JOURNAL OF CHEMOMETRICS
J. Chemometrics 2003; **17**: 400–412
Published online in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/cem.817

Analysis of video images from a gas–liquid transfer experiment: a comparison of PCA and PARAFAC for multivariate image analysis

Stephen P. Gurden, Euler M. Lage, Cristiano G. de Faria, Inés Joekes and Márcia M. C. Ferreira*
Instituto de Química, Universidade Estadual de Campinas (UNICAMP), CP 6541, 13084-971 Campinas, SP, Brazil

Received 7 December 2002; Revised 22 February 2003; Accepted 7 May 2003

The use of chemical imaging is a developing area which has potential benefits for chemical systems where spatial distribution is important. Examples include processes in which homogeneity is critical, such as polymerizations, pharmaceutical powder blending and surface catalysis, and dynamic processes such as the study of diffusion rates or the transport of environmental pollutants. Whilst single images can be used to determine chemical distribution patterns at a given point in time, dynamic processes can be studied using a sequence of images measured at regular time intervals, i.e. a movie. Multivariate modeling of image data can help to provide insight into the important chemical factors present. However, many issues of how best to apply these models remain unclear, especially when the data arrays involved have four or five different dimensions (height, width, wavelength, time, experiment number, etc.). In this paper we describe the analysis of video images recorded during an experiment to investigate the uptake of CO_2 across a free air–water interface. The use of PCA and PARAFAC for the analysis of both single images and movies is described and some differences and similarities are highlighted. Some other image transformation techniques, such as chemical mapping and histograms, are found to be useful both for pretreatment of the raw data and for dimensionality reduction of the data arrays prior to further modeling.

KEYWORDS: multivariate image analysis; PARAFAC; gas–liquid transfer

1. INTRODUCTION

Chemical imaging covers a wide range of measurement techniques all involved with the collection of spatially resolved measurements—images—which can be related to the properties of the chemical system being measured. The development of instruments capable of high resolution in space, time and wavelength is continuing [1–4] and chemical imaging looks set to become a widespread tool for both laboratory and process chemistry applications. Temporal imaging refers to the situation where images are measured on the same system at regular intervals in time. This can be used to provide snapshots of the system in real time, useful for monitoring and/or control of heterogeneous processes, [5] or for recording movies of dynamic processes in order to study reaction kinetics or transport rates.

*Correspondence to: M. M. C. Ferreira, Instituto de Química, Universidade Estadual de Campinas (UNICAMP), CP 6541, 13084-971 Campinas, SP, Brazil.
E-mail: marcia@iqm.unicamp.br
Contract/grant sponsors: State of São Paulo Research Foundation (FAPESP); Brazilian National Research Council (CNPq).

The simplest chemical images are univariate, gray-scale images, such as 2D images from an electron microscope, where each image is a matrix with dimensions *height* × *width*. . Multivariate images are becoming more common, however, with each image being a three-way array with dimensions *height* × *width* × *wavelength*. A color image is a type of multivariate image in which light is measured at three wavelengths corresponding to red, green and blue light, i.e. *height* × *width* × 3. Other types of multivariate images can be produced by spectrophotometric cameras [2] or scanning electrochemical microscopes [4]. A temporal sequence of multivariate images forms a movie, a four-way array with dimensions *height* × *width* × *wavelength* × *time*. As the dimensionality of these data arrays increases, so the amount of data generated becomes huge. For example, a 600 × 800 × 3 color image measured every second generates 86 MB per minute or 5.2 GB per hour. In the case of spectroscopic imaging, this can be increased 100-fold. Clearly, there is a need for multivariate analysis methods capable of extracting the useful information from these huge data arrays, and some applications of chemometrics to image data have already appeared in the last 15 years. Several of these have focused on the use of PCA

and PLS to analyze single multivariate images taken from examples outside of chemistry [6–9], or on the texture analysis of images of powders [10] and food samples [11]. More recently, chemical applications involving the simultaneous analysis of multiple images from secondary ion mass spectrometry (SIMS) [12], X-ray photoelectron spectroscopy (XPS) [13] and fluorescence microscopy [14] have been described.

The data used in this paper come from an investigation into a gas–liquid transfer process in which CO_2 is dissolved by water under controlled conditions [15]. The presence of a pH indicator causes a color change as CO_2 is transported through the system, and this process is recorded using a color video camera. Whilst the aim of the experiment is to determine the influence of temperature and salinity on CO_2 uptake, the data are used here to investigate aspects of multivariate image analysis in general.

Chemical imaging commonly results in three-way, four-way or even higher-order data arrays. When modeling these data, a number of possibilities exist. For example, it is possible to model a movie in its entirety or frame by frame. Different ways of arranging the data array prior to modeling make a fundamental difference to the model obtained [11,16], as can different ways of centering and scaling the data [7,17]. Whilst the benefit of using multivariate analysis to extract important information is apparent, many aspects of how best to apply the models are still unclear. In this paper we focus on the use of two multivariate models already well known in chemistry—PCA and PARAFAC—for the analysis of both single images and movies. Some characteristics of these models, with respect to their application to chemical images, are contrasted and their relative merits are discussed.

In addition to multivariate models such as PCA and PARAFAC, there also exists a huge range of standard image processing and analysis techniques which have been used in the analysis of images within other fields for many years. Prior to statistical analysis of the data, it is necessary to arrange the data in regular, congruent arrays. Some standard image processing techniques are used here in order to prepare the raw image data. The use of mapping and histograms for the analysis of single images is also demonstrated. Although these tools are valuable in their own right for helping to understand the data, they can also be used for reducing the dimensionality of the data, whilst retaining chemical information, prior to further statistical modeling.

The rest of the paper is laid out as follows. In Section 2, PCA and PARAFAC are introduced in the context of multivariate image analysis. In Section 3 the gas–liquid transfer experiment is described, along with some computational issues. In Section 4 the data preparation steps of digitization and reconciliation are described. In Section 5, single multivariate images are analyzed using PCA and PARAFAC. In Section 6, some other single-image analysis techniques—mapping, histograms and mean profiles—are described. In Section 7 the analysis of movies using PCA and PARAFAC is described. Finally, conclusions are given in Section 8.

