Available online at www.sciencedirect.com
ScienceDirect
Pattern Recognition 40 (2007) 1234–1244
PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY
www.elsevier.com/locate/pr

# Eigenmoments

Pew-Thian Yap, Raveendran Paramesran*

*Department of Electrical Engineering, University of Malaya, 50603 Kuala Lumpur, Malaysia*

Received 5 January 2006; accepted 5 July 2006

**Abstract**

Moments and functions of moments are powerful general tools in a vast number of fields, and particularly in the field of image signal processing. In this paper, we present a method for obtaining a set of orthogonal, noise-robust, transformation invariant and distribution sensitive moments, which we call *Eigenmoments* (EM). EM are obtained by performing eigen analysis in the moment space generated by geometric moments (GM). This is done by transforming the *moment space* into the *feature space* where the signal-to-noise ratio (SNR) is maximized. This is equivalent to solving a generalized eigenvalue problem related to a Rayleigh quotient which characterize the SNR. The generalized eigenvalue problem can be decomposed into two eigenvalue problems. In the first eigenvalue problem, the moment space is transformed into the *noise space* where the noise components are removed. In the second eigenvalue problem a second transformation is performed to find the most expressive components. Experiments are performed to gauge the performance of EM and comparisons are made with some well known feature descriptors such as GM, DCT, Legendre moments and Tchebichef moments. The results show that EM give significant improvements in terms of accuracy and noise robustness as predicted by the theoretical framework.


*Keywords:* Moments; Orthogonalization; Image representation; Invariants; Noise robust features; Rayleigh quotient; Generalized eigenvalue problem

## 1. Introduction

Moments and functions of moments are invaluable tools in the statistical literature for the property measurements of a distribution, and they are thus named statistical descriptors. In the field of image analysis, their use as image descriptors was pioneered by Hu [1] when he applied the 2D geometric moments (GM) for the purpose of characterizing the visual patterns in images. For different images, the respective sets of moments are unique and this makes them particularly useful for the task of pattern recognition. This is further added by the advantage of being able to construct moment invariants which are insensitive to rotation, scaling and translation (RST). Hence, GM are effective descriptors for images under different perspectives.

Moments are characterized by their moment kernels. The moment kernels are used to transform the image in a scalar product manner into a multidimensional moment space consisting of moments as points in the space. It is a well known fact that the kernels of GM, i.e. the monomials $\{x^p y^q\}$, are not orthogonal. This impairs the performance of GM in a number of ways. Firstly, non-orthogonal moment kernels do not decorrelate the image in the moment space (i.e. the moments are correlated). Secondly, correlation in the moment space implies information redundancy and hence lower information compaction, that is, the energy in the moment space is not compacted in a few terms but is instead spread out. Thirdly, non-orthogonal moment kernels spread noise and make attempts to remove the noisy components difficult. All these cause impairment on the performance of GM and hence means of rectifying the non-orthogonality problem but at the same time retaining the invariant properties of GM is desirable.

One possible remedy to this is by the orthogonalization of the moment kernels using Gram–Schmidt orthogonalization (GSO) [2]. In the case of monomials, this results in a set of polynomials. For uniform weight, $w(x) = 1$, in the interval $x \in [-1, 1]$, using GSO we can obtain the set of

* Corresponding author. Tel.: +603 79675253; fax: +603 76975316.
*E-mail addresses:* ptyap@time.net.my (P.-T. Yap), ravee@um.edu.my (R. Paramesran).

0031-3203/$30.00 © 2006 Pattern Recognition Society. Published by Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2006.07.003

Legendre polynomials [3]. Legendre moments (LM) were used by Teague [4] in addressing the shortcomings of GM. However, this is not optimal and does not take into account the statistical properties of the images in consideration. Moreover, orthogonalization using the Gram–Schmidt process is subjected to a few weakness: Firstly, the resulting set of orthogonalized functions is dependent on the order of application of GSO to the original non-orthogonal functions. Different order would result in different sets of orthogonalized functions. Secondly, the GSO process is numerically unstable [2] especially when a large number of functions need to be orthogonalized. Thirdly, since no consideration is given to the image distribution, the orthogonalization of the moment kernel does not guarantee that the image will be decorrelated in the moment space. These make the Gram–Schmidt process somewhat non-optimal.

In this paper, we present a method for obtaining a set of orthogonal, noise-robust, RST invariant and distribution sensitive moments which we call *Eigenmoments* (EM). The set of EM is obtained by performing orthogonalization, via eigen analysis, on the moments themselves (i.e. in the moment space) rather than on the moment kernels as discussed previously using GSO. We show that this can be done by maximizing the signal-to-noise ratio (SNR) in the *feature space* in the form of Rayleigh quotient by solving a generalized eigenvalue problem (GEP). The GEP can be decomposed into two eigenvalue problems. In the first eigenvalue problem, the moment space is transformed into the *noise space* where the noise components are removed. In the second eigenvalue problem a second transformation is performed to find the most expressive components. This approach has several advantages:

(1) Since the distribution of the moments in the moment space is dependent on the distribution of the images being transformed, eigen analysis in the moment space ensures the decorrelation of the final feature space.
(2) The ability of EM to take the image distribution in consideration makes it more versatile and adaptable to images of different genres.
(3) The moment kernels generated are orthogonal and this makes the analysis of the moment space much more easier. Transformation with orthogonal moment kernels into the moment space is equivalent to the projection of the image onto a number of orthogonal axes.
(4) Noisy components can be effectively removed as will be evidenced by the empirical results shown later. This makes EM robust in classification task especially when the images are corrupted with noise.
(5) Optimal information compaction can be obtained and hence only a small number of moments are needed to characterize the images.

As mentioned before, one of the motivations for using GM is that moment invariants which are insensitive to RST can be easily obtained using GM. We show that in formulating EM, these properties are still retained and EM give a boost to the performance making the already popular GM more useful and robust to noisy conditions.

Experiments are performed to gauge the performance of EMs. The results are compared to well known feature descriptors such as GM, discrete cosine transform (DCT), LM and the relatively recent Tchebichef moments (TM) [5]. The results are motivating and shows improvement margins of about 80% when compared to the GM. When compared to DCT, LM and TM improvement margins can reach up to more than 30%. Under RST transformations, the improvement is typically 30% over that of GM. These results validated the efficiency of EMs as predicted by the theoretical framework.

The rest of the paper is organized as follows: Section 2 gives the mathematical preliminaries important for later developments of the paper. Section 3 furnishes the details of the problem in consideration and shows how the construction of the set of EM is related to problems such as maximization of a Rayleigh quotient, GEP and simultaneous diagonalization. Section 4 shows how the set of EMs can be derived using the concepts previously discussed. Section 5 provides the empirical validation of the proposed theoretical framework. Section 6 concludes this paper.

## 2. Preliminaries

It would be appropriate to first give a brief summary of the mathematical tools which we are going to use to obtain the set of EM.

