td>
</tr>
<tr>
<td></td>
<td>20% increase</td>
<td>88.26%</td>
</tr>
<tr>
<td></td>
<td>30% increase</td>
<td>83.10%</td>
</tr>
<tr>
<td>Brightness</td>
<td>10% increase</td>
<td>84.79%</td>
</tr>
<tr>
<td></td>
<td>20% increase</td>
<td>82.02%</td>
</tr>
<tr>
<td></td>
<td>30% increase</td>
<td>80.35%</td>
</tr>
<tr>
<td>Noise</td>
<td>10% increase</td>
<td>75.99%</td>
</tr>
<tr>
<td></td>
<td>20% increase</td>
<td>73.63%</td>
</tr>
<tr>
<td></td>
<td>30% increase</td>
<td>72.88%</td>
</tr>
<tr>
<td>Rotation</td>
<td>180 degree rotation</td>
<td>72.68%</td>
</tr>
<tr>
<td>Text Insertion</td>
<td>-</td>
<td>81.74%</td>
</tr>
</tbody>
</table>

Fig. 6. The relationship between accuracy and variations.

As it can be seen in Table 1, our scheme is able to detect duplicate copies with high accuracy for strong variations. Among these attacks, variations in contrast and brightness produce good results as the edge feature are not affected much. However, for contrast variation, accuracy degradation escalates at a faster rate compared to brightness as result from change in energy of wavelet subbands. On the other hand, performance of in presence of noise fared poorly compared to the standard attacks, mainly due to grain injection into each frame which tweaks the spatial information, and results in many errors surfacing when preforming the binary matching.

In attacks that are less common such as scene rotation (Fig. 7), our method has demonstrated vulnerability. In Table 1, the percentage of accuracy is only 72%. Text insertion (Fig. 8) returns a relatively reliable result as shown in Table 1. It should be noted that the average time for a scene fingerprint generation is 2.48 sec and average scene matching is realized in 0.23 sec, therefore our scheme is well adapted for \( \approx \) real-time processing.

IV. CONCLUSION

In this work, we have proposed a fast and robust content-based video fingerprinting method. The proposed technique results in a reduced number of bits required for the fingerprint, making it lightweight. The generated fingerprints do not take up significant storage space and due to the small size, speed of matching improved. Besides, the cluster-based searching algorithm provides an efficient way to do matching, further reducing computation. Our method has demonstrated robustness against common attacks and manages to detect most of the attack copies. As future work, an evaluation and benchmark of the proposed technique with other fingerprinting algorithms will be done. Moreover, new fingerprinting techniques should be investigated in order to cope with attacks like background change.

REFERENCES

Regular Session 4
Radon Transform: Identification of Noise and Instrumental Artifacts

Mario D’Acunto1,2,3, Antonio Benassi2, Davide Moroni2; Ovidio Salvetti2,3
1Istituto di Struttura della Materia, Consiglio Nazionale delle Ricerche, ISM-CNR, via Fosso del Cavaliere 100, I-00133, Rome, Italy
2Istituto di Scienza e Tecnologie dell’Informazione, Consiglio Nazionale delle Ricerche, ISTI-CNR, via G. Moruzzi 1, I-56124, Pisa, Italy
3NanoICT-Lab, via Moruzzi, 1 I-56124, Pisa, Italy
Email: mario.dacunto@ism.cnr.it, antonio.benassi@isti.cnr.it, davide.moroni@isti.cnr.it, ovidio.salvetti@isti.cnr.it

ABSTRACT
Computed Tomography as well as Magnetic Resonance or Positron Electron Tomography are currently the most commonly used medical imaging modalities for the analysis of human body complex structures and organs, where diseases must be recognized and identified. The image reconstruction process used in these tomography techniques is usually based on the Radon Transform (RT). In this paper, an algorithm including correction of noise and of some instrumental artifacts directly from the RT sinograms is presented. The innovative contribution of our algorithm is based on the fact that it is not necessary a priori information on instrumental artifacts or noise sources. In addition, algorithm can be applied to any RT-based medical imaging technologies.

Index Terms: Radon Transform; Computed Tomography; Magnetic Resonance; Noise and instrumental artifacts.

1. INTRODUCTION
An important problem in image processing is to reconstruct a cross section of an object from several images of its projections. A projection is a shadowgram obtained by illuminating an object by penetrating radiation. Each horizontal line shown in this figure is a one-dimensional projection of a horizontal slice of the object. Each pixel of the projected image represents the total absorption of the penetrating radiation along its path from the source to the detector. By rotating the source-detector assembly around the object, projections for several different angles can be obtained. The goal of image reconstruction is to obtain an image of a cross section of the object from a collection of projections. Imaging systems that generate such slice views are called CT (computerized tomography) scanners. The RT is the underlying fundamental concept used for CT scanning, as well for a wide range of other disciplines, including radar imaging, geophysical imaging, nondestructive testing and medical imaging [1].

In the context of biomedical imaging, the reconstruction of complete 3D models of dendritic structures can be an important tool in many imaging applications. In general, a dendritic model can accurately quantify both 3D topology (connectivity) and morphology (radius, length, branching angles, etc) of the dendritic structure. Recently, several algorithms were produced where the main emphasis is put on topology or morphology, or both. Inferential methods have been largely used for the identification of vascular dendritic structures. A powerful way is to use a Bayesian method, such as maximum a posteriori estimation (MAP). Nevertheless, a principal difficulty with Bayesian techniques is a computational one. In fact, such methods require the use of Markov chain Monte Carlo (MCMC) algorithms, which may run for hundreds of thousands of iterations to yield reliable results. A partial solution is given by the usage of methods that combine local likelihood maximization with global structure determination using a general model of structures within a collection of specific samples.

In this paper, we introduce an algorithm that aims at correcting the effects of instrumental aberrations and some noise sources in the RT of 2D-3D phantom structures recorded in conditions of low signal-to-noise ratio (SNR). The algorithm is built to be applied to any tomography-based technique involving RT imaging construction. As a consequence, it is not necessary to have an accurate a priori knowledge of instrumental artifacts or noise sources. The results should improve the resolution of the medical images, making possible to quantify spatial sources observed by indirect measurements as, for example, in the functional Magnetic Resonance (fMRI) applied to brain study, where the measured signal is due to a changes of hemoglobin, a paramagnetic molecule, due to increased brain activity as a response of an external stimulus.

2. RADON TRANSFORM: THEORETICAL BACKGROUND
The Radon transform (RT) of a two dimensional function \( f(x,y) \) is defined as follows. Let \( A \) be a 2D model in \( \mathbb{R}^2 \) and let \( f(x) \) be the bi-dimensional binary function of \( A \), defined as \( f(x)=1 \), when \( x \) lies within the 2D model’s surface, and \( f(x)=0 \) otherwise.