2. THEORY

The use of PCA for the analysis of single, multivariate images is already well known [6,7,17]. The multivariate image with dimensions *height* × *width* × *wavelength* is first 'unfolded' [18] to give a matrix with dimensions *height·width* × *wavelength*. In the case of PCA a bilinear decomposition is then carried out:

$$\mathbf{X} = \sum_{r=1}^{R} \mathbf{t}_r \mathbf{p}_r^{\mathrm{T}} + \mathbf{E} = \mathbf{T}\mathbf{P}^{\mathrm{T}} + \mathbf{E} \tag{1}$$

where R is the number of principal components (PCs) used to describe $\mathbf{X}$ and is typically small (i.e. less than five). The scores $\mathbf{T}$ and loadings $\mathbf{P}$ summarize the important information in the data, whilst discarding the redundant information such as noise. In order to view the scores as a gray-scale image, it is necessary to 'refold' each of the R score vectors, i.e. $\mathbf{t}_r$ (*height* · *width* × 1) is rearranged to $\mathbf{T}_r$ (*height* × *width*). In a similar way it is possible to reconstruct a multivariate image for the PCA residuals. Note that whilst this approach can equally well be used to model a sequence of univariate images measured in time (i.e. *height* × *width* × *time*), extending the model to multivariate movies is not possible, because a meaningful rearrangement of the four-way array with dimensions *height* × *width* × *wavelength* × *time* into a classical two-way matrix does not exist [16].

One important consequence of the above approach is that each pixel in the image is treated as an independent object. This means that whilst between-image correlation [11] (i.e. structure present in the wavelength or time dimension) is modeled, within-image correlation (i.e. correlation between the height and width dimensions) is not. This may be a reasonable approach; real-world images are usually fairly complex and do not necessarily lend themselves well to low-dimensional, linear modeling. By treating each pixel as an independent object, full spatial information is retained and it is left to the user to find meaningful structure in the resultant score images. One possible disadvantage of the use of PCA in this context is that an orthogonal basis for the scores and loadings is usually used in order to identify the model, leading to score images which are 'maximally different' [19]. Real chemical features are rarely orthogonal and so this purely mathematical constraint may make interpretation of the model more difficult.

Multiway models recently introduced into chemistry, such as parallel factor analysis (PARAFAC) [20–22], can be considered as generalizations of the classical, two-way approaches to data of order higher than two. For a three-way array $\underline{\mathbf{X}}$ ($M \times N \times P$), PARAFAC gives a trilinear decomposition which can be expressed as follows:

$$x_{mnp} = \sum_{r=1}^{R} a_{mr} b_{nr} c_{pr} + e_{mnp} \tag{2}$$

where $\underline{\mathbf{X}}$ is decomposed into three sets of loadings, $\mathbf{A}$ ($M \times R$), $\mathbf{B}$ ($N \times R$) and $\mathbf{C}$ ($P \times R$). Arrays of higher order (e.g. four-way, five-way, etc.) are modeled by simply adding extra sets of loadings (i.e. $\mathbf{D}$, $\mathbf{E}$, etc.). Note that a differentiation between score and loading matrices is not made for multiway models; all the reduced-dimension matrices are referred to as loadings. Furthermore, unlike in PCA, PARAFAC components are calculated simultaneously, are not orthogonal and therefore have no particular order in terms of variation explained. As for PCA, it is possible to reconstruct

gray-scale images for each component, in this case by plotting $\mathbf{a}_r\mathbf{b}_r^{\mathrm{T}}$, where $\mathbf{a}_r$ ($height \times 1$) and $\mathbf{b}_r$ ($width \times 1$) are the loadings describing the spatial dimensions for the rth component.

When applied to single, multivariate images, PARAFAC offers an alternative approach to that of PCA, in which the correlation between the height and width dimensions *is* modeled, in addition to the between-image correlation. A disadvantage of this approach is that only structurally simple features are capable of being modeled, i.e. shapes which can be described by simple linear decompositions. Some advantages, however, are the vastly reduced number of model parameters used and the well-known uniqueness property of PARAFAC. The latter has already been shown to yield much more interpretable models in some cases [21,23], as it is not necessary to use orthogonality or maximum variance constraints in order to identify the model. Another advantage of PARAFAC is its increased flexibility in terms of handling multiway arrays, making it suitable for modeling multivariate movies and other higher-order arrays consisting of multiple images.

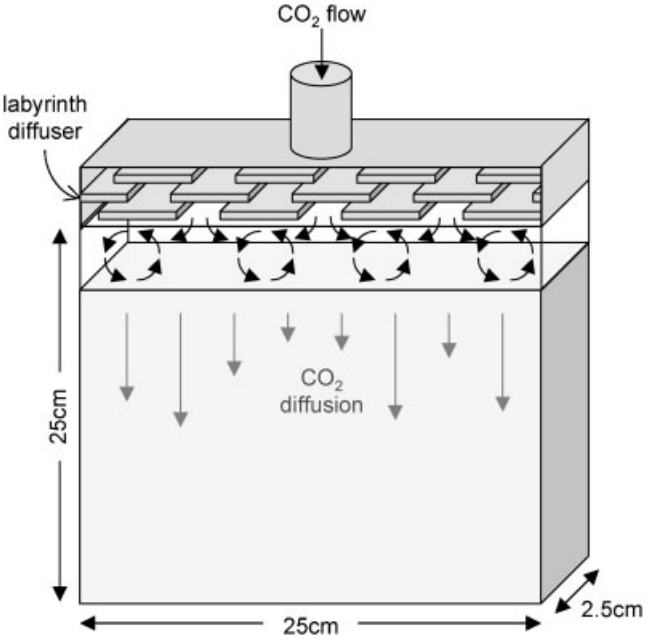

Figure 1. Diffusion tank used to investigate gas–liquid transfer.