### 2.1. The Rayleigh quotient

The optimization that we are concerned with basically has the form:

$$\max_{\mathbf{w}} R(\mathbf{w}) = \max_{\mathbf{w}} \frac{\mathbf{w}^{\mathrm{T}}\mathbf{A}\mathbf{w}}{\mathbf{w}^{\mathrm{T}}\mathbf{B}\mathbf{w}}, \tag{1}$$

where $\mathbf{A}, \mathbf{B} \in \mathscr{R}^{n\times n}$ are both symmetric matrices, and $\mathbf{B}$ positive definite and hence invertible. This is an optimization of a Rayleigh quotient [6,7]. One can see that scaling $\mathbf{w}$ does not change the value of the object function and hence an additional scalar constraint $\mathbf{w}^{\mathrm{T}}\mathbf{B}\mathbf{w}=1$ can be imposed on $\mathbf{w}$ and no solution would be lost when the objective function is optimized. The optimization problem becomes a constrained optimization problem of the form:

$$\max_{\mathbf{w}} \mathbf{w}^{\mathrm{T}}\mathbf{A}\mathbf{w} \quad \text{subject to } \mathbf{w}^{\mathrm{T}}\mathbf{B}\mathbf{w}=1 \tag{2}$$

or by using the Lagrangian $\mathscr{L}(\mathbf{w})$:

$$\max_{\mathbf{w}} \mathscr{L}(\mathbf{w}) = \max_{\mathbf{w}}(\mathbf{w}^{\mathrm{T}}\mathbf{A}\mathbf{w} - \lambda\mathbf{w}^{\mathrm{T}}\mathbf{B}\mathbf{w}). \tag{3}$$

Equating the first derivative to zero leads to

$$\mathbf{A}\mathbf{w} = \lambda\mathbf{B}\mathbf{w}. \tag{4}$$

The optimal value reached by the object function is equal to the maximal eigenvalue, which in this case corresponds to the Lagrangian multiplier $\lambda$. This is the symmetric GEP and will be discussed next.

### 2.2. Generalized eigenvalue problem (GEP)

The GEP [2,6,7] has the form:

$$\mathbf{A}\mathbf{w} = \lambda\mathbf{B}\mathbf{w}. \tag{5}$$

For any pair $(\mathbf{w}, \lambda)$ that is a solution of Eq. (5), $\mathbf{w}$ is called a generalized eigenvector and $\lambda$ is called a generalized eigenvalue. This could be solved as an ordinary but non-symmetric eigenvalue problem by multiplying with $\mathbf{B}^{-1}$ on the left-hand side. We can also convert it to a symmetric eigenvalue problem by decomposing $\mathbf{B} = \mathbf{G}^{\mathrm{T}}\mathbf{G}$, where $\mathbf{G}$ is a non-singular matrix, and defining $\mathbf{v} = \mathbf{G}\mathbf{w}$:

$$\begin{aligned} &\mathbf{A}\mathbf{w} = \lambda\mathbf{G}^{\mathrm{T}}\mathbf{G}\mathbf{w} \\ &\quad\Rightarrow \mathbf{A}\mathbf{G}^{-1}\mathbf{G}\mathbf{w} = \lambda\mathbf{G}^{\mathrm{T}}\mathbf{G}\mathbf{w} \\ &\quad\Rightarrow \mathbf{A}\mathbf{G}^{-1}\mathbf{v} = \lambda\mathbf{G}^{\mathrm{T}}\mathbf{v} \end{aligned} \tag{6}$$

and this by left multiplication with $(\mathbf{G}^{-1})^{\mathrm{T}} = (\mathbf{G}^{\mathrm{T}})^{-1}$, we obtain

$$[(\mathbf{G}^{-1})^{\mathrm{T}}\mathbf{A}\mathbf{G}^{-1}]\mathbf{v} = \lambda\mathbf{v} \tag{7}$$

which is an ordinary symmetric eigenvalue problem. If we let $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ where $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_n$ and by defining $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_n]$ and $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_n]$, where $\mathbf{v}_i$ and $\mathbf{w}_i$ are the eigenvectors of Eq. (7) and Eq. (5), respectively, we can rewrite Eq. (5) in a convenient form as

$$\mathbf{A}\mathbf{W} = \mathbf{B}\mathbf{W}\boldsymbol{\Lambda}. \tag{8}$$

Since $(\mathbf{G}^{-1})^{\mathrm{T}}\mathbf{A}\mathbf{G}^{-1}$ is symmetric, the eigenvectors $\mathbf{v}_i$ are mutually orthogonal and can be chosen to be of unit norm, and thus

$$\mathbf{W}^{\mathrm{T}}\mathbf{B}\mathbf{W} = \mathbf{V}^{\mathrm{T}}\mathbf{V} = \mathbf{I}. \tag{9}$$

This means that the generalized eigenvectors $\mathbf{w}_i$ of a symmetric eigenvalue problem are orthogonal in the metric defined by $\mathbf{B}$. Using Eq. (8), we further have

$$\mathbf{W}^{\mathrm{T}}\mathbf{A}\mathbf{W} = \mathbf{W}^{\mathrm{T}}\mathbf{B}\mathbf{W}\boldsymbol{\Lambda} = \mathbf{I}\boldsymbol{\Lambda} = \boldsymbol{\Lambda}. \tag{10}$$

From Eqs. (9) and (10), the GEP and thus the optimization of the Rayleigh quotient can also be viewed as the *simultaneous diagonalization* [8] of matrices $\mathbf{A}$ and $\mathbf{B}$. We shall see later that these two equations are important for the formulation of EM. The problem reduces to ordinary symmetric eigenvalue problem when $\mathbf{B} = \mathbf{I}$. Note that $\mathbf{w}_1$ gives the maximum value for the Rayleigh quotient, i.e. $R(\mathbf{w}_1) = \lambda_1$. The other $\mathbf{w}_i$ maximizes the Rayleigh quotient subject to the condition [7]:

$$\max_{\mathbf{w}_i^{\mathrm{T}}\mathbf{B}\mathbf{w}_1 = \cdots = \mathbf{w}_i^{\mathrm{T}}\mathbf{B}\mathbf{w}_{i-1} = 0} R(\mathbf{w}_i) = \lambda_i. \tag{11}$$

The proof for this can be found in the Appendix. We will show in the next few sections how the maximization of the Rayleigh can be used to develop the set of EM.

## 3. Problem formulation

Assume that we have a signal vector $\mathbf{s} \in \mathbf{R}^n$ which is taken from a certain distribution having correlation $\mathbf{C} \in \mathscr{R}^{n\times n}$, i.e. $\mathbf{C} = E[\mathbf{s}\mathbf{s}^{\mathrm{T}}]$, where $E[\cdot]$ denotes the expectant value. Since the dimension of signal space, $n$, is often too large to be useful for practical application such as pattern classification, we need to transform the signal space into a space of lower dimension. We perform this by a two-step linear transformation:

$$\mathbf{q} = \mathbf{W}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{s}, \tag{12}$$

where $\mathbf{q} = [q_1, \ldots, q_n]^{\mathrm{T}} \in \mathscr{R}^k$ is the transformed signal, $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_m] \in \mathscr{R}^{n\times m}$ a fixed transformation matrix which transforms the signal into the *moment space*, and $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_k] \in \mathscr{R}^{m\times k}$ the transformation matrix which we are going to determine by maximizing the SNR of the feature space resided by $\mathbf{q}$. For the case of GM, $\mathbf{X}$ would be the monomials. If $m = k = n$, a full rank transformation would result, but more typically, we would have $m \leqslant n$ and $k \leqslant m$. This is especially true when the signal $\mathbf{s}$ is of high dimension, and possibly of infinite dimension, i.e. when $\mathbf{s}$ is continuous. We find $\mathbf{W}$ by maximizing the SNR of the feature space:

$$\mathrm{SNR}_{transform} = \frac{\mathbf{w}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{C}\mathbf{X}\mathbf{w}}{\mathbf{w}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{N}\mathbf{X}\mathbf{w}}, \tag{13}$$

where $\mathbf{N}$ is the correlation matrix of the noise signal. The problem can thus be formulated as

$$\{\mathbf{w}_1, \ldots, \mathbf{w}_k\} = \arg\max_{\mathbf{w}} \frac{\mathbf{w}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{C}\mathbf{X}\mathbf{w}}{\mathbf{w}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{N}\mathbf{X}\mathbf{w}} \tag{14}$$

subject to constraints:

$$\mathbf{w}_i^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{N}\mathbf{X}\mathbf{w}_j = \delta_{ij}, \tag{15}$$

where $\delta_{ij}$ is the Kronecker delta defined as

$$\delta_{ij} = \begin{cases} 1, & i = j, \\ 0, & i \neq j. \end{cases} \tag{16}$$