Let, also, \( \eta = (\cos \theta, \sin \theta) \) be a unit vector in \( \mathbb{R}^2 \) and \( \rho \) a real number. The 2D Radon transformation \( R_2(\eta,\rho) \) of \( f(x) \) is a function which associates to each pair \( (\eta,\rho) \) the integral of \( f(x) \) on the plane \( P(\eta,\rho) = \{ (x,y) | x \cdot \eta = \rho \} \), where the superscript \( t \) indicates vector transposition. This plane is normal to the direction \( \eta \) and a distance \( \rho \) to the origin \( [4] \).

\[
R_2(\eta,\rho) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f(x,y) \delta(x \cos \theta + y \sin \theta - \rho) \, dx \, dy = \int_{-\infty}^{+\infty} f(\rho \cos \theta - l \sin \theta, \rho \sin \theta + l \cos \theta) \, dl \]

(1)

where the \( \delta \)-function converts the two-dimensional integral to a line integral \( dl \) along the line \( x \cos \theta + y \sin \theta = \rho \), see figure 1.

![Figure 1. Schematic sketch of a RT. The arrows go from the photon source to the detectors crossing the sample (green region).](image)

The transformed function \( R_2(\eta,\rho) \) is often referred to as the sinogram of \( f(x,y) \), because a point \( \delta \)-function in \( f \) transforms to a sinusoidal line \( \delta \)-function in \( P \). The generalization to 3D RT of \( f(x,y,z) \) is straightforward defined as:

\[
R_3(\eta,\rho) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f(r) \delta(x \cos \theta + y \sin \theta + z \cos \phi - \rho) r^2 dr \, dx \, dy = \int_{-\infty}^{+\infty} f(\rho \cos \theta - l \sin \theta, \rho \sin \theta + l \cos \theta) \, dl \]

(2)

where the unit vector \( \eta \) using spherical coordinates is now written as: \( \eta = (\cos \phi \sin \theta, \sin \phi \sin \theta, \cos \theta) \) and \( r \) is vector defined in the usual 3D space. Limiting ourselves to bi-dimensional case, since \( P(\eta,\rho+\pi) = P(\eta,\rho) \), we will mostly use a single-valued parameterization assuming that \(-\infty < \eta < \infty, 0 \leq \rho \leq \pi \). In addition, we can consider a class of function \{\( f \)\} to which to apply the projection theorem prescribing that:

\[
P(\chi, \rho) = f(\chi \cos \rho, \chi \sin \rho)
\]

(3)

where

\[
P(\chi, \rho) = \int P(\eta,\rho) e^{ij \eta} \, d\eta
\]

(4.a)

and

\[
f(\chi \cos \rho, \chi \sin \rho) = \int \int P(\rho \cos \theta - l \sin \theta, \rho \sin \theta + l \cos \theta) e^{ij \eta} \, d^2r
\]

(4.b)

The projection theorem (3) implies that in the class \{\( f \)\} we can use the inversion formula:

\[
f(r) = \frac{1}{4\pi} \int_0^{2\pi} \int_0^\infty |\chi| e^{-iz(\|\chi\|)} d\chi \int e^{ij \eta} P(\eta,\rho) d\eta \]

(5)

Eq. 5 describes the reconstruction algorithm of a mathematical model where \{\( f \)\} is the class of analyzed objects, and \( P(\eta,\rho) \) is the array of experimental data.

Now let us discuss the physical assumptions that form the basis for the model described above. Let us have a point radiation source \( M \) and a point detector \( D \). The measurements of \( P(\eta,\rho) \) corresponds to the case when both the source and the detector lie on the line \( (\eta,\rho) \) with the radiation propagating along this line. The above model corresponds to an infinite number of measurements (all \( (\eta,\rho) \)) performed without error by a point detector. It is clear that these conditions are not satisfied in practice, since there are no error/free point detectors and many measurements should require an infinite time to become errors irrelevant. We can consider noise or instrumental induced artifacts assuming the quantities \[5\]:

\[
I_{ij} = \int \Theta(\rho_i - \rho) P(\eta_j, \rho) d\rho
\]

(6)

where \( \Theta(\rho) \) is defined as the Instrumental Function (IF), which at the moment we do not specify. Generally, the IF possesses the following property \( \Theta(\rho) \rightarrow \Theta(\rho) \). In addition, it is defined locally, i.e., \( \Theta(\rho)=0 \) for \( |\rho|>\rho_0 \), where \( \rho_0 \) is a threshold value. The noise can be considered taking the quantities:

\[
I_{ij} = I_{ij} + \Delta I_{ij}
\]

(7)

where \( \Delta I_{ij} \) are random variables. The relation (7) can be easily extended to a 3D case.

### 2.1. Inverse Problem Associated to RT

Let us consider the bi-dimensional case of reconstruction of a real image. The unavailable image that is being solved for, \( f(x,y) \) is an \( N \times N \) image that can be represented by an \( N^2 \times 1 \) vector, \( \Phi \), where \( \Phi=vec(f(x,y)) \) and
The \( f(x,y) \) is a column by column concatenation of \( f(x,y) \), where the total pixels in the image is represented by the number \( N \times N \). Similarly, the sinogram data \( R_\gamma(\eta, \rho) \) is represented by \( \Gamma = \text{vec}(R(\eta, \rho)) \), where

\[
R_\gamma(\eta, \rho) = \begin{pmatrix}
  f(\eta_1, \rho_1) & \cdots & f(\eta_M, \rho_1) \\
  \vdots & \ddots & \vdots \\
  f(\eta_1, \rho_L) & \cdots & f(\eta_M, \rho_L)
\end{pmatrix}
\]  

(9)

here the sinogram data is a \( M \times L \) image, where each element in the vector represents one pixel and the value associated with that pixel is the sum of \( f(x,y) \) along the line \( \rho = \cos \theta \), \( y \sin \theta \).

In absence of noise, the reconstruction problem involves solving the following inverse problem: \( \Gamma = H \Phi \), where \( H \) is the forward transformation operator. The problem can be solved by inverting the transformation \( \Phi = H^{-1} \Gamma \). Such inversion presents various difficulties. The forward transformation operator is often singular or nearly singular, hence, for a given set of projections data there may be many or even no solutions, and the inverse of \( H \) may not exist. Also in the case where \( H^{-1} \) exists, the matrix can be too large to compute. For the reconstruction algorithms, a Bayesian approach can be used. In the Bayesian approach, an important role is played by the prior knowledge of some physiological parameters, for example in the CT tomography, the attenuation coefficients for the various organs [6]. A reconstruction algorithm based on Bayesian approach is detailed in section 3.