3. EXPERIMENTAL

The exchange of CO_2 between the atmosphere and the ocean is of great environmental importance in terms of the global carbon-cycling system [24,25], and new techniques are being sought to improve understanding of this complex process [26]. Despite a large number of experimental investigations both in the field and in the laboratory [27–29], detailed knowledge of the gas–liquid transfer process across a free air–water interface is not currently available, one reason for this being the number and variety of physical and chemical factors which influence the process [30–32]. Air–sea exchange has commonly been modeled as a function of sea surface temperature and wind speed (related to turbulence), but other factors such as salinity, biomatter content, bubbles, surfactants and boundary layer stability are also known to contribute to the gas flux.

The data example used in this paper comes from an experiment in which the uptake of CO_2 by water under controlled temperature and salinity conditions and in the absence of turbulence is investigated [15]. A glass diffusion tank of size $25 \times 25 \times 2.5\,\mathrm{cm}^3$, represented in Figure 1, holds a saline solution into which a mixture of the pH indicator methyl red and the color enhancer methylene blue has been added. Sitting on top of the diffusion tank is a labyrinth diffuser [15], a purpose-developed device used to introduce CO_2 at a constant partial pressure across the surface of the solution. Surrounding the diffusion tank is a water bath used to control the temperature of the solution. A Sony CCD-TR700 VHS video recorder is situated approximately 4 m in front of the diffusion tank. When the gas valve is opened, CO_2 enters the airspace at the top of the diffusion tank and begins to dissolve through the water, changing the color of the solution from green (pH 6.2) to violet (pH 4.2) owing to the formation of carbonic acid, thus allowing the CO_2 uptake to be tracked visually. CO_2 uptake is fast at the boundary layer of the water, and after 1 min an approximately 3 mm layer of dissolved CO_2 is seen at the surface. In the absence of turbulence, diffusion through the rest of the tank is much slower and generally occurs heterogeneously and in chaotic patterns depending, only in part, upon the temperature and salinity of the water. Recording continues until the tank is completely violet, this being a subjective judgment. The experiment is carried out in a dark laboratory, with illumination being provided by a lamp positioned behind the water bath and various precautions being taken to avoid undesired light entering the video camera [15].

Although a number of experimental runs measured at different temperatures and salinities were available, in this paper we will concentrate on the analysis of one experimental run for illustrative purposes. Experimental run 1 was measured at a temperature of 25°C and using distilled water in the diffusion tank. The duration of the experiment was 51 min. Analysis of multiple runs may be described in a future paper.

3.1. Computations

All computations were performed using MATLAB Version 6.1 (The Mathworks, Natick, MA, USA) on a 1500 MHz PC with 512 KB of RAM. Some of the image processing functions were performed using the MATLAB Image Processing Toolbox Version 3.1. PARAFAC models were built using The N-way Toolbox for MATLAB Version 2.0 [33,34].

One of the major problems with modern chemical imaging applications is the large memory requirement caused by the size of the data arrays. Whilst a typical movie described here used only 5 MB of hard disk space, this equaled around 86 MB of computer RAM when stored as unsigned integers (i.e. integers from 1 to 255). However, owing to problems such as rounding errors and overflow, it is not practical to work with integers for computational purposes. After the reconciliation step described below (in which some regions of the images are discarded), the data were converted to double-precision numbers for all subsequent analyses. This required manipulation of multivariate image movies of

around 350 MB and univariate image movies of around 115 MB. In some cases the MATLAB code must be carefully written in order to optimize for memory rather than speed. Examples of this would be replacing rather than creating new arrays during preprocessing steps and performing even simple mathematical operations frame by frame rather than simultaneously on the entire movie array. Computation times were not found to be excessive, however: a PCA on a movie array took around 15 s and a PARAFAC analysis up to 30 min. Although not used here, data compression techniques, e.g. wavelets [35], could be used to reduce memory requirement and computation time.

4. DATA PREPARATION

The experimental data were available as a movie stored on video tape. In order to apply statistical methods to the data, it was necessary to transform the raw data into a form suitable for computational analysis. This involved digitization and reconciliation.

4.1. Digitization

Digitization is the procedure by which the information on the magnetic tape is transformed into a series of numbers capable of being stored by a computer. This was carried out using a Hauppauge WinTV video card and software (Hauppauge Computer Works, Hauppauge, NY, USA). As the movie was played, frames were captured and saved as 'jpg' files at a rate of one frame per minute. Whilst a higher time resolution was possible, current limitations of computer memory during the subsequent computational analyses prohibit this at present. Each frame had a spatial resolution of 768×1024, with three wavelengths (red/green/blue) per pixel. This means that an experimental run lasting 51 min produced a movie array of size $768 \times 1024 \times 3 \times 52$, where the first frame is measured at $t = 0$ [15].

4.2. Reconciliation

Reconciliation is the procedure by which the irregular raw data images are transformed in such a way that they are ready for computational analysis. This involves removing irrelevant information from the images and rotating and resizing the images so they can by stored in regular data arrays. The frames of a movie array should be congruent; that is, in each image, corresponding pixels should refer to the same point in space.

A typical raw data image for this experiment is shown in Figure 2. The diffusion tank is seen in the middle of the image, surrounded by parts of the experimental set-up (gas diffuser, water bath, light screen, etc.). As we are interested only in the uptake process occurring within the diffusion tank, it was necessary to select only this part of the image, discarding the irrelevant information. Prior to this, however, it was necessary to correct for image rotation, caused by the camera not always being entirely straight in relation to the diffusion tank (the camera was removed and replaced in between experimental runs).

To correct for image rotation, the surface of the water in the diffusion tank was used as a reference line. An image showing the edges in the raw image was generated using the

Figure 2. Raw data image from experimental run 4, frame 1 [14].