By letting $\mathbf{A} = \mathbf{X}^{\mathrm{T}}\mathbf{C}\mathbf{X}$ and $\mathbf{B} = \mathbf{X}^{\mathrm{T}}\mathbf{N}\mathbf{X}$, it can be observed that Eq. (14) is actually a Rayleigh quotient as defined in Eq. (1), i.e. Eq. (14) can be written as

$$\{\mathbf{w}_1, \ldots, \mathbf{w}_k\} = \arg\max_{\mathbf{w}} \frac{\mathbf{w}^{\mathrm{T}}\mathbf{A}\mathbf{w}}{\mathbf{w}^{\mathrm{T}}\mathbf{B}\mathbf{w}}, \quad \mathbf{w}_i^{\mathrm{T}}\mathbf{B}\mathbf{w}_j = \delta_{ij}. \tag{17}$$

Hence, one way of solving Eq. (17) is through solving the GEP by Eq. (7), and dimension reduction can be performed by simply choosing the first $k$ components $\mathbf{w}_i$, $i = 1, \ldots, k$, with the highest values for $R(\mathbf{w})$ out of the $m$ components, and discard the rest. This way, the generalized eigenvectors

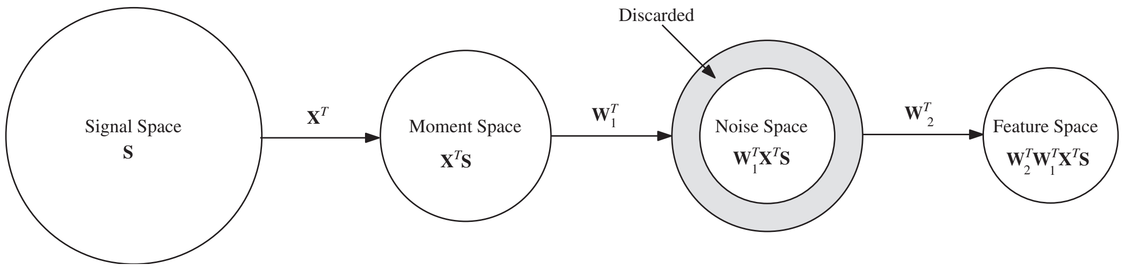


Fig. 1. Conceptual chart for the algorithm of generating the set of Eigenmoments.

can be interpreted geometrically as rotating and scaling the moment space, transforming it into a feature space with maximized SNR. Therefore, the first $k$ components are the components with the highest $k$ SNR values.

However, we choose to use the concept of simultaneous diagonalization (i.e. Eqs. (9) and (10)) instead of the GEP because the former gives us a better understanding and more control in constructing the set of EM. The details are outlined as follows:

(1) Let $\mathbf{A} = \mathbf{X}^{\mathrm{T}}\mathbf{C}\mathbf{X}$ and $\mathbf{B} = \mathbf{X}^{\mathrm{T}}\mathbf{N}\mathbf{X}$ as mentioned and write $\mathbf{W}$ in terms of two separate transformation matrices:

$$\mathbf{W} = \mathbf{W}_1\mathbf{W}_2. \tag{18}$$

$\mathbf{W}_1$ and $\mathbf{W}_2$ can be found by the following steps.

(2) We first diagonalize:

$$\mathbf{P}^{\mathrm{T}}\mathbf{B}\mathbf{P} = \mathbf{D}_{\mathrm{B}}, \tag{19}$$

where $\mathbf{D}_{\mathrm{B}}$ is a diagonal matrix sorted in *increasing* order. Note that since $\mathbf{B}$ is positive definite, thus $\mathbf{D}_{\mathrm{B}} > 0$. It is especially important to retain those eigenvalues which are close to 0, since this means the energy of the noise is close to 0 in this space. On the other hand, we can at this stage choose to discard the eigenvectors with large eigenvalues. This will make EM insensitive to the effect of noise in the image. Let $\widehat{\mathbf{P}}$ be the first $k$ columns of $\mathbf{P}$, now

$$\widehat{\mathbf{P}}^{\mathrm{T}}\mathbf{B}\widehat{\mathbf{P}} = \widehat{\mathbf{D}}_{\mathrm{B}}, \tag{20}$$

where $\widehat{\mathbf{D}}_{\mathrm{B}}$ is the $k \times k$ principal submatrix of $\mathbf{D}_{\mathrm{B}}$.

(3) Let $\mathbf{W}_1 = \widehat{\mathbf{P}}\widehat{\mathbf{D}}_{\mathrm{B}}^{-1/2}$, and hence:

$$\mathbf{W}_1^{\mathrm{T}}\mathbf{B}\mathbf{W}_1 = (\widehat{\mathbf{P}}\widehat{\mathbf{D}}_{\mathrm{B}}^{-1/2})^{\mathrm{T}}\mathbf{B}(\widehat{\mathbf{P}}\widehat{\mathbf{D}}_{\mathrm{B}}^{-1/2}) = \mathbf{I}. \tag{21}$$

Thus $\mathbf{W}_1$ whiten $\mathbf{B}$ and reduces the dimensionality from $m$ to $k$. For convenience, we call the transformed space resided by $\mathbf{q}' = \mathbf{W}_1^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{s}$ the *noise space*.

(4) Diagonalize $\mathbf{W}_1^{\mathrm{T}}\mathbf{A}\mathbf{W}_1$:

$$\mathbf{W}_2^{\mathrm{T}}\mathbf{W}_1^{\mathrm{T}}\mathbf{A}\mathbf{W}_1\mathbf{W}_2 = \mathbf{D}_{\mathrm{A}}, \tag{22}$$

where $\mathbf{W}_2^{\mathrm{T}}\mathbf{W}_2 = \mathbf{I}$. $\mathbf{D}_{\mathrm{A}}$ is the matrix with the eigenvalues of $\mathbf{W}_1^{\mathrm{T}}\mathbf{A}\mathbf{W}_1$ at its diagonal. We may retain all the eigenvalues and their eigenvectors since the most of the noise components are already discarded in the previous step and the current space is minimally affected by noise. Or we can discard the components with small eigenvalues if we further want to reduce the dimensionality.

(5) The transformation matrix is then given by (as noted before)

$$\mathbf{W} = \mathbf{W}_1\mathbf{W}_2. \tag{23}$$

$\mathbf{W}$ diagonalizes both the numerator and denominator of the SNR:

$$\mathbf{W}^{\mathrm{T}}\mathbf{A}\mathbf{W} = \mathbf{D}_{\mathrm{A}}, \quad \mathbf{W}^{\mathrm{T}}\mathbf{B}\mathbf{W} = \mathbf{I}. \tag{24}$$

(6) Hence, the transformation of the signal $\mathbf{s}$ to the feature space given by

$$\mathbf{q} = \mathbf{W}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{s} = \mathbf{W}_2^{\mathrm{T}}\mathbf{W}_1^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{s}. \tag{25}$$

A summary of the above algorithm is given in Fig. 1.

**Remark 3.1.** If $k = m$, i.e. no component is discarded, then the solution $\mathbf{W}$ is exactly the same as that obtain using Eq. (7).