3. DENOISING APPROACH

The tomography images may suffer from different kinds of noise and artifacts. The most significant source of noise occurs in the photon detection process. Artifacts can occur from the physics of the system, from partial volume effects or from patient artifacts. Image noise occurs with the statistical fluctuation of photon detection. Generally, this type of noise can be modeled in two main ways, as Poissonian noise or Gaussian noise. The general form of a Poisson distribution can be written as:

\[
P(y) = \frac{\lambda^y e^{-\lambda}}{y!}
\]

(10)

where \( P(y) \) is the probability of a random variable \( y \), having mean \( \lambda \). The variance of the distribution is proportional to the average number of photons since photons are not emitted or detected uniformly. Poisson distribution reveals important characteristics, the mean and the variance are identical, as a consequence, the signal-to-noise ratio (SNR) is proportional to the square root of the detector measurements counts producing a much better SNR.

A Gaussian distribution is given as follows:

\[
P(y) = \frac{e^{-\frac{(y-\mu)^2}{2\sigma^2}}}{\sqrt{2\pi\sigma}}
\]

(11)

where \( \sigma \) is the variance. In the case of a Gaussian distribution, the number of photon counts is not proportional to variance.

On the contrary, image artifacts have different sources. For example, beam hardening: when an electromagnetic radiation, like X-ray, passes through an object, the lower energy photons are absorbed much quicker than the higher energy photons, the beam is said to become harder. This artifact in CT system causes dark bands or streaks. Other artifacts can be caused by photon starvation. Photon starvation occurs with objects that are very dense, such as bone. In CT, during the photon starvation, most photons are absorbed and very little reach the detectors. This results in very noisy projections that produce horizontal streaks in the image.

Patient artifacts can be due to metallic objects in CT or PET or by patient motion. In functional MR, also fluctuations of blood fluxes can cause artifacts. In addition patient motion can cause mis-registration of attenuation values; these artifacts can occur as shading or slight streaking in the entire image.

Let \( S, X, Y, N \) be respectively the matrices corresponding to biomedical source, observed signal, output 2D image (the approach is valid for all the tomography techniques, DT, MR, PET or SPECT) and noise. Then, the RT and the approximation of its inversion set following relationships among matrices \( X=RT(S+N) \) and \( Y=IRT(X) \), where \( IRT \) denotes the inverse RT.

The noise can be correlated or not. Generally, \( Y=IRT(RT(S)+N)=S+IRT(N) \), and the noise for the image \( Y-S=IRT(N) \) is strongly correlated, this can be a problem for Poissonian noise. To develop a denoising algorithm independent by the noise type, we apply a filter \( H \) to be specified so that the de-noised image can be written as \( \tilde{Y}=IRT(H(RT(X))) \).

A Shepp-Logan Phantom (SLP) and a real X-ray CT vascular brain system were used to test the algorithms for removing the noise. The signal-to-ratio of the noisy sinogram, in relation to the original sinogram, denoted as \( \xi_{bno} \), can be calculated as follows:
\[ \xi(h, \rho) = \frac{1}{100} \sum_{\eta=0}^{N-1} \sum_{\rho=0}^{N-1} \frac{|P(\eta, \rho)|^2}{\sigma^2} \]

where \( P(\eta, \rho) \) is the original sinogram. The sinograms of the Shepp-Logan and the X-ray CT barian vascular system (tree-like dendritic structure) were generated using 64 projections at the angles varying from \( 0^\circ \) to \( 180^\circ \), see figures 2 and 3.

**Figure 2.** On the left, 256×256 Shepp-Logan Phantom. On the middle image, the filtered backprojection, with the variance of the noise as \( \sigma^2=0.001 \) and SNR=16.67. On the right, the filtered backprojection, with the variance of the noise as \( \sigma^2=0.05 \) and SNR=10.05.

**Figure 3.** On the left, the real x-ray CT vascular brain system. On the middle image, the filtered backprojection, with the variance of the noise as \( \sigma^2=0.001 \) and SNR=16.67. On the right, the filtered backprojection, with the variance of the noise as \( \sigma^2=0.05 \) and SNR=10.05.

Initial results show that the denoising approaches feel again the incidence of the noise and the SNR. It is evidenced that in the Shepp-Logan Phantom, Figure 2, some residual noise effects are still present. In the case of the x-ray CT image, Figure 3, between the middle- and the right-image no relevant difference is outlined. Denoising sinograms can be improved also by incorporating prior knowledge of the tissues being scanned, This is currently our next research direction for reconstruction algorithms.

4. IMAGE RECONSTRUCTION UNDER LOW SNR CONDITIONS

The global procedure for image reconstruction from stacks is displayed in Figure 4, the case study is an osteoblast cell imaged with confocal microscopy. These images are used instead of a 3D phantom. A sinogram of single stacks is made both for an as-recorded image, both for Poisson-like noise removed image. Then, the IRT procedure is applied to the sinograms. The results are displayed in Figure 4.

On the left, we have an optical noisy low SNR cell image from confocal microscopy. Its sinogram is simulated applying a RT function and then is denoised using the algorithm developed in Section 2. Finally, 3D image reconstruction is made using the MCMC approach as described in this section (central image). Special features corresponding to inner nucleoli are enhanced with the same approach removing some possible discontinuities. No instrumental artifacts were taken in consideration, because specific information on the standard performance of the instrument used was not available.

**Figure 4.** Example of 3D reconstruction of images from stacks. From a-to-f, we have an optical noisy low SNR cell image from confocal microscopy (a). Its sinogram is simulated applying a RT function and then is denoised using the algorithm developed in Section 2. (b). Special features corresponding to inner nucleoli are enhanced with the same approach removing some possible discontinuities (c). Finally, the nucleoli and surrounding region are extracted by the 3D structure (d-f).