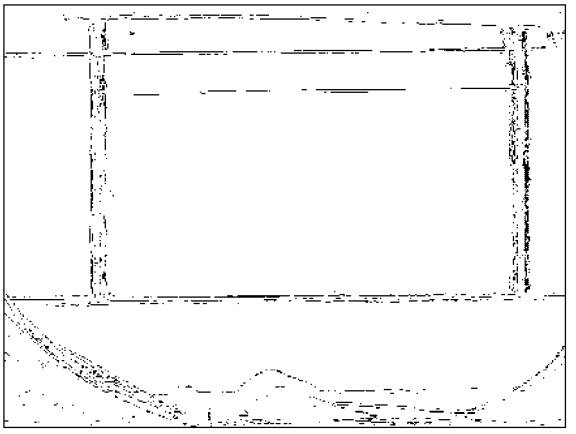

Figure 3. Sobel edge detection of the image shown in Figure 2.

Sobel method [36] with a threshold level of 0.015, as shown in Figure 3. A rectangular area around the surface of the water in the diffusion tank was then selected and the Radon transform [37] was applied to this subimage. The Radon transform determines the angles of the principal edges within the image by projecting the image intensities onto a radial line oriented within a given range, e.g. $\pm 2^\circ$. In this case the angle at which a maximum projected intensity is found corresponds to the angle of deviation of the line defined by the water's surface from the top border of the image, found to be 0.525° for Figure 2. This angle was then used to rotate all the images within the same experimental run, it being assumed that the first frame is representative of the entire movie.

After rotation the next step was to remove the background region not useful for analysis of the uptake process, a procedure known as cropping. Again the first frame was used to represent subsequent frames within the same movie. The cropping was done manually, the aim being to include as much of the diffusion tank as possible whilst excluding irregularities at the sides and bottom of the tank.

Finally, after rotation and cropping, it was found that movies from different experimental runs were left with different sizes along the height and width dimensions. In order to standardize the movies, each frame was resized using bicubic interpolation to have spatial resolution 300×600. Thus, after reconciliation, experimental run 1 gave a data array of size $300 \times 600 \times 3 \times 52$.

Four frames from experimental run 1 are shown in Figure 4. The entry of CO_2 into the tank from a thin layer close to the water's surface is seen after 2 min (Figure 4(a)). After 4 min (Figure 4(b)), three distinct mushroom-shaped regions can be seen which, after 9 min (Figure 4(c)), start to fill the entire tank with the exception of a space on the right-hand side. The diffusion of CO_2 throughout the tank continues and, after 30 min (Figure 4(d)), CO_2 is spread homogeneously throughout the tank.

5. ANALYSIS OF SINGLE IMAGES

In this section the use of PCA is compared with PARAFAC as a tool for the analysis of single, multivariate images. Frame 5 from experimental run 1, shown in Figure 4(b), is used as an example image for demonstrating the results of the analyses. This frame is held in a data array with dimensions $300 \times 600 \times 3$.

5.1. Principal component analysis

A PCA was performed on the frame, first unfolding the array into a $180\,000 \times 3$ matrix, and two components, explaining 99.99% of the variation in the data, were retained. No centering or variable rescaling was used, although the preprocessing of image data is discussed later on in Section 7.1. As no mean centering of the data was performed prior to PCA, a very high percentage of the data is described by the first PC alone [19]. This PC (not shown here) describes the average intensity image, similar to a gray-scale version of the original image. The reconstructed score image for the second PC is shown in Figure 5, where the mushroom-shaped diffusion of CO_2 into the saline solution and the thin layer of high CO_2 concentration near the surface are shown in high contrast. This can be understood by looking at the PCA loadings, shown in Figure 6. The loadings for both PCs, but especially PC 2, clearly distinguish between green and violet (red/blue), these being the two colors present in the diffusion tank. PC 2 shows more clearly than the original image the contrast between the high- and low-pH areas. The model residuals (not shown here) are very noisy and distinguish between red and cyan (blue/green).

5.2. PARAFAC

A PARAFAC was performed directly on the same multivariate image, with no unfolding being necessary. The core consistency diagnostic [38] indicated that two PCs were optimal. These explained 99.87% of the variation in the data—still very high, but lower than for PCA. The PARAFAC loadings describing the height, width and wavelength dimensions are shown in Figure 7. The loadings describing the wavelength dimension, **C**, are almost identical to those found for PCA; both components, but especially the second (broken line), distinguish between the colors green and violet.

PARAFAC gives two sets of loadings, **A** and **B**, which describe the height and width dimensions respectively. Although it is possible to plot these loadings individually, as in Figure 7, it is also possible to reconstruct images from the PARAFAC loadings in a similar way as for PCA, but in this case by plotting $\mathbf{a}_r\mathbf{b}_r^{\mathrm{T}}$. The first PARAFAC component image, shown in Figure 8(a), describes an area of high light intensity centered at the top center of the image, with the bottom left and right corners of the image being darker. This component represents a background effect present throughout this movie caused by the inability of the light source (situated behind the diffusion tank) to provide uniform illumination across the image space. Although not clearly apparent from the original image (Figure 4(b)), this background effect can be seen more clearly in some of the other frames from this and other experimental runs.

The second PARAFAC component is shown in Figure 8(b) and describes the CO_2 uptake, in particular the thin layer of high concentration at the top of the tank, the wider region below this layer and the three vertical bands where CO_2 is moving towards the bottom of the tank. The image is rather abstract, owing to the limited ability of low-dimensional linear models to describe complex shapes, and features such as the mushroom shape of the bands are lost (cf. Figure 5).

5.3. Discussion

In finding structure in the wavelength dimension, both PCA and PARAFAC yield very similar results. Both models find the main feature present, i.e. the contrast between the high-pH (green) and low-pH (violet) factors present in the image. Although the second PCA component provides a useful enhancement of the original images, in general, the advantages of PCA are more apparent when images with a higher number of wavelengths per pixel are being analyzed, such as spectrophotometric images, where the advantages of a compression of the wavelength dimension are greater. In the case of color images, which have only three wavelengths per pixel, it is possible to augment the spectral range by adding transformed variables [9], but this was not thought to be useful here.