**Remark 3.2.** If no component is intended to be discarded in Step 4, that is, if $\mathbf{W}_2$ has full rank, then Step 4 can be skipped if Euclidean distance is used as a distance measure in the feature space. This is because $\mathbf{W}_2$ is orthogonal, i.e. $\mathbf{W}_2^{\mathrm{T}}\mathbf{W}_2 = \mathbf{I}$ and preserves the Euclidean distance. We can then set $\mathbf{W} = \mathbf{W}_1$.

## 4. Eigenmoments (EM)

The above framework is general enough for different applications. But we shall show here one example of how to use the above framework to formulate a set of noise robust moments which we call EM. The derivation for 1D signal is first discussed and then followed by the 2D version. Following the discussion in the previous section, and extending them to the continuous case, if we let $\mathbf{X} = [1, x, x^2, \ldots, x^{m-1}]$, i.e. the monomials, after the transformation $\mathbf{X}^{\mathrm{T}}$ we obtain the GM, denoted by the vector $\mathbf{M}$, of signal $\mathbf{s} = [s(x)]$, i.e. $\mathbf{M} = \mathbf{X}^{\mathrm{T}}\mathbf{s}$. In practical situations, it is often difficult to estimate the correlation of the signal due to insufficient number

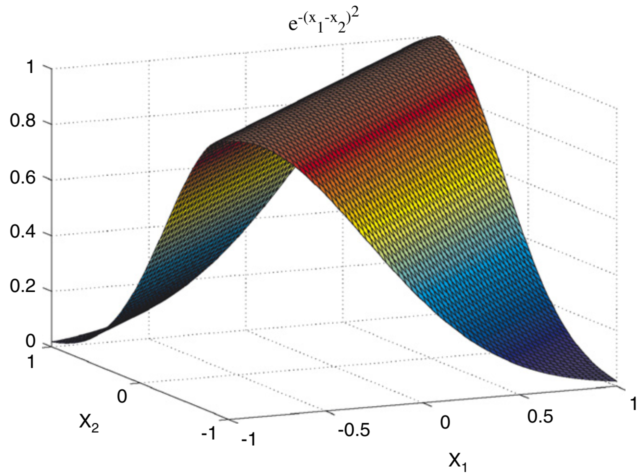


Fig. 2. The correlation of the signal for $c = 1$.

of samples. Hence, we use a parametric approach and let the correlation of the signal be

$$r(x_1, x_2) = r(0, 0)\mathrm{e}^{-c(x_1-x_2)^2}, \tag{26}$$

where $r(0, 0) = E[\mathrm{tr}(\mathbf{ss}^{\mathrm{T}})]$.[1] A plot of $r(x_1, x_2)$ is shown in Fig. 2. This model of correlation caters for general natural images and can be replaced by other models whenever appropriate. Since the result will not be affected by $r(0, 0)$, we will drop it for conciseness. We then have

$$\mathbf{A} = \mathbf{X}^{\mathrm{T}}\mathbf{C}\mathbf{X} = \int_{-1}^{1}\int_{-1}^{1}\left[x_1^j x_2^i \mathrm{e}^{-c(x_1-x_2)^2}\right]_{i,j=0}^{i,j=m-1} \mathrm{d}x_1\,\mathrm{d}x_2 \tag{27}$$

which can be determined fairly easily. Since the correlation of the noise can be modeled as $\sigma_n^2\delta(x_1, x_2)$, where $\sigma_n^2$ is the energy of the noise and $\delta(x_1, x_2)$ the delta function, dropping $\sigma_n^2$, we have

$$\begin{aligned}\mathbf{B} = \mathbf{X}^{\mathrm{T}}\mathbf{N}\mathbf{X} &= \int_{-1}^{1}\int_{-1}^{1}[x_1^j x_2^i \delta(x_1, x_2)]_{i,j=0}^{i,j=m-1}\,\mathrm{d}x_1\,\mathrm{d}x_2 \\ &= \int_{-1}^{-1}[x_1^{i+j}]_{i,j=0}^{i,j=m-1}\,\mathrm{d}x_1 = \mathbf{X}^{\mathrm{T}}\mathbf{X}.\end{aligned} \tag{28}$$

Using $\mathbf{A}$ and $\mathbf{B}$ and applying the algorithm discussed in the previous section we can find $\mathbf{W}$ and a set of transformed monomials $\mathbf{\Phi} = [\phi_1, \ldots, \phi_k] = \mathbf{XW}$ (hence polynomials) which forms the moment kernels of EM. The moment kernels of EM decorrelate the image correlation $\mathbf{C}$, i.e.,

$$\mathbf{\Phi}^{\mathrm{T}}\mathbf{C}\mathbf{\Phi} = (\mathbf{XW})^{\mathrm{T}}\mathbf{C}(\mathbf{XW}) = \mathbf{D}_{\mathrm{C}}, \tag{29}$$

and are orthogonal:

$$\begin{aligned}\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi} &= (\mathbf{XW})^{\mathrm{T}}\mathbf{XW} \\ &= \mathbf{W}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{XW} \\ &= \mathbf{W}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{NXW} \\ &= \mathbf{W}^{\mathrm{T}}\mathbf{BW} \\ &= \mathbf{I}\end{aligned} \tag{30}$$

since $\mathbf{W}^{\mathrm{T}}\mathbf{BW} = \mathbf{I}$ according to Eq. (9). As an example, if we take $c = 1$, the dimension of moment space as $m = 5$ and the dimension of feature space as $k = 4$, we have

$$\mathbf{W} = \begin{bmatrix} 0 & -0.9574 & 0 & 0.6972 \\ -1.7159 & 0 & -2.5359 & 0 \\ 0 & 3.1041 & 0 & -7.6955 \\ 0.8528 & 0 & 4.5987 & 0 \\ 0 & -0.8964 & 0 & 9.2363 \end{bmatrix} \tag{31}$$

and

$$\begin{aligned}\phi_1 &= -1.7159x + 0.8528x^3, \\ \phi_2 &= -0.9574 + 3.1041x^2 - 0.8964x^4, \\ \phi_3 &= -2.5359x + 4.5987x^3, \\ \phi_4 &= 0.6972 - 7.6955x^2 + 9.2363x^4.\end{aligned} \tag{32}$$

Hence, the signal $\mathbf{s}$ can be transformed using Eq. (25), i.e.,

$$\mathbf{q} = \mathbf{\Phi}^{\mathrm{T}}\mathbf{s} = \mathbf{W}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{s} = \mathbf{W}^{\mathrm{T}}\mathbf{M}, \tag{33}$$

where $\mathbf{M}$ is a column vector containing the GM of $\mathbf{s}$, and $\mathbf{q}$ is a column vector containing the set of EM.

### *4.1. EM for 2D signals*

For 2D image signals, the above results can be employed directly to the conventional GM to obtain the set of 2D EM. The definition of GM of order $(p + q)$ is [1]

$$m_{pq} = \int_{-1}^{1}\int_{-1}^{1} x^p y^q f(x, y)\,\mathrm{d}x\,\mathrm{d}y. \tag{34}$$

If we denote $\mathbf{M} = \{m_{j,i}\}_{i,j=0}^{i,j=m-1}$, then we can obtain the set of EM by

$$\mathbf{\Omega} = \mathbf{W}^{\mathrm{T}}\mathbf{M}\mathbf{W}, \tag{35}$$

where $\mathbf{\Omega} = \{\Omega_{j,i}\}_{i,j=0}^{i,j=k-1}$ is a matrix which contains the set of EM, and

$$\Omega_{j,i} = \sum_{r=0}^{m-1}\sum_{s=0}^{m-1} w_{r,j} w_{s,i} m_{r,s}. \tag{36}$$

### *4.2. EM invariants (EMI)*

To obtain a set of moment invariants, we simply need to replace $\mathbf{M}$ with the set of normalized GM, denoted as $\widehat{\mathbf{M}}$.

