Medipix Imaging - evaluation of datasets with PCA

J. S. Butzer[1,2], A. P. H. Butler[3,4,5], P. H. Butler[1,5], P. J. Bones[4], N. Cook[6], L. Tlustos[5]

[1]Physics & Astronomy, University of Canterbury, Christchurch, New Zealand.
[2]Karlsruhe Institute of Technology, Karlsruhe, Germany.
[3]Department of Radiology, University of Otago, Christchurch, New Zealand.
[4]Electrical & Computer Engineering, University of Canterbury, Christchurch, New Zealand.
[5]European Organisation for Nuclear Research (CERN), Geneva, Switzerland.
[6]Medical Physics & Bio-Engineering, Canterbury District Health Board, Christchurch, New Zealand.
Email: jochen@physikag.de

Abstract

Spectral datasets of a watch and a fetal hand have been acquired with the energy-resolving 2D x-ray imaging detector Medipix. We applied principal component analysis (PCA) to evaluate the spectral information in the data. PCA is useful as it identifies the relevant information in a few derived variables that account for most of the variance of the dataset. A scattergram and cluster analysis allow us to group pixels with similar spectral characteristics. With our data, three derived variables display the most relevant information of the full dataset which can be represented in one RGB image. We have begun to apply this method to CT reconstructed slices to separate different materials. Our approach applies PCA to the energy domain and should not be confused with widely used applications of PCA in pattern recognition where it is applied to the spatial domain.

Keywords: image processing, principal component, spectroscopic x-ray imaging, Medipix

1 Introduction

Most current radiographs are images obtained by recording the intensity of photons interacting with an x-ray detector, without recording the energy or wavelength of the incident photons. In recent years, attempts have been made to add spectroscopic (energy) information to radiographic data. The presumed benefit of this spectral information is based on the knowledge of the energy dependency of the transmission coefficient $\mu(E)$ for different materials. Therefore with spectral imaging it is hoped to provide better contrast and novel information to differentiate materials.

A recently developed method for obtaining spectral information is to use a 2D detector capable of recording the energy of incoming photons. Medipix-2 is an example of such a x-ray detector that relies on advances in CMOS technology enabling the construction of hybrid photon counting detectors. As a member of the Medipix-3 collaboration, we have focussed on collecting and analysing data from such photon processing detectors. In this paper we apply PCA to 2D data obtained with a Medipix-2 detector with a longer term goal of applying PCA to 3D spectroscopic data from our CT scanner, dubbed MARS (Medipix All Resolution System) [1].

A traditional technique for obtaining spectroscopic data is to acquire a series of images from different x-ray spectra. Major disadvantages are the need for multiple exposures and overlapping x-ray spectra. The most common form of this technique is to use two different x-ray tube peak voltages. Therefore, this method is often referred to as 'dual-energy' [2]. It is currently being introduced to clinical CT scanners by most of the major medical imaging suppliers. In some non-destructive testing applications, 'multi-energy imaging' describes acquisitions from many different tube voltages. PCA has been applied to multi-energy x-ray data acquired using these techniques [3].

There are several other ways of analysing spectral data. Groups within the Medipix collaborations [4] as well as suppliers of medical x-ray systems [5] have developed algorithms to reconstruct an image of the base materials using knowledge about the materials present in the sample and their energy-dependent absorption coefficients. Remarkable results have been achieved in isolating the locations of contrast agent. Another approach is to look for weighting factors that optimise the signal-to-noise-ratio (SNR) and can be used to construct

978-1-4244-2582-2/08/$25.00

an optimised 'energy weighted' images from the spectral dataset [6].

The approach we describe in this paper uses PCA as a way of datamining spectroscopic images. The goal of PCA is to display most of the variance of the data in a few derived variables. This enhances the contrast and at the same time reduces the dimensionality of the data by describing the important features of the dataset in only a few images that can be easily interpreted. A brief definition of the terms will be given, followed by a description how to apply the method to spectroscopic datasets of a watch and a fetal hand. A simple method of cluster analysis will be used to demonstrate that pixels with similar spectra correspond to regions within the image containing similar materials. Finally, we display the relevant information of the dataset in only one RGB image.