The way in which PCA and PARAFAC model the spatial features of the image is very different. By treating each pixel as an independent element, ignoring contextual information within the image, PCA effectively focuses on finding structure in the wavelength rather than in the spatial dimensions. This means that whilst PCA is good at highlighting regions of the image with different spectral features, it does not discriminate between different structural features in the image, this being left to the user. The PARAFAC model looks for structure in both the wavelength and the spatial modes. This means that far fewer parameters are used by the model (PARAFAC uses 903 parameters per component; PCA uses 180 003), but that less variation in the data is explained. PARAFAC is able to find only structurally simple factors in the data: either smooth background effects or simplified representations of physicochemical factors such as the bands seen in Figure 8(b). However, the ability of the PARAFAC model to identify factors without forcing the images to be orthogonal was seen to uncover an important aspect of the data not exposed by PCA, i.e. the background illumination effect.

6. OTHER SINGLE-IMAGE TRANSFORMATIONS

Multivariate modeling techniques such as PCA and PARAFAC reduce the dimensionality of large, collinear data sets.

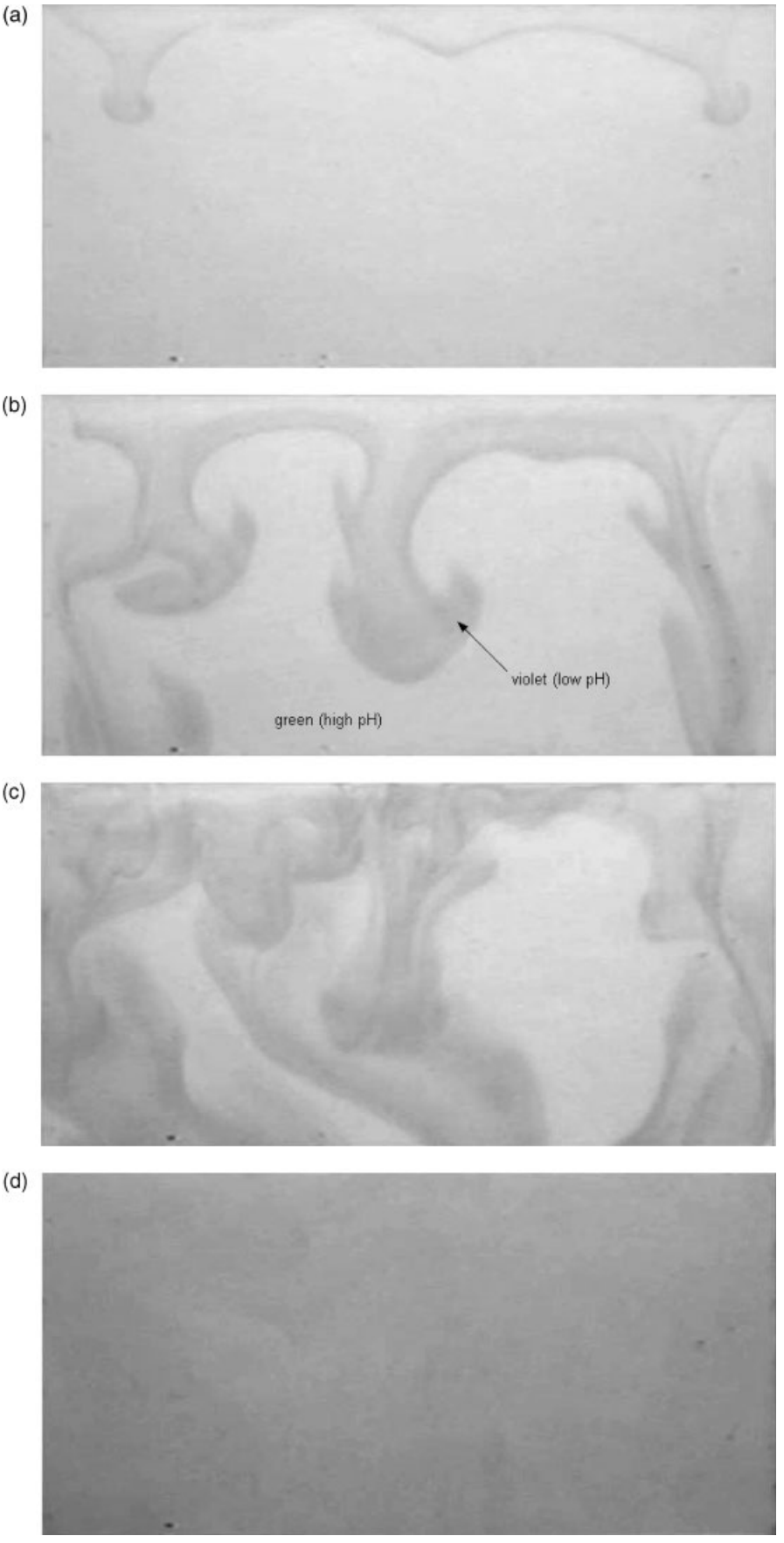

Figure 4. Frames (a) 3, (b) 5, (c) 8 and (d) 31 of experimental run 1 [14].

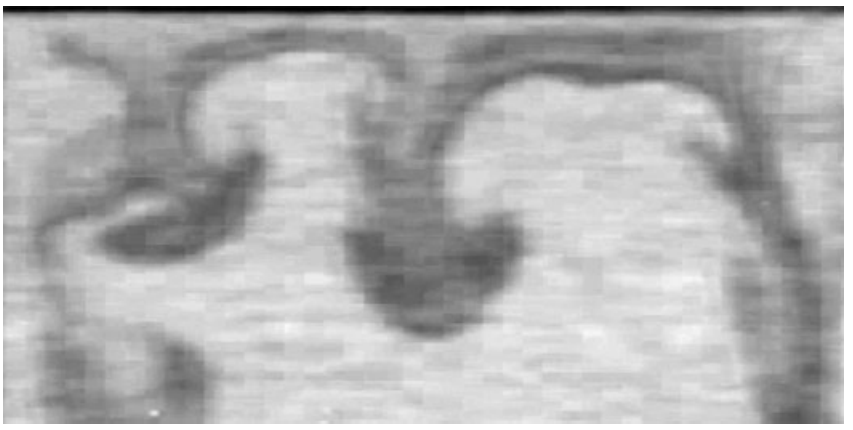

Figure 5. Reconstructed score image for PC 2 (0.15%) from a PCA on the image shown in Figure 4(b).