[1] We are aware of the existence of other correlation models such as $\lambda^{|x_1-x_2|}$ [9] and $\mathrm{e}^{-c|x_1-x_2|}$ [10], but our form makes it easier to obtain the integral.

Table 1
Information loss, $\eta$, for different values of $m$ and $k$

| $\eta$ (%) | | | | | | | |
|---|---|---|---|---|---|---|---|
| $k$ | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| $m=k+1$ | 59.80 | 59.76 | 58.56 | 58.56 | 57.63 | 57.63 | 56.87 |
| $m=k+2$ | 83.09 | 81.89 | 80.95 | 80.01 | 79.24 | 78.48 | 77.83 |
| $m=k+3$ | 93.99 | 93.05 | 92.61 | 91.83 | 91.44 | 90.80 | 90.45 |

In our case, the set of GM are normalized so that they are invariant to RST by

$$\hat{m}_{pq} = \alpha^{p+q+2} \int_{-1}^{1}\int_{-1}^{1} [(x-x^c)\cos(\theta)+(y-y^c)\sin(\theta)]^p \times [-(x-x^c)\sin(\theta)+(y-y^c)\cos(\theta)]^q \times f(x,y)\,\mathrm{d}x\,\mathrm{d}y, \tag{37}$$

where $(x^c, y^c) = (m_{10}/m_{00}, m_{01}/m_{00})$ is the centroid of image $f(x, y)$, and

$$\alpha = [m^S_{00}/m_{00}]^{1/2}, \tag{38}$$

$$\theta = \frac{1}{2}\tan^{-1}\frac{2m_{11}}{m_{20}-m_{02}}, \tag{39}$$

where $m^S_{00}$ is a scaling factor depending on the nature of the images in consideration. $m^S_{00}$ is normally set to 1 for binary images with pixel values 1 and 0 (where $m_{00}$ is typically small), but it is appropriate to set this to a higher value when the images are in grayscale format with pixel values ranging from 0 to 255 (where $m_{00}$ is typically much larger) so that the value of the higher order moments would not be scaled to a meaninglessly small value by the factor $\alpha^{p+q+2}$. In our case, we set $m^S_{00} = 100$. The angle $\theta$ can be made unique by considering different signs of the numerator and denominator.

### 4.3. Simplification of algorithm

We can make some simplifications to the EM algorithm by gaining some insights from the derivation of DCT. The kernels of DCT are *asymptotic* eigenvectors of the correlation matrix:

$$\mathbf{C} = [\rho^{|i-j|}]_{i,j}. \tag{40}$$

A signal with such correlation matrix is said to be a stationary Markov-1 signal [9]. The eigenvalues of Eq. (40) is somewhat complicated and can be simplified by making $\rho \to 1$ [9], and the end results would be DCT. This is equivalent to setting the signal correlation matrix to a constant matrix, i.e. each and every element of the correlation matrix has the same constant value. We show that we can do the same with EM and simplify the algorithm. This is done by setting $c$ in Eq. (26) to 0 and we have $r(x_1, x_2) = r(0, 0)$. The calculation of $\mathbf{A}$ can be simplified as:

$$\mathbf{A} = \int_{-1}^{1}\int_{-1}^{1} \left[x_1^j x_2^i\right]_{i,j=0}^{i,j=m-1} \mathrm{d}x_1\,\mathrm{d}x_2. \tag{41}$$

In this case $\mathbf{A}$ will degenerate into a matrix of rank 1. In spite of this, the procedure of obtaining EM detailed in Section 3 is still applicable. Although in this case only the first eigenvalue is nonzero, the eigenvectors of the zero eigenvalues still give discriminant information. This is due to the fact that natural images generally divert from the ideal case of $c = 0$.

### 4.4. Information loss

Some information of the image will be lost when we discard the noise axes. This can be measured by

$$\eta = 1 - \frac{\mathrm{trace}\,(\mathbf{W}_1^{\mathrm{T}}\mathbf{A}\mathbf{W}_1)}{\mathrm{trace}\,(\mathbf{D}_{\mathrm{B}}^{-1/2}\mathbf{P}^{\mathrm{T}}\mathbf{A}\mathbf{P}\mathbf{D}_{\mathrm{B}}^{-1/2})} = 1 - \frac{\mathrm{trace}\left(\widehat{\mathbf{D}}_{\mathrm{B}}^{-1/2}\widehat{\mathbf{P}}^{\mathrm{T}}\mathbf{A}\widehat{\mathbf{P}}\,\widehat{\mathbf{D}}_{\mathrm{B}}^{-1/2}\right)}{\mathrm{trace}\,(\mathbf{D}_{\mathrm{B}}^{-1/2}\mathbf{P}^{\mathrm{T}}\mathbf{A}\mathbf{P}\mathbf{D}_{\mathrm{B}}^{-1/2})}. \tag{42}$$

Some values are shown in Table 1. Although the numbers seem unfavorably large, we shall see, as proven by the results of the experiments, that these discarded information is irrelevant for the purpose for classification. A related discussion about the distinction between the most expressive features (MEF) and the most discriminant features (MDF) can be found in Ref. [11].

## 5. Experimental studies

In this section, we gauge the performance of the set of EM by performing a number of experiments. The outcomes of the experiments show that EM give positive significant improvements when compared to GM [1], LM [4], TM [5] and DCT [9].

### 5.1. Images

The images utilized in our experiments are taken from the COREL photograph data set, which are used in Ref. [12]. The database consists of 1000 color images which are stored in JPEG format with size 384 × 256 or 256 × 384.

Table 2
Categories of images

| ID | Category name |
|---|---|
| 1 | African people and villages |
| 2 | Beaches |
| 3 | Buildings |
| 4 | Buses |
| 5 | Dinosaurs |
| 6 | Elephants |
| 7 | Flowers |
| 8 | Horses |
| 9 | Mountains and glaciers |
| 10 | Food |

The images can be divided into 10 categories, each consisting of 100 images, as listed in Table 2. Most categories include images containing the specific objects. Images in "African people and villages" class are those depicting the African people during their various activities in their village. The "buildings" class refers to a variety of buildings of different architectural styles. The "beach" class refers to sceneries of coasts. Some sample images are shown in Fig. 3. Each image is converted to grayscale and rescaled to a standard size of $128 \times 128$ before its features are extracted.

### 5.2. Method of evaluation

For each image a feature vector of length $L$ can be constructed:

$$\mathbf{\Gamma} = [\gamma_1, \ldots, \gamma_L]. \tag{43}$$

This feature vector serves as a compact *descriptor* of the image. Using $\mathbf{\Gamma}$ we can perform a series of classification tasks to gauge the performance of the set of EM. For all cases we use the squared Euclidean distance:

$$d = \sum_{k=0}^{L} (\gamma_{k,i} - \gamma_{k,j})^2 \tag{44}$$

as a measure of similarity or dissimilarity between the images. Images with small distances are classified as similar. The accuracy of classification is measured by

$$\eta = \frac{N_{correct}}{N_{total}}, \tag{45}$$

where $N_{correct}$ denotes the number of correctly classified images and $N_{total}$ the number of images used in the test.

### 5.3. Length of feature vector

A longer feature length basically means more information is extracted from the image and hence should results in a better classification rate. To ensure fair comparisons, the length of feature vector of each method considered is kept at the same value. For GM, LM and TM, if the degree of the monomials/polynomials are $p$ and $q$ for the $x$-axis and the $y$-axis, respectively, we limit the constituent of the feature vector to those moments with $p, q < T$, where $T$ is a constant. For DCT, we choose the coefficients of the square block $T \times T$ which corresponds to the spectrum of the lowest frequencies. For EM, we choose $k = T$. Hence, for each method used in comparison the feature vector length is $L = T \times T$.