5. CONCLUSIONS

Radon Transform is the basic tool for the image reconstruction in many medical diagnostic devices, such as CT, fMRI or SPECT, etc. In any of such applications, the capability to discriminate the noise is a main target to be reached. In this paper, we applied some algorithms for removing noise and instrumental artifacts from Phantoms and CT vascular images. In addition, a 3D image reconstruction is made after artifacts have been removed. The original contribution of this method is given by its generality. It can be applied to all tomographic methods using Radon Transform and it can be further improved for image-based diagnostic techniques.
REFERENCES


NON-STATIONARY MODELING FOR THE SEPARATION OF OVERLAPPED TEXTS IN DOCUMENTS

Anna Tonazzini, Pasquale Savino, Emanuele Salerno

Istituto di Scienza e Tecnologie dell’Informazione - CNR
Via G. Moruzzi, 1, I-56124 PISA, Italy
name.surname@isti.cnr.it

ABSTRACT

In this paper, we address the removal of severe back-to-front interferences in archival documents, when recto and verso images of the page are available. The problem is approached from a modeling point of view, considering the ideal images of the two separated texts as individual source patterns that overlap in the observed images through some parametric mixing operator. Earlier approaches were based on linear mixtures of the ideal reflectance maps, or of the ideal optical densities and absorbance maps, through unknown coefficients or blur kernels. Some approximations and/or partial user supervision were then adopted to jointly estimate the sources and the model parameters. Nevertheless, a feasible and reliable data model for this problem should at least be non-linear and space-variant, to cope with occlusions, ink saturation, and large variability of the mixing level. This is especially true for ancient documents affected by ink seeping (bleed-through). The search for such a model is still far from being concluded, or even impossible to pursue, due to the unavailability of information about the chemical and physical processes at the origin of the phenomenon. Hence, here, we propose the use of pixel-dependent parameters, within a model additive in the optical densities, to compensate not only for non-stationarity, but also for the lack of the precise knowledge of the non-linearity, and for modeling errors more in general.

Index Terms— Document restoration, non-stationary data model, back-to-front interferences

1. INTRODUCTION

Archival documents written or printed on both sides of the page are typically degraded by back-to-front interferences (or see-through, show-through/bleed-through). These are undesired patterns in the background, caused either by transparency or seeping of the ink of the text in the reverse side of the page. Such distortion can significantly disturb the readability of the document by the interested scholars, make difficult the analysis of its content, or prevent the successful application of OCR techniques.

The most effective approaches for see-through reduction are methods that exploit the information from scans or images of both sides (recto and verso). Whereas the earliest proposed solutions were based on thresholding [1], wavelet techniques for enhancing the foreground strokes and smearing the interferences [2], or segmentation-classification [3], more recently the interest on model-based techniques has increased noticeably. More precisely, the recto and verso appearances of the degraded document are modeled as two parametric superimpositions of the uncorrupted front and back side images. This has led to solution methodologies based on source separation techniques, either fully blind, where the unknown model parameters are estimated along with the source, ideal images, or partially supervised, where the parameters are estimated off-line from the data, possibly with some user intervention. When a linear mixing model is assumed, fast fully blind separation algorithms are available, such as those based on Independent Component Analysis (ICA) and data decorrelation [4]. Within this linear modeling approach, compensations for the apparent non-linearity and/or non-stationarity of the physical phenomenon have used regularization techniques [5][6], or penalized Non-negative Matrix Factorization (NMF) [7]. In [8], based on the physical model of modern scanners, a non-linear convolutional mixing model is derived and solved for both the show-through Point Spread Function (PSF) and the ideal images, by decoupling the two recto-verso equations and approximating the ideal recto and verso patterns with the observed ones. In [9], the non-linearity is assumed known, and a total variation stabilizer is adopted for regularizing the problem. A quadratic mixing model accounting for a blur kernel on the show-through patterns is assumed in [10], and solved through maximum likelihood, whereas, in [11], variational approaches, based on nonlinear diffusion and wavelet transforms, have been proposed to model and then remove bleed-through. In [12] [13], we extended the original model in [8] to make it suitable also for the more
invasive bleed-through distortion. A constrained Maximum Likelihood technique was then adopted for blindly estimating the source images and the see-through PSF, assumed space-invariant throughout the image, or the model parameters were estimated off-line. An attempt to overcome the stationarity limitation was finally performed in [14], within a regularization framework.

Although improved results can be obtained when non-linearities are used in place of linearity, a comprehensive, feasible model for text overlapping in ancient documents is still far from being available. The main difficulties in this respect are to design a suitable non-linearity to describe text superposition in the occlusion areas, and to properly manage the non-stationarity of the degradation. We argue here that the use of pixel-dependent parameters, even within an additive or whatever (reasonable) model, could compensate not only for non-stationarity, but also for the lack or the imprecise knowledge of this non-linearity, and for modeling errors more in general. Indeed, it could act as a way to express the non-linearity in tabular form rather than as an analytical function. In this paper, we experiment such an approach, and base it on a non-stationary linear model combining, in a space-variant manner, the ideal optical densities of the two individual texts. The challenge in this data modeling is to estimate the pixel-dependent coefficients of the linear mixture. We propose to estimate them from the data alone, based on simple and intuitive criteria.

The paper is organized as follows. In Section 2, we describe our data model and the scheme adopted to estimate the two ideal front and back texts when the model parameters are assumed known. In Section 3, we discuss in detail our strategy to estimate the model parameters from the data. Section 4 analyzes some real experimental results. Finally, Section 5 concludes the paper, by discussing some ideas that could help to overcome the limitations of the method.

2. DATA MODEL AND IMAGE ESTIMATION

Assuming pre-registered recto and verso images, the data model we consider is the following:

\[
D^{\text{obs}}_r(t) = D_r(t) + q_r(t) [h_r(t) \otimes D_v(t)] \\
D^{\text{obs}}_v(t) = D_v(t) + q_v(t) [h_v(t) \otimes D_r(t)] \\
t = 1, 2, \ldots, T
\]  

with

\[
D(t) = -\log \left( \frac{s(t)}{R} \right)
\]

where \(D^{\text{obs}}_r(t)\) and \(D^{\text{obs}}_v(t)\) are the observed optical densities, and \(D_r(t)\) and \(D_v(t)\) are the ideal optical densities, of the front and back side, respectively, at pixel \(t\). Each density is related to the corresponding (ideal or observed, recto or verso) reflectance \(s\) through relation (2), involving a suitable constant \(R\) that represents, in the two instances \(R_r\) and \(R_v\), the mean reflectance values of the background in the recto and verso side, respectively. The model includes also two Point Spread Functions (PSF), \(h_r\) and \(h_v\), describing the smearing of ink that penetrates or shines through the paper, thus allowing a pattern in a side to match the corresponding one in the opposite side. These PSFs are stationary throughout the image but characterized by different gains \(q_r(t)\) and \(q_v(t)\), which, while representing our way to account for the non-stationary and non-linearity of the actual data model, have also the physical meaning of interference level from the front to the back and from the back to the front, respectively, at each pixel. With this notation, \(h_r\) and \(h_v\) are intended to be of unitary sum.