2 Principal component analysis

The goal of PCA is to identify linearly independent patterns of variance within a data set. According to I. Jollife [7], the roots of principal component analysis date back to the late 19th century when both Beltrami and Jordan independently derived the Singular Value Decomposition (SVD) in a form that underlies PCA. However, the first description of PCA in its modern form was by Pearson in 1901 and later Hotelling in 1933. Therefore it is sometimes referred to as Hotelling transform; another name is the Karhunen-Loève transform, after Kari Karhunen and Michel Loève.

One of the first uses in imaging was by Turk and Pentland in 1991 who applied the method to pattern recognition and coined the term 'eigenimaging'. Since then it has been used in many fields ranging from voice recognition to medical imaging, e.g. [8].

In our case the analysis is performed on acquisitions taken of the same object in different spectral bands. This is different to applications in pattern recognition where images of different objects that vary in spatial configuration are compared.

In the following, the mathematical formulation of principal component analysis will be described in order to introduce the notation we use. More detailed derivations and their proofs can be found in literature, e.g. [7]. A good description of the application of PCA to multi-energy data acquired from different x-ray spectra is given in [3].

For every chosen energy, $i = 1 \dots k$, we measure an image of p pixels. Each image at a particular energy is represented by the row vector $\mathbf{x}_i$ in which each element x_{ij} contains the intensity value of one pixel where $j = 1 \dots p$. For spectroscopic x-ray images there are typically far fewer energy bins than pixels. That is, $k \ll p$.

In order to simplify further calculations, each pixel's intensity value is centred by subtracting $\langle \mathbf{x}_j \rangle$, the average intensity over all energies. Thus

$$a_{ij} = x_{ij} - \langle \mathbf{x}_j \rangle = x_{ij} - \sum_{i=1}^{k} x_{ij}/k. \qquad (1)$$

For further evaluation, it is beneficial to combine all data into one matrix A where each element is given by a_{ij}. In this representation, the j^{th} column is the centred energy spectrum for the j^{th} pixel, while the i^{th} row of A is the centred image $\mathbf{a}_i$ at the i^{th} energy.

Having established a notation for our dataset, we can now proceed by performing the relevant steps for principal component analysis. The matrix of co-variances for a sample distribution of k centred images is given by equation 2. It gives a measure for the linear dependencies amongst the vectors $\mathbf{a}_i$. Its rank, at most k, gives the number of independent components within the data [7].

$$C = \frac{1}{k-1} A \cdot A^T \qquad (2)$$

It can be shown that the eigenvectors corresponding to the highest eigenvalues of C account for directions of maximal co-variance in the data. When arranged in decreasing order of eigenvalues λ_i, the k eigenvectors $\mathbf{u}_i$ can be used as base vectors to transform the data matrix A onto a new coordinate system given by $\mathbf{z}_1, \dots \mathbf{z}_k$ as expressed in equation 3. This vector space can be referred to as eigenspace.

$$\mathbf{z}_i = \mathbf{u}_i^T A = \sum_{l=1}^{k} \mathbf{u}_{il}\, \mathbf{x}_l \qquad (3)$$

In engineering literature, the eigenvectors $\mathbf{u}_i$ are typically called 'principal components', although Joliffe notes that some authors call $\mathbf{z}_i$ the principal component and refer to $\mathbf{u}_i$ as the vector of loading [7]. We adopt the notation from engineering literature calling $\mathbf{u}_i$ 'principal components'. To avoid confusion, we name the transformed vectors $\mathbf{z}_i$ 'PC-images'. It is worth noting that $\mathbf{z}_i$ have a mean of zero and a variance of λ_i.

3 Application

Radiographs have been acquired with the energy-discriminative x-ray detector Medipix-2 [9]. With almost no dark current and low electronic noise, this detector shows promising features for imaging

applications. Furthermore, each of the 256 by 256 pixels allows the setting of a lower energy threshold above which photons will be counted. Once calibrated to a common threshold amongst pixels, images can be obtained in different energy bins.