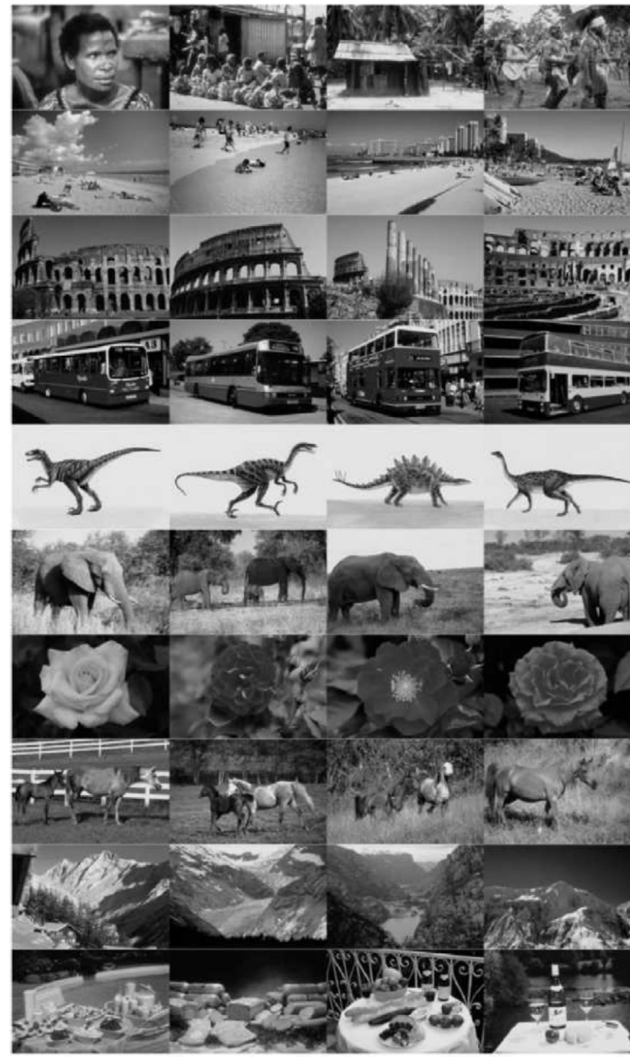

Fig. 3. Some sample images.

### 5.4. Selection of the values of k for EM

We found that it is generally sufficient to discard only the noisiest component (i.e. we discard only one component) in the noise space in order to achieve significant improvement in performance. Hence, in all cases of the experiments performed we let $k = m - 1$ for EM.

### 5.5. Classification under noisy conditions

We test the classification rate of EM under different degrees of Gaussian noise. The training set constitutes of the

Table 3
Average classification rates $\eta$ (%) for images under different degrees of Gaussian noise

| $T$ | $\sigma^2$ | GM | LM | TM | DCT | EM |
|---|---|---|---|---|---|---|
| 4 | 0.05 | 66.06 | 95.54 | 95.56 | 97.09 | 99.82 |
| | 0.10 | 34.39 | 83.59 | 83.59 | 86.79 | 97.61 |
| | 0.15 | 19.75 | 72.21 | 72.22 | 75.86 | 93.05 |
| | 0.20 | 12.85 | 53.72 | 53.72 | 59.64 | 86.11 |
| | 0.25 | 8.45 | 36.72 | 36.73 | 40.63 | 74.13 |
| | 0.30 | 5.68 | 24.76 | 24.79 | 27.85 | 56.74 |
| 5 | 0.05 | 68.49 | 98.03 | 98.03 | 98.70 | 99.91 |
| | 0.10 | 38.35 | 88.67 | 88.68 | 91.82 | 98.97 |
| | 0.15 | 21.33 | 79.57 | 79.57 | 83.54 | 96.09 |
| | 0.20 | 14.18 | 67.71 | 67.75 | 73.92 | 92.45 |
| | 0.25 | 9.65 | 51.85 | 51.91 | 59.01 | 85.94 |
| | 0.30 | 6.12 | 35.43 | 35.45 | 41.20 | 74.13 |
| 6 | 0.05 | 69.57 | 98.88 | 98.88 | 99.44 | 99.97 |
| | 0.10 | 40.13 | 91.63 | 91.63 | 94.54 | 99.53 |
| | 0.15 | 22.43 | 83.75 | 83.74 | 86.54 | 97.71 |
| | 0.20 | 14.38 | 75.20 | 75.24 | 80.04 | 95.18 |
| | 0.25 | 9.81 | 62.36 | 62.38 | 69.52 | 91.84 |
| | 0.30 | 6.67 | 45.48 | 45.49 | 53.41 | 83.96 |
| 7 | 0.05 | 70.54 | 99.32 | 99.32 | 99.69 | 100.00 |
| | 0.10 | 42.16 | 94.12 | 94.13 | 96.25 | 99.81 |
| | 0.15 | 23.55 | 86.13 | 86.13 | 89.82 | 98.79 |
| | 0.20 | 15.05 | 79.57 | 79.59 | 83.72 | 96.81 |
| | 0.25 | 10.32 | 68.71 | 68.76 | 75.65 | 94.02 |
| | 0.30 | 6.88 | 53.47 | 53.52 | 60.71 | 88.96 |
| 8 | 0.05 | 71.58 | 99.65 | 99.65 | 99.84 | 100.00 |
| | 0.10 | 43.32 | 95.59 | 95.60 | 97.16 | 99.95 |
| | 0.15 | 24.24 | 87.90 | 87.91 | 92.23 | 99.43 |
| | 0.20 | 15.47 | 82.39 | 82.42 | 86.41 | 98.03 |
| | 0.25 | 10.48 | 73.97 | 74.01 | 79.45 | 95.76 |
| | 0.30 | 7.28 | 60.07 | 60.12 | 67.16 | 91.90 |
| 9 | 0.05 | 72.14 | 99.76 | 99.76 | 99.97 | 100.00 |
| | 0.10 | 44.27 | 96.75 | 96.77 | 97.89 | 100.00 |
| | 0.15 | 25.18 | 90.25 | 90.29 | 94.13 | 99.69 |
| | 0.20 | 16.38 | 84.62 | 84.62 | 88.16 | 98.65 |
| | 0.25 | 10.72 | 77.70 | 77.75 | 82.28 | 97.14 |
| | 0.30 | 7.33 | 63.83 | 63.87 | 71.15 | 94.34 |
| 10 | 0.05 | 72.74 | 99.88 | 99.89 | 99.99 | 100.00 |
| | 0.10 | 45.17 | 97.39 | 97.39 | 98.50 | 100.00 |
| | 0.15 | 25.82 | 92.19 | 92.24 | 95.41 | 99.91 |
| | 0.20 | 16.36 | 86.08 | 86.10 | 89.73 | 99.27 |
| | 0.25 | 10.73 | 80.10 | 80.15 | 84.37 | 97.85 |
| | 0.30 | 7.60 | 68.05 | 68.12 | 74.20 | 96.05 |

clean images which are not contaminated by noise. The feature vector for each image in the training set is generated and stored. This means that for each of the method considered, i.e. GM, LM, TM, DCT, we have 1000 feature vectors per method. The set of 1000 images are then degraded with increasing degree of Gaussian noise by increasing the noise variance, i.e. $\sigma^2 = 0.05, 0.10, \ldots, 0.30$. For each degree of Gaussian noise added, the 1000 noise contaminated images, used as the testing set, are classified by comparing their feature vectors with those stored for the 1000 clean images. For this purpose the distance (44) is used. The average results for 10 repeated trials are shown in Table 3. It can be seen that the size of the feature vector $L = T \times T$ increases, the classification rates increase. Among the methods considered, EM performed consistently better. For drastic noise condition such as when $\sigma^2 = 0.30$, an example of the noise degraded image shown in Fig. 4, the classification accuracy of EM is still remarkably high and the accuracy is higher than 90% for $T \geqslant 8$. In these cases, the classification accuracy of EM is better than the other methods used in comparison: approximately 20% higher when compared to DCT, 30%