We propose to solve the system in eqs. (1) for \(D_r\) and \(D_v\), once \(q_r(t)\) and \(q_v(t)\), \(\forall t\), and \(h_r\) and \(h_v\) have been estimated off-line once for all, through a few iterations of the following iterative scheme:

\[
D^{(n+1)}_r(t) = D^{\text{obs}}_r(t) - q_r(t) \left[ h_r(t) \otimes D^{(n)}_v(t) \right] \\
D^{(n+1)}_v(t) = D^{\text{obs}}_v(t) - q_v(t) \left[ h_v(t) \otimes D^{(n+1)}_r(t) \right]
\]

by setting \(D^{(0)}_r = D^{\text{obs}}_r\), and constraining the density to be positive at each cycle, in order to make the reconstructed recto and verso images to have background values no greater than \(R_r\) and \(R_v\), respectively.

3. MODEL PARAMETER ESTIMATION

To estimate off-line the model parameters a minimum intervention of the user is required. \(R_r\) and \(R_v\) can be estimated by averaging the pixel intensities within two pure background areas. To estimate \(h_r\) and \(h_v\), two small areas need to be isolated, one in the recto and the other in the verso, where pure see-through is present (\(D_r = 0\) and \(D_v = 0\), respectively), and where the interference levels are about constant. Let us consider the recto side area. We have:

\[
D^{\text{obs}}_r = q_r \cdot h_v \otimes D^{\text{obs}}_v
\]

from which, by constrained least squares, \(q_r \cdot h_v\) can be estimated and then normalized to obtain just \(h_v\) [13]. A dual procedure can be used to obtain \(h_r\). Note that, for the show-through case, usually caused by the scanning process, there would be no reason to assume an anisotropic diffusion of light in the document support, so that we could assume \(h_r = h_v\) and thus compute it in a single side only. However, there could be a small residual displacement, \(t_0 = (i_0, j_0)\), between \(D_r\) and \(D_v\), due to some registration error. In this case, if \(h(t)\) is the see-through PSF for perfect registration, then we have \(h_r(t) = h(i + i_0, j + j_0)\), and \(h_v(t) = h(i - i_0, j - j_0)\). This relaxes the requirement for perfectly registered recto-verso pairs, since a small translational misalignment between recto and verso can be compensated, and allows us to still estimate the PSF on one side only, and derive the PSF on the other side by just translating it.
As per the estimation of the space-variant maps \( q_r(t) \) and \( q_v(t) \) at each pixel \( t \), we adopted the following criterion:

\[
q_r(t) = \frac{D^{obs}(t)}{h_r(t) \otimes D^{est}(t) + \epsilon}
\]

\[
q_v(t) = \frac{D^{obs}(t)}{h_v(t) \otimes D^{est}(t) + \epsilon}
\]

(5)

where \( \epsilon \) is a small positive number to avoid indeterminacies or infinity.

It is apparent that, in consideration of eqs. (1), eqs. (5) would make sense only when the ideal density is zero, that is in the pixels corresponding to the background in both sides, and to the pixels of pure see-through in one side and foreground text in the opposite side. However, with the aid of the few, intuitive precautions described below, we can use them everywhere.

Ideally, in the background pixels, we would like eqs. (5) to give null interference levels, so that the reconstructed images retain there the same values of the observed images. Nevertheless, due to small fluctuations around the mean background values, and to the fact that the density is always close to 0 in those areas, even high values of \( q_r(t) \) and \( q_v(t) \) can be obtained. However, in general, this is not a problem when restoration is performed (see eqs. (3)), since possible negative values of the estimated density are set to zero.

For the pixels of foreground text in one side and see-through in the opposite side, the equations would return a wrong high interference level that must be ignored. Indeed, this corresponds to interpreting the foreground text as an enhanced rather than attenuated see-through pattern coming from the actual see-through pattern in the opposite side. However, avoiding this is trivial. Indeed, it suffices, at each \( t \), to maintain the smallest between the two interference levels, and set the other to zero.

The last possible situation occurs in the occlusion pixels, i.e. when the densities in the two sides are both high since the two texts overlap. When the see-through darkens the foreground text, according to the data model of eqs. (1), a correct estimation of the interference levels should account for the ideal density, which is however unknown. Assuming instead the ideal density to be zero, as done in eqs. (5), would unavoidably produce holes in correspondence of the occlusion areas. However, when the degradation, although non-stationary, is not extreme, in the sense that the foreground text saturates the see-through, and provided that the two inks in the two sides reflect similarly under the same wavelength, we can expect that the densities of the occlusion areas in the two sides are close to each other. Since only the pixels in the occluding areas have this property, they can be located, and the corresponding interference levels can be set to zero therein. In the most obvious way, this can be accomplished with the aid of manually selected thresholds. However, in many cases, we have found that a simple procedure allows to automatically discriminate the occlusion pixels. In short, if the graylevels of the two foreground texts are similar, we compute the absolute difference between the reflectance maps of the recto and verso observations, and then apply an Otsu binarization [15] on this difference map. This will return a binary map where the zero pixels corresponds either to the occlusion pixels or to the background pixels, where both the two interference levels will be set to zero.

The algorithm in its whole is very fast and, in all the experiments, we obtained stabilization of the solutions in less than 10 iterations of the iterative scheme of eqs. (3).

4. DISCUSSION OF THE EXPERIMENTAL RESULTS

We experimented the method proposed above on the images of an online database of high resolution grayscale images of ancient documents affected by bleed-through [16]. This database comprises 25 registered recto-verso sample grayscale image pairs, taken from larger high resolution manuscript images, with varied degrees of bleed-through. In addition, for each image a binary ground-truth mask of the foreground text is provided. Although these ground truth images are synthetic, i.e. manually created, they can be useful for a quantitative analysis of the results.

Figures 1(a) and 1(d) show a manuscript belonging to the Allan and Maria Myers Academic Centre, University of Melbourne. The results of using the stationary non-linear model of [13] are shown in Figures 1(b) and 1(e), whereas Figures 1(c) and 1(f) show the much better results obtained by using the non-stationary, linear model proposed in this paper. From a visual point of view, the higher quality of the reconstructions in this case can be appreciated especially in the recto side. Indeed, in Figure 1(b) it is well visible the lower ink density in the occlusion areas, which does not affect the homologous image of Figure 1(c). This is confirmed by a quantitative analysis, performed by applying an automatic Otsu binarization to the reconstructions, and then counting the number of pixels that are different from their homologous in the corresponding ground-truth images. For the recto side we obtained a percentage of wrong pixels of 2.66 in the stationary case and of 1.93 in the non-stationary case, whereas for the verso side the percentage of wrong pixels was 1.82 in the stationary case and 1.63 in the non-stationary case. Of course, these measures of quality can only be considered as indicative, in that, as already said, the ground-truth images were built manually, so that they are not fully trustable. On the other hand, more sophisticated thesholding techniques could also give better binary images.

