Quantized Principal Component Analysis with Applications to Low-Bandwidth Image Compression and Communication

David Wooden, Magnus Egerstedt
{wooden,magnus}@ece.gatech.edu
School of Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, GA 30332, USA

Bijoy K. Ghosh
ghosh@netra.wustl.edu
Department of Electrical and Systems Engineering
Washington University in St. Louis
St. Louis, MO 63130, USA

Abstract

In this paper we show how Principal Component Analysis can be mapped to a quantized domain in an optimal manner. In particular, given a low-bandwidth communication channel over which a given set of data is to be transmitted, we show how to best compress the data. Applications to image compression are described and examples are provided that support the practical soundness of the proposed method.

1 Introduction

Principal Component Analysis (PCA) is an algebraic tool for compressing large sets of statistical data in a structured manner. However, the reduction results in real-valued descriptions of the data. In this paper, we take the compression one step further by insisting on the use of only a finite number of bits for representation. This is necessary in a number of applications where the data is transmitted over low-bandwidth communication channels. In particular, the inspiration for this work came from the need for multiple mobile robots to share visual information about their environment.

Assuming that the data $x_1, \ldots, x_N$ takes on values in a d-dimensional space, one can identify the d principal directions, coinciding with the eigenvectors of the covariance matrix, given by

$$C = \frac{1}{N}\sum_{i=1}^{N}(x_i - m)(x_i - m)^T, \tag{1}$$

where m is the mean of the data.

If our goal is to compress the data set to a set of dimension $n < d$, we would pick the n dominant directions, i.e. the directions of maximum variation of the data. This results in an optimal (in the sense of least squared error) reduction of the dimension from d to n. For example, if $n = 0$ then only the mean is used, while $n = 1$ corresponds to a 1-dimensional representation of the data. The fact that the reduction can be done in a systematic and optimal manner has lead to the widespread use of PCA in a number of areas, ranging from process control [7], to weather prediction models [3], to image compression [2]. In this paper, we focus on and draw inspiration from the image processing problem in particular, even though the results are of a general nature.

2 Principal Component Analysis

Suppose we have a stochastic process with samples $x_k \in \mathbb{R}^d, k = (1, \ldots, N)$, where N is the number of samples taken. Let

1. $m = \frac{1}{N}\sum_{k=1}^{N} x_k$ be the mean of the input data.
2. $e_i \in \mathbb{R}^d$ be the ith principal direction of the system, where $i \in \{1, \ldots, d\}$
3. $a_i \in \mathbb{R}^d$ be the ith principal component, i.e.

$$a_i = {e_i}^T(x - m),$$

associated with the sample x. We can then reconstruct x perfectly from its principal components and the system's principal directions as

$$x = m + \sum_{i=1}^{d} a_i e_i. \tag{2}$$

If we wish to reduce the system complexity from a d-dimensional data set to n dimensions, only the n principal directions (corresponding to the n largest eigenvalues) should be chosen.

The main contribution in this paper is not the problem of reducing the dimension of the data set, but rather the problem of communicating the data. Given $n \leq d$ number of transmittable real numbers, the optimal choice for the reconstruction of x from these numbers is simply given by the n largest (in magnitude) principal components. But because the a_i are all real-valued, we are required to quantize them prior to communication, and we wish then to transmit only the most significant quanta. In this paper we derive a mapping of the PCA algorithm to a quantized counterpart.

3 Quantized Components

Let $r \in \mathbb{N}$ be the resolution of our system. For example, if $r = 10$, then we are communicating decimal integers. If $r = 16$, then we are communicating nibbles (i.e. half-bytes). In other words, $r = 2^b$, where b is the number of bits used for quantization. Also, let $K \in \mathbb{Z}$ be the largest integer exponent of r such that

$$\frac{max(|a_i|)}{r^K} \in \mathbb{R}_{[1,r-1]}.$$

With this definition of r and K, we are equipped to define the quantities by which we will decompose the principal components. We name the first quantity the *quantized component*,

$$z_i = \arg\min_{\zeta \in Z} \Big(|a_i - \zeta| \Big), \tag{3}$$

where $Z = r^K\{-r+1, ..., r-1\}$. As a result, we have that

$$0 \leq |z_i| \leq (r-1)r^K, \tag{4}$$

In other words z_i is taken as the integer in range $[-r+1, +r-1]$, which, when scaled by r^K, minimzes the distance to the principal component a_i.