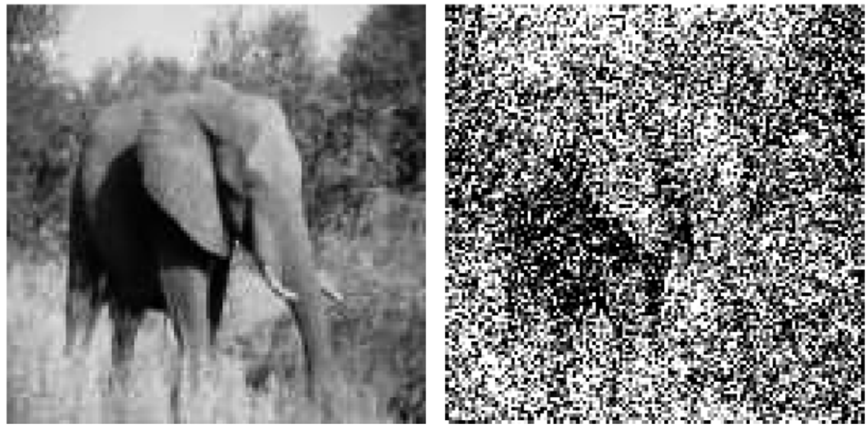

Fig. 4. A sample of the clean images and its noise contaminated version. The noise added is Gaussian noise with $\sigma^2 = 0.30$.

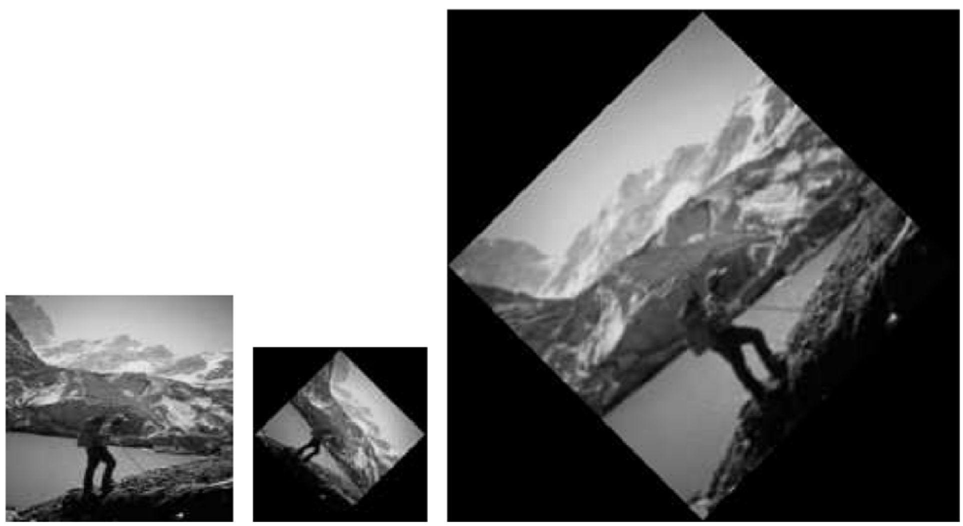

Fig. 5. Some samples of the transformed images. The left most is the original image.

when compared to LM and TM and over 80% when compared to GM.

### 5.6. Classification under RST and noise distortions

In this experiment we test the invariant property of the EMI. We selected five images from each category as training images, resulting in a training set of 50 images. The testing images comprise of the transformed and noise contaminated versions of the 50 selected images.[2] The images are first scaled with factors $s = 0.5$, 1.0, 1.5 and rotated at angles $\theta = 0°, 45°, \ldots, 315°$. The transformed images are then added with Gaussian noise of variance $\sigma^2 = 0.00, 0.02, \ldots, 0.10$. Some examples of the transformed images are shown in Fig. 5. These transformed and noise contaminated image are then classified according to their feature vectors. The average results for 10 trials are shown in Table 4. Also shown for comparison are the results for Hu's moment invariants (using principal axis, denoted as GMI) [1]. It can be seen that EMI perform consistently better in all cases with typical improvement margin of approximately 30% in cases where $T \geqslant 8$.

### 5.7. Information compaction and noise robustness

Table 3 shows that EM are capable of giving high classification rates of over 90% under noisy conditions. This is achieved by using a compact feature vector which has a typical length of $L = T \times T = 8 \times 8 = 64$. Compared to the information contained in the original image $128 \times 128 = 16,384$, this is an apparent compaction in information energy, reaching a compaction factor of $16,384/64 = 256$. This means that the classification task can be performed very much faster since only comparisons of feature vectors need to be done, instead of the pixel values. Furthermore, a direct comparison

[2] We apparently could not use all the 1000 images, this will cause a huge test set which can easily reach 100,000 images if 100 transformed images were generated from each training image.

Table 4
Average classification rates $\eta$ (%) for images under RST and noise distortions

| $T$ | Feature | $\sigma^2$ | | | | | |
|---|---|---|---|---|---|---|---|
| | | 0.00 | 0.02 | 0.04 | 0.06 | 0.08 | 0.10 |
| 4 | GMI | 100.00 | 52.22 | 47.04 | 40.88 | 31.27 | 26.52 |
| | EMI | 99.70 | 69.54 | 55.93 | 50.23 | 43.87 | 34.94 |
| 5 | GMI | 100.00 | 50.23 | 47.27 | 39.85 | 32.64 | 28.79 |
| | EMI | 99.70 | 80.27 | 61.15 | 52.87 | 48.00 | 43.30 |
| 6 | GMI | 100.00 | 50.57 | 47.20 | 40.27 | 33.32 | 28.75 |
| | EMI | 99.30 | 86.09 | 73.31 | 60.15 | 51.50 | 47.58 |
| 7 | GMI | 100.00 | 50.37 | 46.98 | 40.12 | 32.99 | 29.13 |
| | EMI | 99.30 | 86.75 | 81.36 | 66.84 | 57.27 | 51.97 |
| 8 | GMI | 100.00 | 50.61 | 47.10 | 39.82 | 33.14 | 29.00 |
| | EMI | 99.30 | 87.54 | 83.26 | 76.92 | 65.97 | 58.96 |
| 9 | GMI | 100.00 | 50.45 | 46.96 | 39.92 | 33.26 | 29.28 |
| | EMI | 99.30 | 86.15 | 83.07 | 78.78 | 72.88 | 66.11 |
| 10 | GMI | 100.00 | 50.39 | 46.75 | 39.95 | 33.25 | 29.13 |
| | EMI | 99.30 | 85.88 | 83.02 | 78.31 | 73.85 | 66.94 |

Table 5
Computation complexity of Eigenmoments

| | Number of operations | Special case $m=k+1$ | Typical value ($k=8$) |
|---|---|---|---|
| + | $k(m-1)(m+k)$ | $k^2(2k+1)$ | 1088 |
| × | $km(m+k)$ | $k(k+1)(2k+1)$ | 1224 |

of the pixel values are often susceptible to the effect of the noise, whereas by using EM the noise components are effectively discarded.