3.1 Watch dataset

The first dataset that we used for testing this technique was a series of radiographs of a watch. They were obtained with a conventional x-ray tube in 56 different energy bins with the lower energy threshold of the sensor ranging from 15 to 70 keV with a stepsize of 1 keV.

Figure 1 shows an example of the data. In the right image, where only photons above 68 keV contributed, the gear-wheels behind the cover are better visible than in the 'broad spectrum' image on the left. This is due to the fact that many of the high-energy photons are transmitted through dense areas, while low-energy photons are mostly absorbed. Therefore detecting only high-energy photons in this case leads to an improved contrast and better detail recognition.

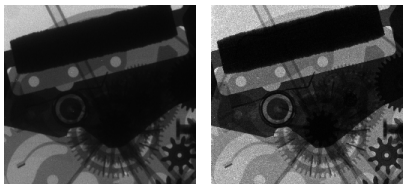

Figure 1: Spectral dataset counting photons above 15 keV (left) and only photons above 68 keV (right)

We now test if the method of PCA, by aiming at maximising the variance in the data and at the same time reducing redundancies, can find such significant features in a large dataset. Following the notation introduced in section 2, the centred data matrix A has a size of $k = 56$ by $p = 256 \times 256 = 2^{16}$ elements, where p gives the number of pixels in each of the k images.

Figure 2 shows a plot of the first three derived principal components. Since the i^{th} element of each principal component corresponds to the i^{th} energy bin it is possible to label the x-axis of this plot with energy. Similarly to the convention in face recognition we have coined the term "eigenspectrum" to refer to these principal components. Any pixel's spectrum can be reconstructed as a linear combination of the eigen-spectra plus the average spectrum. Three non-noisy eigen-spectra could be identified which implies that there are at least three patterns of independent variance within the spectral data.

According to equation 3, the elements of the i^{th} eigenvector give a weighting factor for each centred image which determines its contribution to the i^{th} PC-image $\mathbf{z}_i$.

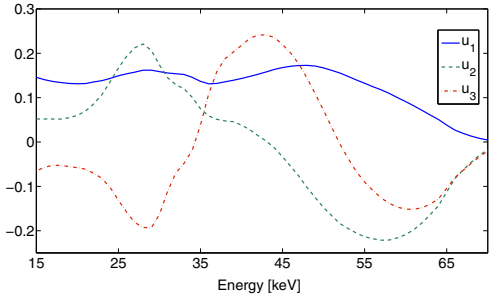

Figure 2: The first three eigenvectors

The 2^{16} elements of the vector $\mathbf{z}_i$ can be reshaped to a 256×256 pixel matrix leading to a representation of the image in the i^{th} eigenspace, as can be seen in figure 3. While these images are closely related to eigenimages commonly found in pattern recognition they are not identical to them. In particular, we have centred our data in the energy domain rather than the spatial domain before analysing the variance.

Figure 3: the first three PC-images

From the k PC-images, only the first three contain usable information. The others show a few features, but mainly noise. So the high redundancies of the original 56 images have been removed leading to a representation of the relevant information in the first few eigenspace images.

The first PC-image shows a much better signal-to-noise ratio than any of the single acquisitions. Looking at the first eigenvector in figure 2 reveals that all original images have almost equally contributed to this component which leads to a reduction of the statistical noise.

Each eigenvector is by definition linearly independent and therefore the PC-images highlight subtle changes in the dataset. Some regions change their intensity values significantly between two PC-images while the other regions remain at similar intensities. This can be seen for example in the darkest region of the first PC-image in comparison to the third PC-image in figure 3. It highlights the effect of beam-hardening which is the main difference that occurs in the whole dataset, as explained above.