The second quantity, called the *remainder component*, is simply defined as

$$y_i = a_i - z_i, \tag{5}$$

and therefore,

$$0 \leq |y_i| < \frac{1}{2}r^K. \tag{6}$$

The remainder component is equivalent to the round-off error between z_i and a_i.

With these definitions, we define the quantized version of the original prinicipal components to be $a_i^Q \equiv z_i$. And in a manner similar to the reconstruction of x from its principal components, we can reconstruct a quantized version of x from its quantized principal components:

$$x^Q = m + \sum_{i=1}^{d} a_i^Q e_i. \tag{7}$$

Now, the question remains, if we may only transmit one quantized component, which one should we pick?

Problem 1. Identify the quantized component which minimizes the error between x^Q and x. In other words, solve

$$\arg\min_{k \in \{1,...,d\}} \left\| \Big(m + \sum_{i=1}^{d} \delta_{ik} a_i^Q e_i \Big) - x \right\|^2 \tag{8}$$

where

$$\delta_{ik} = \begin{cases} 1, & if\ i = k \\ 0, & otherwise \end{cases}$$

4 Main Result

Define

$$S = \{1, \ldots, d\}$$

and

$$S_z = \Big\{ s \in S \Big| \ |z_s| \geq |z_m|, \forall m \in S \Big\}.$$

Theorem 1. If $|S_z| = 1$, i.e. $\exists\ n \in S$ such that $|z_n| > |z_m|\ \forall m \in S,\ m \neq n$, then n is the solution to Problem 1, i.e. z_n is the optimal component to transmit.

Proof. Define the cost function

$$\begin{aligned} J(a^Q) &= \left\| x^Q - x \right\|^2 = \\ &= \left\| (m + \sum_{i=1}^{d} a_i^Q e_i) - (m + \sum_{i=1}^{d} a_i e_i) \right\|^2 = \\ &= \left\| \sum_{i=1}^{d} e_i (a_i^Q - a_i) \right\|^2 = \\ &= \sum_{i=1}^{d} (a_i^Q)^2 - 2a_i^Q a_i + a_i^2 = \\ &= \sum_{i=1}^{d} z_i^2 - 2z_i(z_i + y_i) + a_i^2 = \\ &= -\sum_{i=1}^{d} \Big(z_i^2 + 2z_i y_i \Big) + \sum_{i=1}^{d} a_i^2. \end{aligned} \tag{9}$$

Clearly, as the last summation in the equation above is constant for a given system, it will not affect the optimization. Now, define a similar cost function

$$J_k(a^Q) = \left\| m + \sum_{i=1}^{d} \delta_{ik} a_i^Q e_i - x \right\|^2.$$

Hence,

$$\begin{aligned} J_k(a^Q) &= \sum_{i=1}^{d} \delta_{ik} \Big((a_i^Q)^2 - 2a_i^Q a_i \Big) + a_i^2 = \\ &= (a_k^Q)^2 - 2a_k^Q a_k + \sum_{i=1}^{d} a_i^2 = \\ &= -\Big(z_k^2 + 2z_k y_k \Big) + \sum_{i=1}^{d} a_i^2. \end{aligned} \tag{10}$$

Again, the last summation is immutable, and can be ignored when minimizing J_k.

Taking $n, m \in S$, we may extend Equation 10 to write

$$\begin{aligned} J_n - J_m = & -(z_n^2 + 2z_n y_n) + (z_m^2 + 2z_m y_m) \\ = & -z_n^2 - 2|z_n||y_n| sgn(z_n) sgn(y_n) + \\ & + z_m^2 + 2|z_m||y_m| sgn(z_m) sgn(y_m) \end{aligned} \tag{11}$$

where $sgn(z_i)$ indicates the sign (+ or −) of z_i. Since $sgn(z_i)sgn(y_i) \in \{-1, +1\}$, we may write

$$J_n - J_m \leq -z_n^2 + 2|z_n||y_n| + z_m^2 + 2|z_m||y_m| \tag{12}$$

Now, assume that $|z_n| > |z_m|$, which gives us

$$|z_n| = |z_m| + \alpha r^K, \tag{13}$$

where $\alpha \in \mathbb{Z}_+$. Hence, we wish to show that $J_n - J_m < 0$.