Finally, Figure 2 shows the maps of the estimated interference levels \( q_r \) and \( q_v \), and of the estimated occlusion pixels. Darker pixels in the interference level maps indicate higher values of \( q_r \) and \( q_v \). In the third map, black pixels represent occlusion areas.
Fig. 1. Application of the proposed method to a real recto-verso pair: (a) original degraded recto; (d) original degraded verso; (b) recto restored with the method in [13] (stationary non-linear model); (e) verso restored with the method in [13] (stationary non-linear model); (c) recto restored with the proposed method (linear non-stationary model); (f) verso restored with the proposed method (linear non-stationary model). Original images (a) and (d): reproduction by courtesy of The Allan and Maria Myers Academic Centre, University of Melbourne, digitized by Irish Script On Screen (www.isos.dias.ie).

Fig. 2. Maps of the estimated interference levels $q_r$ (a) and $q_v$ (b), and of the estimated occlusion pixels (c), for the images in Figures 1(a) and 1(d). Darker pixels in the interference level maps indicate higher values; black pixels in the third map represent occlusion areas.
5. CONCLUSIONS

We proposed linear, non-stationary mixtures of the optical densities to describe text overlapping in recto-verso images of archival documents, affected by show-through or bleed-through. The two fundamental features of this data model are non-stationarity, which allow to describe both penetration and transparency of the ink through the medium, as well as the variability of the degradation strength, definite acting as a sort of surrogate of a pixel-dependent non-linearity, and the inclusion of a PSF, which allows a pattern in a side to match the corresponding one in the opposite side. Based on this model, we derived a simple, partially supervised algorithm to separate the two texts. The algorithm acts in two steps: in the first one, the model parameters are estimated off-line, by a minimum and intuitive intervention of the user; in the second step, a few iterations of a classical constrained maximum likelihood technique allows to recover the restored images. The experimental results on real recto-verso printed documents and manuscripts show that the parameters involved in the data model can be estimated with satisfactory accuracy, so as to obtain clean images of the two document sides, where only the foreground text is correctly present.

Currently, we are working at the extension of the method to the multichannel, e.g. RGB, case, by restoring the various pairs of recto-verso channels independently. Another issue regards the present limitation due to the assumption of ink saturation in the occlusions. We plan to relax this constraint, which, although valid in the majority of the documents we examined, could be violated when the strokes are not fully black.

6. REFERENCES


SEGMENTATION AND RECOGNITION TECHNIQUES FOR HANDWRITTEN DEVANAGARI SCRIPT

Deepti Khanduja and Neeta Nain

MNIT Jaipur

ABSTRACT
Segmentation and classification of handwritten text is a major challenge in document recognition and analysis task. It becomes more complex and tedious task when handwritten text is written in skewed, tilted and intermixed lines. Incorrect segmentation propagates to error in character recognition. This paper presents a novel approach for line segmentation for Hindi handwritten text based on depth first search of connected components identified using clustering mechanism on the pixels of the image. Line segmentation is followed by the steps of slant and skew normalization and character segmentation is performed using projection profile of connected components thresholded by the average width of characters computed adaptively. In latter steps feature extraction of segmented character can be effectively done by using a mathematical curve fitting method on the segmented and partitioned character images. Resultant feature vector is fed to the neural network for classification. The proposed techniques have been tested on several Devanagari text document images and the tested results prove an accuracy of above 90% for each of the approaches. The proposed methods are adaptive to writer variations and are thus independent of external parameters compared to existing techniques.

Index Terms— Connected Components, Projection Profile, HCR, Shirorekha, Depth First Search, Zoning, Curve fitting, Neural Network.

1. INTRODUCTION
Hindi is the language that incorporates a wide range of character as well as word formations of complex nature. These formations make typing of Hindi text a time consuming task and require essentially an expert. Whereas writing in Hindi is easy and consumes less time as compared to typing. Here comes the major need and importance of a Hindi specific character recognizer where quickly written text can be fed into a recognizer to convert it into a digital text. This saves ample amount of time as needed in typing the same, where a typical Hindi character entry requires three - four key presses. Secondly, in the field of character recognition, recognition of Hindi handwritten text is more difficult task than machine printed document recognition. This is due to the wide range of variety in the formations of words and characters in Hindi and other text characteristics like text size, font, alignment and orientations present in handwritten text. This two fold need make it an obvious choice to design an efficient method of segmentation and recognition of Hindi text.

Character recognition consists of three major stages: pre-processing, segmentation (line and character), and recognition. Line Segmentation is the most crucial and challenging step in HCR (Handwritten Character Recognition) process. An error in line segmentation leads to incorrect character segmentation which further propagates to recognition stage. The challenges that need to be addressed by segmentation include unequal skew angle among lines, irregular spaces between lines and words, and fluctuating lines.

A number of methods have been previously proposed for segmentation, which includes, Projection profile based approach [1], Hough transform based approach [2], shredding text lines [3] and smearing approach [4]. However most of these methods are dedicated for machine printed documents and straight lines. Techniques proposed for feature extraction includes gradient based [6], Zernike moment based [7] and neural network based feature extraction techniques [8]. The main purpose of this paper is to present a complete handwritten character segmentation and recognition process which overcomes all the above mentioned obstacles and results in properly segmented and recognized documents.

The paper is organized as follows: Section 2 describes the characteristics of Hindi language. Section 3 includes the proposed segmentation and normalization techniques used for segmenting Devanagari handwritten document. A novel feature extraction technique has been proposed in section 4. In section 5 the experimental setup and the test results of the approach are presented followed by conclusions in Section 6.

2. CHARACTERISTICS OF HINDI LANGUAGE
Hindi is the third most popular language in the world and is written in Devanagari script. It is written from left to right and does not distinguish between lower and upper case. De-
vanagari consists of 11 vowels and 33 consonants. These are called basic characters. Complexity in recognition of Hindi characters increases as the characters also have half forms known as conjuncts. Vowels may also be combined with consonants causing them to change their shape. Such vowels are called Modifiers. Consonants when combined with other consonants lead to formation of Compound characters. The details of Devanagari script are reported by Bansal and Sinha in literature[5].

A Devanagari word is formed when a combination of consonants, vowels, modifiers and compound characters are bind together by a horizontal line, known as Shirorekha. Each word consists of three zones: Upper zone, middle zone and lower zone as shown in Figure 1.

3. SEGMENTATION

Segmentation can be considered as a pipeline process typically consisting of binarization, line segmentation, skew and slant normalization, and finally character segmentation as ordered steps for isolating and separating out each character of the script such that it can be a frequent input to the recognizer whose overall efficiency is primarily dependent upon the correct input. Each pipelined step of the segmentation process proves to be a refinement of the original script.