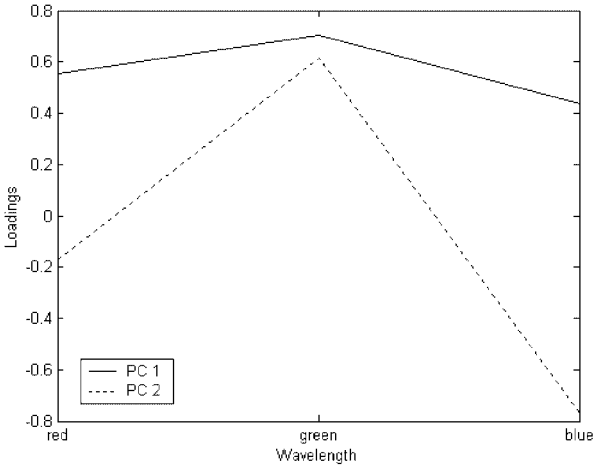

Figure 6. Loadings from a PCA on the image shown in Figure 4(b).

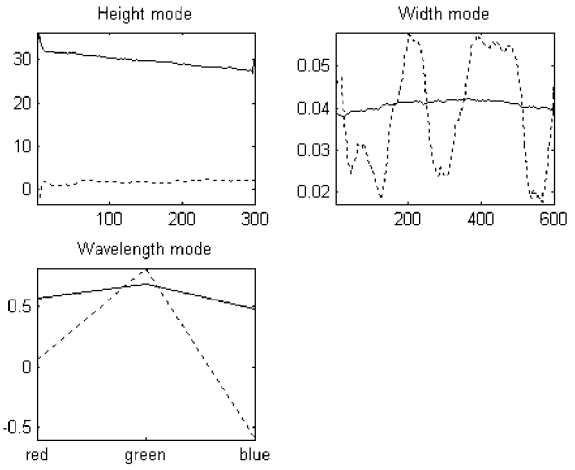

Figure 7. Loadings from a PARAFAC on the image shown in Figure 4(b).

As demonstrated for color images, PCA is not very efficient as it reduces only the dimensionality of the wavelength mode, which is already low. PARAFAC reduces the dimensionality of both the wavelength and spatial modes, but it is only able to describe structurally simple features. Some other single-image transformations which were found to be useful are now described.

6.1. Mapping

One important use of chemical image analysis is the ability to build maps for chemical properties of specific interest. The use of multivariate calibration is already well established in chemistry as a method for determining 'difficult-to-measure' chemical responses, such as compound concentration, using 'easy-to-measure' information, such as near-infrared spectra. This approach can be applied within chemical imaging to give information about the distribution of a chemical substance (or a chemical property) within an inhomogeneous region. Examples may be the distribution of an active compound within the filler during a powder blending process; moisture content of a food sample; or the presence of a fouling substance on a surface catalyst surface. If a good calibration model can be made between the chemical response of interest and the spectral information provided by the imager, then it is relatively simple to construct a univariate (or, in the case of multiple responses, multivariate) map from a multivariate image.

In the example of the CO_2–water transfer experiment, we are interested in the distribution of dissolved CO_2 within the diffusion tank, this being directly related to pH which, in turn, is related to the color of the solution. Thus the idea is to transform the multivariate color images into univariate pH maps [15] (C. G. de Faria and E. M. Lage, in preparation). For each set of experimental runs a separate data set was recorded consisting of images of the same water/indicator solution at known pHs. For the data considered here, seven color/pH calibration images were available. A one-component partial least squares (PLS) [39] model was sufficient to describe 98.94% of the color information and 95.49% of the pH response information (i.e. $R^2 = 0.9549$). In order to correct for intensity differences due to the uneven background illumination (see previous section), pixelwise normalization [17] was used prior to PLS modeling, in which each pixel is scaled to unit length:

$$x^*_{mnp} = \frac{x_{mnp}}{\sqrt{\sum_{p=1}^{3} x^2_{mnp}}} \tag{3}$$

The PLS model was then used to transform the color image, pixel by pixel, into the pH map shown in Figure 9.

It is found that the pH map gives a more accurate representation of the differences in pH throughout the tank