Substituting Equation 13 into Equation 12,

$$\begin{aligned} J_n - J_m \leq \quad & -(|z_m| + \alpha r^K)^2 + 2|y_n|(|z_m + \alpha r^K) + \\ & + z_m^2 + 2|z_m||y_m|. \end{aligned}$$

Using Equation 6, we conclude that

$$\begin{aligned} J_n - J_m \quad &< \quad -(|z_m| + \alpha r^K)^2 + (|z_m| + \alpha r^K) r^K + \\ & \quad + z_m^2 + |z_m| r^K \\ &< \quad 2|z_m| r^K - 2|z_m| \alpha r^K - r^{2K}\Big(\alpha^2 - \alpha\Big) \\ &< \quad \Big(1 - \alpha\Big)\Big(2|z_m| r^K + \alpha r^{2k}\Big). \end{aligned}$$

And finally, recalling that $\alpha \geq 1$,

$$J_n - J_m < 0, \tag{14}$$

and the theorem follows. □

Now, define $S_= = \Big\{ s \in S_z \;\Big|\; sgn(z_s) = sgn(y_s) \Big\}$ and $S_{\neq} = S_z \backslash S_=$. Moreover, define $S_y^+ = \Big\{ s \in S_= \;\Big|\; |y_s| \geq |y_l|, \forall l \in S_= \Big\}$. In other words, S_y^+ refers to the principal component(s) with the largest quantized and remainder components which also have equal signs.

Theorem 2. If $|S_z| > 1$ and $|S_y^+| > 0$, i.e. $\exists\ n \in S_y^+$ such that $|y_n| \geq |y_m|$, $|z_n| = |z_m|\ \forall m \in S_z$ and $sgn(z_n) = sgn(y_n)$, then n is the solution to Problem 1, i.e. z_n is the optimal component to transmit.

Proof. By assumption, $|S_z| > 1$, and it is a direct consequence of Theorem 1 that we should choose between the elements of S_z for a quantized component to transmit. Recalling Equation 11,

$$\begin{aligned} J_n - J_m \quad &= \quad -z_n^2 - 2z_n y_n + z_m^2 + 2z_m y_m \\ &= \quad -2z_n y_n + 2z_m y_m \\ &= \quad -2|z_n|(|y_n| - |y_m| sgn(z_m) sgn(y_m)). \end{aligned}$$

It is true that either $m \in S_{\neq}$ or $m \in S_=$. When $m \in S_{\neq}$, $sgn(z_m)sgn(y_m) = -1$ and therefore

$$J_n - J_m = -2|z_n|(|y_n| + |y_m|) < 0. \tag{15}$$

On the other hand, when $m \in S_=$, $sgn(z_m)sgn(y_m) = +1$ and

$$J_n - J_m = -2|z_n|(|y_n| - |y_m|). \tag{16}$$

Note again that $n \in S_y^+$ (i.e. $|y_n| \geq |y_m|$). Hence, Equation 16 becomes

$$J_n - J_m \leq 0. \tag{17}$$

Furthermore, $J_n - J_m = 0$ only when $|y_n| = |y_m|$ and $|z_n| = |z_m|$. In other words, $|a_n| = |a_m|$ and clearly the two cost funtions are equal. □

Finally, define $S_y^- = \Big\{ s \in S_{\neq} \Big|\ |y_s| \leq |y_l|, \forall l \in S_{\neq} \Big\}$.

Theorem 3. If $|S_z| > 1$ and $|S_y^+| = 0$, i.e. $sgn(z_s) \neq sgn(y_s)\ \forall s \in S_z$ and $\exists\ n \in S_y^-$ such that $|z_n| = |z_m|$ and $|y_n| \leq |y_m|\ \forall m \in S_z$, then n is the solution to Problem 1, i.e. $|z_n|$ is the optimal component to transmit.