3.1. Binarization

The first stage in the pipeline of segmentation is image binarization also called image thresholding. It is the conversion of gray scale scanned image to a 2-color code scheme (i.e. binary image) with 1 and 0 representing foreground object pixels and background pixels respectively. Here, binarization is performed using Otsu’s method [9]. This step reduces the amount of information to be handled. On this binarized image further steps of segmentation are applied.

3.2. Line Segmentation

A hand written line is the first major element to be dealt with to estimate the size or font of the script. A line consists of variety of characteristics ranging from its width, length, skewness from horizontal axis, extent of intermixing with each other to variability in spacing between lines. These all characteristics are essential to be considered and affect the result to greater extent while segmenting the lines. The proposed Line segmentation algorithm takes as input a binarized image and results in a vector containing labels that specify the sequence of segmented lines.

The algorithm for line segmentation can be summarized as follows:

2. Label pixels belonging to the same cluster components with same number to isolate it from other clusters. Each cluster is called a Connected Component (CC).
3. Find the Centroid of each CC by calculating the mean of all pixel positions in x and y direction, belonging to the component using Equation 1.

\[
(x_i) = \frac{1}{N} \sum x_i \quad (y_i) = \frac{1}{N} \sum y_i
\]  

where, \(x_i, y_i \in CC_i\) and N is the number of pixels in \(CC_i\).

4. Calculate the Bounding Box (BB) around each component as shown in Figure 2 and find extreme coordinates \(X_{min}, X_{max}, Y_{min}, Y_{max}\) of the corresponding BB.

5. Perform correction of the overlapping BB as shown in Figure 3 (a), using Equation 2.

\[
\text{Merge } CC_1 \text{ and } CC_2 \text{ if } \left\{ \begin{array}{l}
CC_1 \text{ is encapsulated in } CC_2, \text{vice-versa} \\
|Y_{1max} - Y_{2min}| < 5 \\
|Y_{1min} - Y_{2max}| < 5
\end{array} \right.
\]

where 5 is the heuristically determined modulus threshold for merging two \(CC_i\). The correction of overlapped BBS of Figure 3 (a) is shown in Figure 3 (b).
6. Recalculate the BBs as the correction causes changes in centroid and co-ordinates of BBs.

7. Next compute a distance matrix \(D(i,j)\) to determine the normalized distances between the CCs as shown in Equation 3.

\[
D(i,j) = \begin{cases} 
|y_i - y_j| & \text{if } BB_i \text{ and } BB_j \text{ are horizontally parallel.} \\
|y_i - y_j| & \text{if } BB_i \text{ and } BB_j \text{ are vertically parallel.} \\
\sqrt{x_i - x_j^2 + y_i - y_j^2} & \text{otherwise.} 
\end{cases}
\]

8. Calculate threshold \(D_{min}\), i.e. minimum distance between two vertically parallel bounding boxes. \(D_{min}\) is required to identify components belonging to same line.

9. Create a vector containing labels of CCs, sorted according to the distance of their centroids from the origin.

10. Apply Depth First Search (DFS) as explained in Equation 4 to the sorted vector components resulting in final sequential vector with labels of CCs of segmented lines.

Let, \(i\) be the initial CC, \(j\) be the CC with minimum value of \(D(i,j)\), \(i'\) is the CC preceding \(i\) in sorted vector, \(c_{yk}\) is the y co-ordinate of centroid for \(CC_k\).

\[
D = D'' - D', \quad \text{where } D'' = cy_k - cy_{j'}, \quad D' = cy_i - cy_{j}, \quad \text{and} \\
if(D < D_{min}) \text{ then add } j \text{ to sequential vector containing } i. 
\]

Figure 4 shows an example handwritten Hindi text document having lines with different skews within a document, both positive and negative. The labels generated after application of the above DFS approach using Equation 4 are depicted as numbers on the words. The vector specifying the correct sequencing of the words of the document in Figure 4 is shown in Figure 5.

3.3. Skew and Slant Normalization

In handwritten text, skewness could be defined as the deviation of text from horizontal axis. The normalization of skewed text is one of the important step in the process of segmentation. The algorithm presented here is the modification of the approach presented in [12] and provides good results for images skewed in the range of -90° to +90° (excluding both).

The skew angle is calculated using the following steps.

1. Split the word in two equal parts.

2. Calculate centroid of each part i.e. \((x_1, y_1)\) and \((x_2, y_2)\) using its pixel co-ordinates.

3. Compute the difference, say \(D\), between vertical co-ordinates of the centroids of both parts using Equation 5.

\[
D = (y_2 - y_1) 
\]

4. Rotate the word by +1°, -1° and compute the difference for rotated texts as +\(D_1\) and -\(D_1\), respectively.

5. If +\(D_1 < D\) then (-ve skew in text) rotate original text by +1°.

Else If -\(D_1 < D\) then (+ve skew in text) rotate original text by -1°.

Repeat steps 1-5 until \(D == 0\) (skew corrected).

Figure 6(a) shows a sample text image skewed at an angle of +60°. This image has a +ve skew, which would be corrected by -ve rotations by 60°. Figure 6 (b) - (f) shows the intermediate results also, till the final skew correction as in Figure 6(g). Note that this approach does not require any explicit baseline detection [1] [5]. The skew angle and skew correction is adaptively determined without any external input. This method exploits orthographic projections for skew
Slant is the deviation of characters from vertical direction occurring due to writing styles. The method used to normalize this deviation is referred as Slant Angle Correction. The approach used here to deslant text is Vertical Projection Profile [11]. The basic idea behind this approach is that a straight written word will have high and distinct peaks as compared to a slanted word.

The algorithm proceeds by repeating steps (i) and (ii) at various shear angles ranging from $-45^\circ$ to $+45^\circ$, as it is the most common range of slant angles found in handwritten text.

(i) The first step is to obtain a vertical histogram of the text.

(ii) Next take the average of top three peaks of the histogram.

The text is transformed by the angle resulting in maximum value of peak average. Highest three peaks have been chosen for averaging by experimenting on 120 different words written by 50 writers, with the observation that a Hindi word has at least one consonant leading to three peaks. A value $> 3$ may generate false positives because, then the valleys may be misinterpreted as peaks in a single consonant. An example of the proposed slant normalization on a word, along with its vertical projection profile and calculation of peak average is shown in Figure 7(a), the slant corrected word id shown in Figure 7(b).

### 3.4. Character Segmentation

The $CC$ resulted from line segmentation algorithm is fed to character segmentation step after processing through the stages of skew and slant normalization. The character segmentation method proposed in this paper is based on Vertical Projection Profile as presented in the following steps.