3.2 Cluster analysis of watch data

In order to further investigate differences in the transformed variables, we plotted the location of the first three principal components in eigenspace as a scattergram, as can be seen in figure 4. Each point represents one pixel, the location on the x-axis is given by the intensity value of the first PC-image (left image in figure 3), the intensity value of the second and third PC-images give the location on the y-axis and the colour respectively.

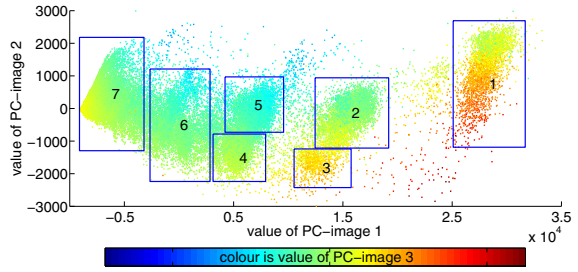

Figure 4: Each datapoint represents one pixels intensity value in PC-image z_1, z_2 and z_3

Most of the variance can be found along the $\mathbf{z}_1$ axis, but the second and third PC-images also contribute to a separation of individual clusters. e.g. separating region '2' from region '3' in figure 4.

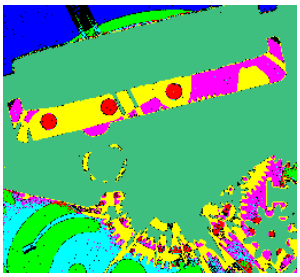

Figure 5: colour map of identified clusters

The identified clusters can each be given a colour and then combined in a single image, as in figure 5. Each cluster represents a group of pixels that have similar spectra. It is a non-trivial result that clustering in eigenspace leads to a grouping of neighbouring pixels. As can be seen from the example of region '2' (green) and '3' (red), areas that could not been distinguished from any single images of the unprocessed dataset, can now be separated. Furthermore, we have demonstrated that at least three patterns of variance are contained within the spectral data.

3.3 PCA of fetal hand

A spectroscopic dataset consisting of seven acquisitions with variation of the lower energy threshold from 4-27 keV of the hand of a 20 week old misscarried fetus were acquired with an x-ray tube with a peak voltage of 35 kV. One can see the bones that are not yet fully developed.

The principal component analysis has been applied in the same way as described above for the watch dataset. This time, the first three PC-images have been combined into one RGB image, figure 6. In this image $\mathbf{z}_1$, $\mathbf{z}_2$ and $\mathbf{z}_3$ are used for red, green, and blue respectively.

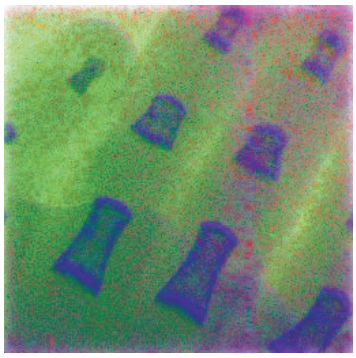

Figure 6: fetal hand with the RGB colour map given by the first three PC-images $\mathbf{z}_1$, $\mathbf{z}_2$ and $\mathbf{z}_3$

The representation of the dataset as one RGB image leads to advanced contrast that combines all relevant information of the dataset. It can clearly been seen that the bone (blue) has a different energy respone than the soft tissue (green). The red area at the top right of the image is thought to relate to variations in the detector response.

4 Discussion

The technique presented so far has only been applied to radiographs, therefore it can not be used to uniquely identify materials, as their thickness is unknown. Our group has built a computed tomographic scanner equipped with the Medipix detector [1] to further investigate the application of PCA to reconstructed CT slices. Current work is focused on multi contrast agent recognition of spectroscopic CT data. The next version of the Medipix detector [10] will allow simultaneous measurement in up to eight energy bins. Simultaneous measurements will give identical Poissonian noise across all energy bins, which will improve PCA.

Another area of investigation is to study the possibility of identifying principal components (eigenspectra) from a small data set and then using these as basis vector to describe spectra from a larger data set. Such a method is more akin to the application of PCA to face recognition where principal components are derived using a small training set of face images. New faces, not from the training set, are then described as a linear combination of the principal components. This method would be particularly help for large 3D data sets, where

applying PCA to the entire dataset would be computationally expensive.