Proof. As was the case under Theorem 2, $|S_z| > 1$, but $S_=$ and S_y^+ are empty (indicating that the signs of the quantized and remainder components differ for all those in S_z). We prove then that the optimal quantized component to transmit is the one with largest $|z_i|$ and the smallest $|y_i|$. Recalling again Equation 11,

$$\begin{aligned} J_n - J_m \quad &= \quad -z_n^2 - 2z_n y_n + z_m^2 + 2z_m y_m \\ &= \quad -2z_n y_n + 2z_m y_m \\ &= \quad 2|z_n|(|y_n| - |y_m|) \\ &< \quad 0, \end{aligned}$$

and the theorem follows. □

Theorem 1 tells us that we should transmit the quantized component largest in magnitude. If this is not unique, Theorem 2 tells us to send the z_i which also has the largest remainder component which points in the same direction as its quantized component (i.e. $sgn(z_i) = sgn(y_i)$). According to Theorem 3, if no such remainder component exists (i.e. all y_i point opposite of their quantized counterparts), then we send the z_i with the *smallest* remainder component.

When a unique largest (in magnitude) quantized component exists, it is a direct result of Equations 5 and 6 that it corresponds to the largest (in magnitude) principal component. In other words, Theorem 1 tells us to send the quantized component of the largest principal component.

As a consequence of Equation 5, when the remainder component has the same sign as the quantized component, the corresponding principal component is larger (again, in magnitude) than the quantized component. Moreover, given $|z_i|$ and that z_i and y_i have matching signs, $|a_i|$ is maximized by the largest value possible for $|y_i|$. Theorem 2 tells us that, given the $|z_i|$ which are largest in magnitude, transmit the z_i which has matching remainder and quantized component signs and has the largest $|y_i|$. In other words, Theorem 2 tells us to transmit the z_i corresponding to the largest $|a_i|$.

Similarly, given $|z_i|$ and mismatching signs of z_i and y_i, $|a_i|$ is maximized by the *smallest* value possible for $|y_i|$. Theorem 3 tells us that, given the $|z_i|$ which are largest in magnitude, transmit the z_i which has mismatching remainder and quantized component signs and has the smallest $|y_i|$. In other words, Theorem 3 tells us to transmit the z_i corresponding to the largest $|a_i|$.

Theorem 1, 2, and 3 tell us, given a set of principal, quantized, and remainder components, which quantized component should be transmitted. This optimal quantized component is always the one corresponding the principal component largest in magnitude.

5 Image Compression Examples

We apply the proposed quantization scheme to a problem in which images are to be compressed and transmitted over low-bandwidth channels. Two separate data sets are used. The first set is comprised of 12 352x352 pixel grayscale images of very similar scenes (Figure 1). The second is comprised of 118 64x96 pixel grayscale images of very different scenes (Figure 2).

The images themselves are represented in two different ways. In the first method, as is common practice in image processing [5] [4], the images are broken into 8x8 pixel blocks. Principal components and directions are computed over these 64-pixel pieces, and the quantization is computed for each block. At each additional iteration of the algorithm, one more quantized component is added for each block of the image. In the second method, principal directions and components are computed over the entire image, as a whole. At each iteration under this method, only a single quantized component is added.

There are computational advantages/disadvantages associated with employing either method. In the first case, the mean-squared error (MSE) drops off at a slower rate, but computing the principal directions is easy, and the memory required to hold them is small. In the second case, the MSE drops off dramatically fast, but computing and maintaining the principal directions can be utterly intractable. In other words, the burden associated with computing larger and larger principal directions seems to be inversely proportional to the payoff of low MSE in the reconstructed image, as seen in Figure 4.

As shown in Figure 4, whether the image sets are broken into blocks or not, the MSE of any image drops off at approximately a log-linear rate with respect to the number of quantized components transmitted. Given this behavior, it may be possible to predict how many quantized components are required so that the expected MSE of a transmitted image is below a certain threshold.

As stated, we have considered two image representation cases for our image sets; either we break the image into subimages (8x8 blocks) or we leave it as a whole image. The latter is equivalent to breaking the image into blocks the same size as the image. In the first case, we have sacrificed the error drop-of rate per quantized component for reduced computational burden. In the second, we have done the reverse. We should expect a middle ground to exist between the two cases where we try to balance these two competing objectives. That is, the images should be broken into blocks of some certain size that takes sufficiently little time to compute, but requires a relatively small number of quantized components.