### 5.8. Computation complexity

The transformation from the set of conventional GM, **M**, to the set of EM takes only ordinary matrix multiplication as noted in Eq. (35) (Table 5). Since the sizes of **W** and **M** are $m \times k$ and $m \times m$, respectively. We can calculate EM with just a small number of extra mathematical operations, i.e. $km(m-1) + k^2(m-1) = k(m-1)(m+k)$ additions and $km^2 + k^2m = km(m+k)$ multiplications. For our case, we let $k = m - 1 \Rightarrow m = k + 1$, hence, the number of extra operations can be written in terms of $k$: $k^2(2k+1)$ additions and $k(k+1)(2k+1)$ multiplications. A summary and the typical values for the case of $T = 8$ are shown in Table 5. Note that the extra operations are a small price to pay considering the benefits which can be reaped from EM.

## 6. Conclusion

In this paper, we present a method for obtaining a set of orthogonal, noise-robust, transformation invariant and distribution sensitive moments, which we call *Eigenmoments* (EM). EM are obtained by performing eigen analysis in the moment space generated by geometric moments (GM). This is done by transforming the *moment space* into the feature space where the signal-to-noise ratio (SNR) is maximized. This is equivalent to solving a generalized eigenvalue problem related to a Rayleigh quotient which characterizes the SNR. The GEP can be decomposed into two eigenvalue problems. In the first eigenvalue problem, the moment space is transformed into the *noise space* where the noise components are removed. In the second eigenvalue problem a second transformation is performed to find the most expressive components.

This approach has several advantages. (1) Since the distribution of the moments in the moment space is dependent on the distribution of the images being transformed, eigen analysis in the moment space ensures the decorrelation of the final feature space. (2) The ability of EM to take the image distribution in consideration makes it more versatile and adaptable to images of different genres. (3) The moment kernel generated is orthogonal and this makes the analysis of the moment space much more easier. Transformation with orthogonal moment kernel into the moment space is equivalent to the projection of the image onto a number of orthogonal axes. (4) Noisy components can be effectively removed as is evidenced by the empirical results. This makes EM robust in classification task especially when the images are corrupted with noise. (5) Optimal information compaction can be obtained and hence only a small number of moments are needed to characterize the images.

Experiments are performed to gauge the performance of EM and comparisons are made with some well known feature descriptors such as geometric moments (GM), DCT, Legendre moments (LM) and Tchebichef moments (TM).

Table 6
Comparison of eigenmoments (EM) with geometric moments (GM)

| | EM | GM |
|---|---|---|
| Optimal information compaction | ✓ | × |
| Noise robustness | ✓ | × |
| Adaptable to different image distributions | ✓ | × |
| Decorrelates the feature space | ✓ | × |
| Projection onto orthogonal axes | ✓ | × |

The results show that EM give significant improvements in terms of accuracy and noise robustness as predicted by the theoretical framework. When compared to DCT, LM and TM improvement margins can reach up to more than 30%. When compared to GM the improvements are over 80%. Under RST transformations, the improvement is typically 30% over that of GMI. The advantages of EM are summarized in Table 6.

## Acknowledgment

The authors gratefully acknowledge the anonymous reviewers for the many helpful comments and criticism on this paper.

## Appendix A. Eigenvalues and the maximization of the Rayleigh quotient

We want to prove that

$$\max_{\mathbf{w}^{\mathrm{T}}\mathbf{B}\mathbf{w}_1=\cdots=\mathbf{w}^{\mathrm{T}}\mathbf{B}\mathbf{w}_{i-1}=0} R(\mathbf{w}) = \lambda_i. \tag{46}$$

We first let $\mathbf{w} = \mathbf{W}\mathbf{z}$, where $\mathbf{z} \in \mathscr{R}^n$ is the vector of a set of properly defined coefficients. Then the Rayleigh quotient can be written as

$$\frac{\mathbf{z}^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}\mathbf{A}\mathbf{W}\mathbf{z}}{\mathbf{z}^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}\mathbf{B}\mathbf{W}\mathbf{z}} = \frac{\mathbf{z}^{\mathrm{T}}\mathbf{\Lambda}\mathbf{z}}{\mathbf{z}^{\mathrm{T}}\mathbf{z}} = \frac{\lambda_1 z_1^2 + \cdots + \lambda_n z_n^2}{z_1^2 + \cdots + z_n^2}. \tag{47}$$

Since $\mathbf{w}^{\mathrm{T}}\mathbf{B}\mathbf{w} = 1$, we have

$$\mathbf{w}^{\mathrm{T}}\mathbf{B}\mathbf{w} = \mathbf{z}^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}\mathbf{B}\mathbf{W}\mathbf{z} = \mathbf{z}^{\mathrm{T}}\mathbf{z} = 1. \tag{48}$$

That is,

$$z_1^2 + \cdots + z_n^2 = 1. \tag{49}$$

Hence, since $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_n$, the maximum of $R(\mathbf{w})$ is clearly:

$$\lambda_1 = \frac{\lambda_1(z_1^2 + \cdots + z_n^2)}{z_1^2 + \cdots + z_n^2} \geqslant \frac{\lambda_1 z_1^2 + \cdots + \lambda_n z_n^2}{z_1^2 + \cdots + z_n^2}. \tag{50}$$

But if there are constraints:

$$\mathbf{w}^{\mathrm{T}}\mathbf{B}\mathbf{w}_1 = \cdots = \mathbf{w}^{\mathrm{T}}\mathbf{B}\mathbf{w}_{i-1} = 0 \tag{51}$$

then it should be fairly clear that

$$z_1 = \cdots = z_{i-1} = 0 \tag{52}$$

and hence the maximum value becomes

$$\lambda_i = \frac{\lambda_i(z_i^2 + \cdots + z_n^2)}{z_i^2 + \cdots + z_n^2} \geqslant \frac{\lambda_i z_i^2 + \cdots + \lambda_n z_n^2}{z_i^2 + \cdots + z_n^2}. \tag{53}$$

This ends the proof.

## References

[1] M. Hu, Visual pattern recognition by moment invariants, IRE Trans. Inf. Theory IT-8 (1962) 179–187.
[2] G.H. Golub, C.F.V. Loan, Matrix Computations, second ed., The Johns Hopkins University Press, Baltimore, MD, 1989.
[3] R. Askey, Gabor Szegö: Collected Papers, vols. I–III, Birkhäuser, Boston, 1982.
[4] M. Teague, Image analysis via the general theory of moments, J. Opt. Soc. Am. 70 (8) (1980) 920–930.
[5] R. Mukundan, S. Ong, P. Lee, Image analysis by Tchebichef moments, IEEE Trans. Image Process. 10 (9) (2001) 1357–1364.
[6] T. De Bie, N. Cristianini, R. Rosipal, Eigenproblems in pattern recognition, in: E. Bayro-Corrochano (Ed.), Handbook of Computational Geometry for Pattern Recognition, Computer Vision, Neurocomputing and Robotics, Springer, Heidelberg, 2004.
[7] G. Strang, Linear Algebra and Its Applications, second ed., Academic Press, New York, 1980.
[8] K. Fukunaga, Introduction to Statistical Pattern Recognition, Computer Science and Scientific Computing, second ed., Academic Press, New York, 1990.
[9] K.R. Rao, P. Yip, Discrete Cosine Transform—Algorithms, Advantages and Applications, Academic Press, New York, 1990.
[10] C. Teh, R. Chin, On image analysis by the method of moments, IEEE Trans. Pattern Anal. Mach. Intell. 10 (4) (1988) 496–513.
[11] D.L. Swets, J.J. Wang, Using discriminant eigenfeatures for image retrieval, IEEE Trans. Pattern Anal. Mach. Intell. 18 (8) (1996) 831–836.
[12] J.Z. Wang, J. Li, G. Wiederhold, SIMPLIcity: semantics-sensitive integrated matching for picture libraries, IEEE Trans. Pattern Anal. Mach. Intell. 23 (9) (2001) 947–963.