5 Conclusion

First applications of the principal component analysis to Medipix-2 images included a non-destructive testing sample (watch) and a clinical subject (fetal hand).

With an original dataset of images from 56 energy bins (watch dataset), redundancies were expected. The PCA method leads to a representation of the data in three components that span a majority of the variance. PC-images show improved contrast and highlight areas where changes in the original dataset occur. The maximisation of the variance is even more obvious, when performing a cluster analysis, where regions of similar intensity in the original dataset can now be separated from each other.

RGB remapping techniques can be used to enhance contrast further and to display the relevant information in only one image. Overall, the principal component analyis is a promising method for enhancing contrast in the spectral information. One major benefit of the method is that it requires no a priori knowledge of how the spectral information varies within the data.

6 Acknowledgements

We would like to thank the whole Medipix-Team at CERN for their help in acquiring the dataset. Nigel Anderson (Canterbury District Health Board), Nanette Schleich, Juergen Meyer and Richard Watts (University of Canterbury) have greatly contributed to the project. Furthermore, we'd like to thank the whole Medipix collaboration for their support. The author would like to thank Simone Dunkl for useful suggestions and proofreading.

References

[1] J. S. Butzer, A. P. Butler, N. J. Cook, P. H. Butler, F. Ross, N. Schleich, J. Selkirk, R. Watts, J. Meyer, N. Scott, P. J. Bones, D. van Leeuwen, S. Hemmingsen, T. P. Melzer, and N. Anderson, "Mars: A 3d spectroscopic x-ray imaging device based on medipix," *Accepted for presentation at the IEEE Medical Imaging Conference*, Oct. 2008.

[2] T. Asaga, S. Chiyasu, S. Mastuda, H. Mastuura, H. Kato, M. Ishida, and T. Komaki, "Breast imaging: Dual-energy projection radiography," *Radiology*, vol. 164, pp. 869–870, 1987.

[3] A. R. Kalukin, M. Van Geet, and R. Swennen, "Principal component analysis of multienergy x-ray computed tomography of mineral samples," *IEEE Transactions on Nuclear Science*, vol. 47, no. 5, pp. 1729–1736, Oct. 2000.

[4] M. Firsching, T. Michel, and G. Anton, "First measurements of material reconstruction in x-ray imaging with the medipix2 detector," *Proceedings Nuclear Science Symposium, Honolulu*, pp. 2736–2740, 2007.

[5] E. Roessl and R. Proksa, "K-edge imaging in x-ray computed tomography using multi-bin photon," *Physics in Medicine & Biology*, vol. 52, pp. 4679–4696, 2007.

[6] J. Giersch, D. Niederloeher, and G. Anton, "The influence of energy weighting on x-ray imaging quality," *Nuclear Instruments and Methods in Physics Research A*, vol. 531, pp. 68–74, 2004.

[7] I. Jollife, *Principal Component Analysis*, 2nd ed. Springer-Verlag, 2002.

[8] P. J. Bones, A. P. H. Butler, and M. Hurrell, "Enhancement of chest radiographs using eigenimage processing," *In Image Reconstruction from Incomplete Data IV, Proceedings SPIE*, vol. 6316, p. 12pp, 2006.

[9] X. Llopart, M. Campbell, R. Dinapoli, D. San Segundo, and E. Pernigotti, "Medipix2: A 64-k pixel readout chip with 55-μm square elements working in single photon counting mode," *Nuclear Science, IEEE Transactions on*, vol. 49, no. 5, pp. 2279–2283, Oct 2002.

[10] R. Ballabriga, M. Campbell, E. Heijne, X. Llopart, and L. Tlustos, "The medipix3 prototype, a pixel readout chip working in single photon counting mode with improved spectrometric performance," *Nuclear Science Symposium Conference Record, 2006. IEEE*, vol. 6, pp. 3557–3561, 2006.