Figure 3 shows successive iterations of the algorithm on selected images from our two data sets. Each pairing in the figure ((a) with (b), (c) with (d), etc.) shows two progressions of an image, where the first progression is based on block representation of the image, and the second is based on whole representation. The value of r was set to 16, meaning that at each iteration, 4 additional bits of information per block were transmitted.

The PCA algorithm and our Quantized PCA algorithm rely on the agreed knowledge between the sender/reciever of what the data mean and principal directions are. Though these values may indeed not be stationary, it is possible to update them in an online fashion. This can be accomplished, for example, by using the Generalized Hebbian Algorithm [6], [1]. Hence, even in a changing environment, as is common for a group of mobile robots, a realistic model of the surroundings can be updated and maintained.

6 Conclusions

We have proposed a method for transmitting quantized versions of the principal components associated with a given data set in an optimal manner. This method was shown to be applicable to the problem of image compression and transmission over low-bandwidth communication channels.

References

[1] L. Chen, S. Chang: An adaptive learning algorithm for principal component analysis, *IEEE Transactions on Neural Networks*, v6, i5, pp.1255-1263, 1995.

[2] X. Du, B.K. Ghosh, P. Ulinski: Decoding the Position of a Visual Stimulus from the Cortical Waves of Turtles, *Proceedings of the 2003 American Control Conference*, v1, i1, pp.477-82, 2003.

[3] R. Duda, P. Hart, D. Stork: *Pattern Classification*, John Wiley and Sons, Inc., N.Y., 2001.

[4] M. Kunt: Block Coding of Graphics: A Tutorial Review, *Proc. of IEEE*, v68, i7, pp.770-86, 1980.

[5] M. Marcellin, M. Gormish, A. Bilgin, M. Boliek: An Overview of JPEG-2000, *Proc. of IEEE Data Compression Conference*, pp.523-541, 2000.

[6] T. D. Sanger: Optimal unsupervised learning in a single-layer linear feed-forward neural network, *Neural Networks*, v2, i6, pp.459-473, 1989.

[7] C. Undey, A Cinar: Statistical Monitoring of Multistage, Multiphase Batch Processes, *IEEE Control Systems*, v22, i5, pp.40-52, 2002.

Fig. 1: Image Set 1.

Fig. 2: Image Set 2.