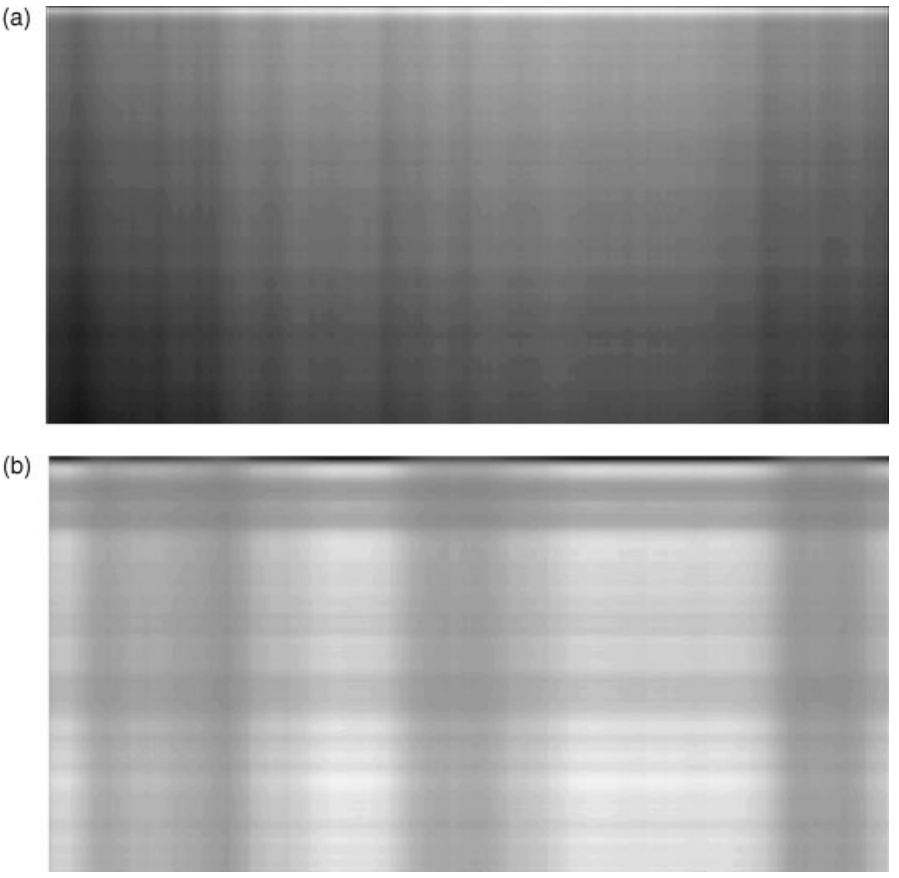

Figure 8. Reconstructed images for (a) PC 1 and (b) PC 2 from a PARAFAC on the image shown in Figure 4(b).

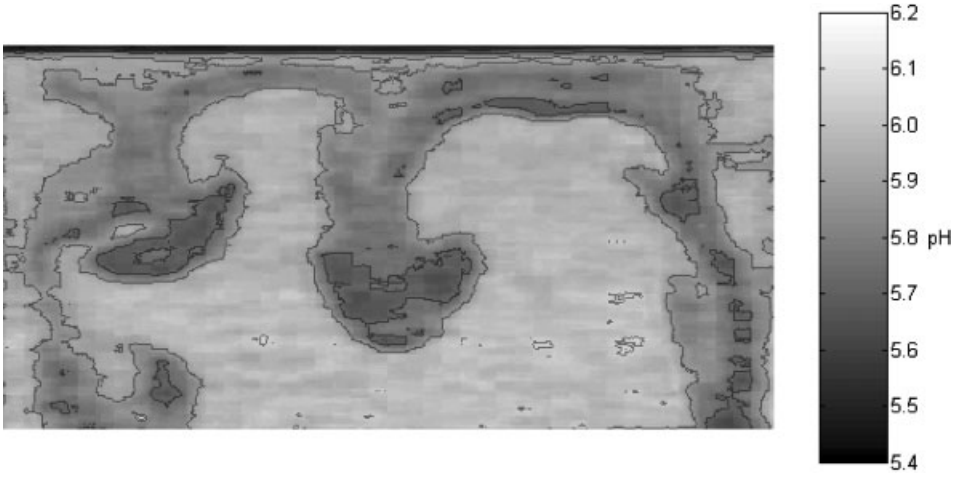

Figure 9. pH map of the image shown in Figure 4(b).

than is possible by a simple visual inspection of the original color image [14] (C. G. de Faria and E. M. Lage, in preparation). The pH map is also more efficient than the PCA representation given in the previous section, because it is a univariate representation of the one characteristic of specific interest: pH. Note that, like the PCA method used previously, mapping reduces the wavelength mode of a multivariate image—from three to one in the case studied here—whilst retaining the full spatial detail.

6.2. Histograms

In classical image analysis, histograms are often used as a measure of the heterogeneity of a univariate image. Histograms of the pH map shown in Figure 9 and another taken from the same experimental run, but at a later point in time, are shown in Figure 10. The first histogram ($t = 4$ min) represents the heterogeneous distribution as the CO_2 begins to penetrate the tank. The second histogram ($t = 30$ min) represents the more homogeneous distribution as the CO_2 almost reaches an equilibrium between air and water. The most important feature of an image histogram is that spatial information is discarded. This means that information about how much CO_2 has been absorbed is present, but that information about the distribution of the CO_2 within the diffusion tank has been lost. This allows a large compression of the data—from a 300×600 image matrix into a 64×1

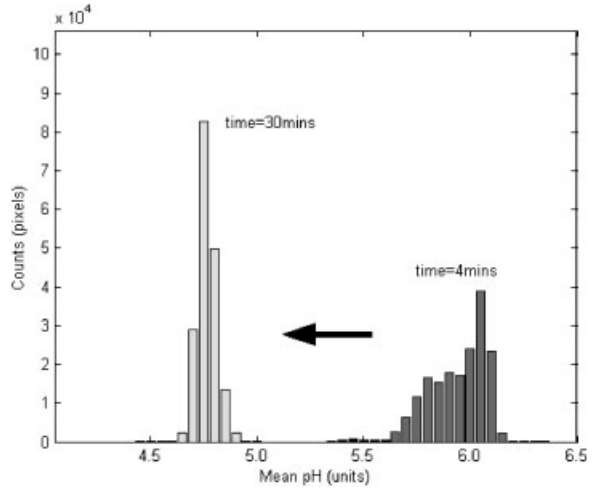

Figure 10. Histograms taken from the pH maps for experimental run 1, frames 5 and 25.

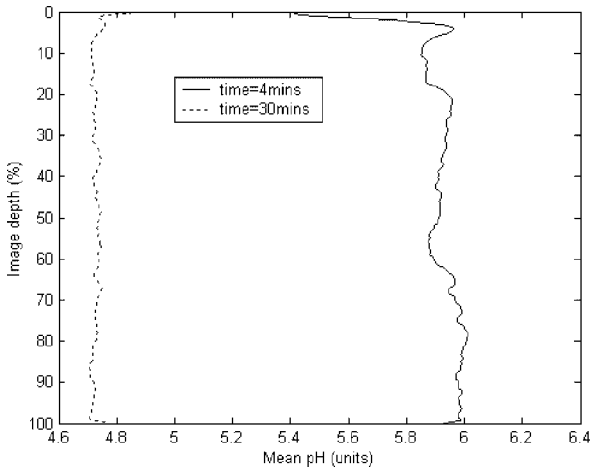

Figure 11. Mean intensity profiles taken from the pH maps for experimental run 1, frames 5 and 25.

vector—and could be used to increase the computation speed for analyses where spatial information is not important.