1. Compute the skeleton $\text{Skel}_i$ for the $i^{th}$ input connected component $CC_i$.

2. Calculate vertical projection ($VP_i$) of $\text{Skel}_i$.

3. Segment $CC_i$ at positions where value of pixel in $VP_i$ is (0/1) and store the boundary pixels of segment $CC_i$ in vectors $\text{CharStart}_i$ and $\text{CharEnd}_i$.

4. Calculate distance between successive segments using Equation 6 and store the result in vector $\text{Diff}_i$.

$$\text{Diff}_i = \text{CharStart}_{i+1} - \text{CharEnd}_i \quad (6)$$

5. Compute the average width ($\text{Avg}$) of character in $CC_i$ by finding the mean of vector $\text{Diff}_i$ using Equation 7.

$$\text{Avg} = \frac{1}{N} \sum \text{Diff}_i \quad (7)$$

where, $\text{Diff}_i \in \text{Diff}$ and $N$ is the size of $\text{Diff}$.

6. Repeat for each element of $\text{Diff}_i$.

- If $\text{Diff}_i < \text{Avg}$, then merge segments $i$ and $i + 1$ (same character) set $\text{Diff}_i = 0$.

- If $\text{Diff}_i > 0$ then print pixels from $\text{CharStart}_i$ to $\text{CharEnd}_i$ (character is segmented).
Figure 8 illustrates an example for character segmentation using the above average width threshold for character segmentation.

4. FEATURE EXTRACTION AND CLASSIFICATION

Feature extraction is the method of converting highly redundant, variable and diverse data to a small size, robust, abstract and complete set of data that conveys all features of the original data. Generally data with greater size increases complexity and becomes time consuming for applications to process. Hence mapping to a corresponding small set of data makes task easier.

In character recognition task extracting those features that are essential to differentiate among characters is a tedious task. A simple method that generates most appropriate and complete set of features is always needed. Here we present a simple mathematical approach based on a combination of zoning and polynomial curve fitting approaches. Input of the algorithm is a binary segmented character and output is a feature vector consisting of 100 features to be fed to the neural network for further classification.

1. Resize input image \( Img \) to dimensions 100x100 and partition \( Img \) in 5x5 block using equal partitioning, thus resulting in 25 blocks each of size 20x20. Figure 9 renders the zoning approach, where, original image and the corresponding zoned image is shown in Figure 9(a) and (b) respectively.

2. Now for each part of \( Img \) find the character skeleton using averaging of boundary positions of drawn character.

3. Using each part separately find the number of dark pixels for each partition of the character. It constitutes a vector of 25 entries for complete image.

4. Using curve fitting mechanism over the skeleton of character generate a degree-2 polynomial using equation 8 with \( k = 2 \).

\[
f(x) = a_0 + a_1 x + a_2 x^2 + a_3 x^3 \ldots a_n x^n
\]

5. Polynomial generated in previous step gives \( 3 \) values of corresponding coefficients \( (a_0, a_1, \text{ and } a_2) \) constituting 75 entries \((25x3)\) in the previous vector.

6. Hence a complete feature vector consists of 100 entries that give a wide range of features to differentiate among characters.

7. Finally this feature vector can be fed as input to the neutral network for training and for testing as well.

The feature vector extracted is fed to the classifier for training and testing. The classifier is the decision making part of the recognition system based on the feature vector extracted from the previous stage. The classifier used here is a feed forward back propagation neural network having one hidden layer of 15 neurons obtained heuristically. The hidden layers use log sigmoid activation function, and the output layer uses a purelin transfer function. The number of input neurons is 100 as determined by the length of the feature vector. The total numbers of characters in the dataset 49 determines the number of neurons in the output layer. The network training parameters are:

- Input nodes: 100
- Hidden nodes: 15
- Output nodes: 49
- Training algorithm: trainrp(Resilient Backpropagation)
- Perform function: Mean Square Error
- Training goal achieved: 0.001
- Training epochs : 5000

An example of polynomial curve fitting is illustrated in Figure 10(a)-(l). Each representing the data points and corresponding curve for non-empty blocks of the image zoned into 5x5 blocks as shown in Figure 9(b).
5. EXPERIMENTAL RESULTS

The proposed line segmentation technique is tested on IAM database [13] containing 100 handwritten text documents having 7 lines each, thus on 700 text lines. Since there is no open source database for handwritten Hindi text, we generated a sample of 186 lines generated by scanned handwritten samples from 10 users. The complete test set configuration is enumerated in Table 1.

Line segmentation algorithm provides nearly 100% accuracy on documents ranging from printed lines to skewed and curved lines. The algorithm for skew and slant normalization have been tested on 120 word images and detects correct angle of skew and slant in 95% and 97% of images efficiently. The normalized images are tested for character segmentation which provides an accuracy of about 92%. The results for line segmentation and character segmentation using above approaches have been summarized in Tables I and II respectively.

The approach used for character recognition has been tested on an open source database set “Devnagari numeral and character database for offline handwritten character recognition” [14] with an overall accuracy of 96%. The database contains 20305 isolated character samples of 49 different character sets, taken from 750 different users. The regression curve between target vector and detected character output using neural network is shown in Figure 11, depicts the accuracy of feature extraction and character recognition approach.

6. CONCLUSIONS

In this paper we have used a simple and adaptive method of segmentation of Hindi handwritten text by the application of CCs and DFS for line segmentation. Skew and slant correction is performed using simple distance measures of the centroids of CCs, a novel orthographic projection for skew correction, and projection profile based method for slant correction. Character segmentation is based on calculating the average width of characters through vertical projection. The average calculation is adaptive catering to different user character sizes. Feature extraction is performed by zoning the character image followed by the application of 2-degree polynomial curve fitting techniques resulting in feature vector of 100 features. The feature vector is then fed to the classifier for training and testing. The aim of the paper is to develop an efficient algorithm for segmentation and recognition of Hindi handwritten text. Our results generated by the experiments on the test bed images conform to the greater level of efficiency of the methods. The success of the meth-
ods is impressive, particularly in view of methods simplicity and adaptability to user variations. The character segmentation does not remove modifiers as the modifiers get attached to the characters by bounding box merging done in line segmentation itself. The proposed methods are independent of external parameters compared to existing techniques [2] [3], which requires baseline computations for skew correction and the modifiers also get dropped at character segmentation stage [1] [5] [10]. The algorithm for feature vector extraction proves a greater level of accuracy, and efficiency as compared to previous approaches [8] [6].

Acknowledgment

This paper has been sponsored by Department of Science and Technology (DST), Rajasthan

7. REFERENCES