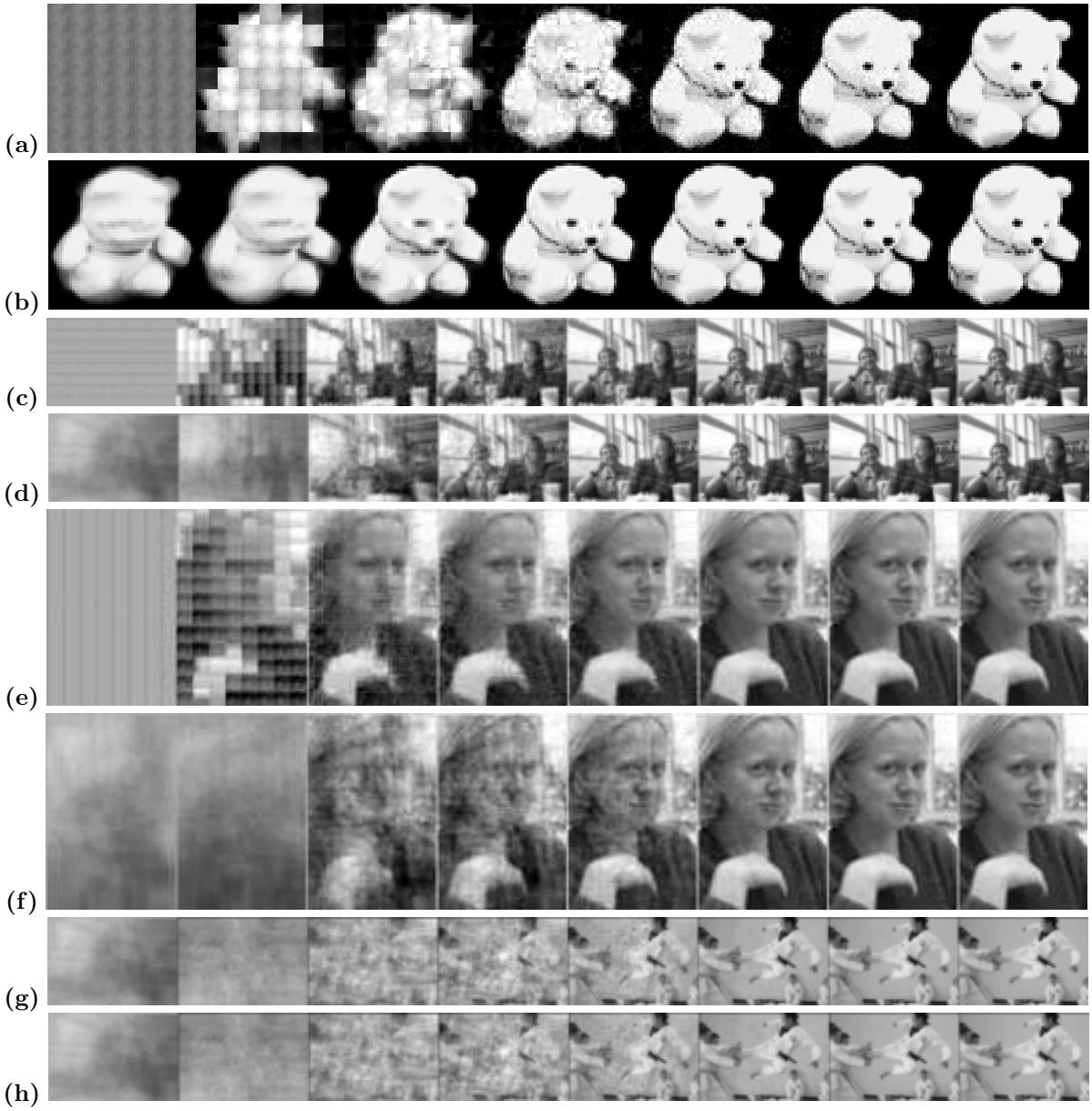
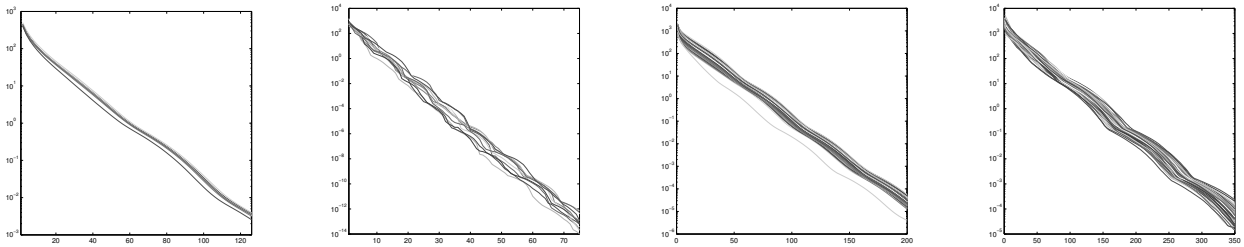


Fig. 3: Progression of Quantized Images - { Version (Image Set, Image Number) Iteration } (a) Block (1,4) 1 3 9 27 40 100 (b) Whole (1,4) 1 3 8 14 20 25 (c) Block (2,23) 1 5 10 15 30 50 80 (d) Whole (2,23) 1 10 24 50 90 140 200 (e) Block (2,80) 1 5 10 15 30 50 80 (f) Whole (2,80) 1 10 24 50 90 140 200 (g) Block (2,120) 1 5 10 15 30 50 80 (h) Whole (2,120) 1 10 24 50 90 140 200



(a) **(b)** **(c)** **(d)**

Fig. 4: *log*(Mean Squared Error) per Iteration of QPCA for (a) Image Set 1 - 8x8 blocks (b) Image Set 1 - Whole (c) Image Set 2 - 8x8 Blocks (d) Image Set 2 - Whole.