6.3. Mean profiles

Another simple transformation which may be useful when the top/bottom or left/right orientations have a chemical significance is the use of mean profiles calculated by taking the average across either the height or width dimension of a univariate image. In the case of the pH maps used here, it is possible to calculate a mean intensity profile which gives information on the average concentration of CO_2 at a given depth in the diffusion tank. These profiles are shown in Figure 11, for the top half of the tank, for the frames at 4 and 30 min. It can be seen that the layer near the surface of high CO_2 concentration which is present after 4 min actually disappears as the uptake process proceeds, and that after 30 min there is actually a thin layer of lower concentration near the surface.

7. ANALYSIS OF MOVIES

Up to now we have focused on the analysis of single images—frames of a movie. However, for dynamic processes, in which rates and patterns of chemical change are of interest, it is necessary to consider the movie as a whole. Given that the resolution of the time mode is sufficiently high, we would expect to find autocorrelation between the images which can be modeled and provide information on the most important changes occurring in time.

A number of possibilities exist for modeling multivariate movie arrays. As with the single images discussed in Section 5, a choice can be made as to whether to model the spatial dimensions explicitly, using a quadrilinear PARAFAC model, or whether to first unfold the spatial dimensions and build a trilinear PARAFAC model on the *height* · *width* × *wavelength* × *time* array. In cases where images have different dimensions (uncommon for movie arrays, but possible when comparing multiple images in general), a further approach based upon the simultaneous decomposition of a series of covariance matrices has been described in the literature [14].

One disadvantage of approaches which use the entire data array during the modeling stage is the large memory and computational burden. If possible, it is advantageous to first use image processing techniques, such as those described in the previous section, in order to transform the data to lower-dimensional arrays whilst retaining the essential information content present in the data. For the data discussed here, the pH mapping technique provides a meaningful preprocessing of the data in which the wavelength dimension is eliminated but the information pertaining to the distribution of CO_2 is retained. For this resultant univariate movie consisting of only three dimensions (*height* × *width* × *time*), the two options used here are (a) to build a trilinear PARAFAC model or (b) to unfold the data and build a PCA model on the *height* · *width* × *time* array.

7.1. Centering and scaling

Before describing the modeling of the movie array, a few words about the centering and scaling of movie arrays is appropriate. The subject of how to scale and center multiway arrays in general has already been addressed in the literature [23,40,41]. In the context of multivariate images, one form of scaling which as already been described here is pixelwise normalization, a technique for removing variation due to inhomogeneous illumination across the image space, although other types of scaling also exist [17]. Centering of multivariate image arrays is not always performed in the literature. However, for image arrays in which a time dimension is present, centering across this dimension can be useful. Two particularly useful centering options for chemical movies in which there is a sense of progression *to* or *from* a state of equilibrium are the subtraction of either the *last* or *first* image from every frame in the movie. For the data studied here, where the system moves towards a state of equilibrium in which the CO_2 is homogeneously distributed throughout the final image, centering using this image was applied and was found to lead to the most interpretable model.

7.2. PARAFAC

The multivariate movie was transformed into a movie of univariate pH maps using the mapping procedure. The last image was then subtracted from each frame and a

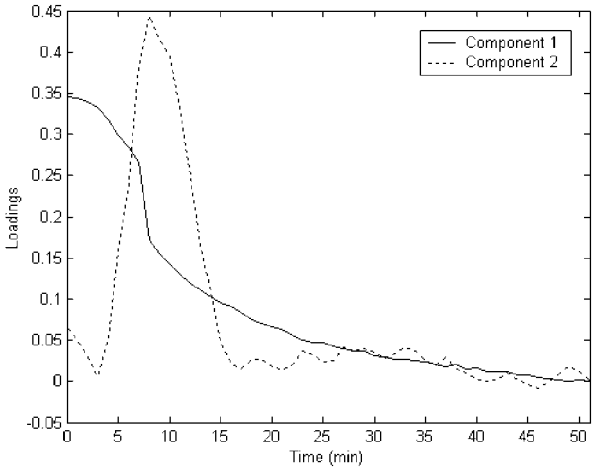

Figure 12. Loadings describing the time dimension from a PARAFAC of the movie.

two-component PARAFAC model was calculated, found to explain 98.82% of the data. The loadings for the time dimension are shown in Figure 12 and the reconstructed images are shown in Figure 13. The two components are found to describe two distinct features of the movie. The first component (Figure 13(a)) describes almost exclusively the pH difference between a thin layer (approximately 3 mm) of high CO_2 concentration at the surface of the water and the rest of the tank. This difference starts at a maximum and gradually decreases towards zero at the end of the run, where the CO_2 distribution is homogeneous (see component 1 in Figure 12). The component corresponds to known theory that gas uptake occurs in two stages: first, the air–water boundary is crossed; next, the gas is transported through the system by diffusion and/or turbulence. In the absence of turbulence, as in this investigation, the second stage is relatively slow, occurring in sporadic 'bursts' in an apparently chaotic process.

The second component (Figure 13(b)) describes an effect specific to this experimental run, that of the uneven spread of CO_2 throughout the tank leading to a 'hole' right of center, as seen in the original movie frame at $t = 7$ min (see Figure 4(c)). The loadings profile for this component moves to zero towards a peak at $t = 9$ min, after which it fades back to zero as the CO_2 distribution becomes more homogeneous (see component 2 in Figure 12).

Although not presented here, movies from other experimental runs, analyzed using the same methodology, exhibited the same pattern: one component describing the boundary layer and the other component describing a feature particular to the specific experimental run being analyzed. In some cases a component describing a general CO_2 flux from the top to the bottom of the tank is found, although for the run described here this is not the case (with only a slight gradient from top to bottom being visible in Figure 13(a)).

7.3. PCA

As a comparison, PCA was performed on the same data by first unfolding the array. The model described 99.06% of the data, and the loadings and score images are shown in

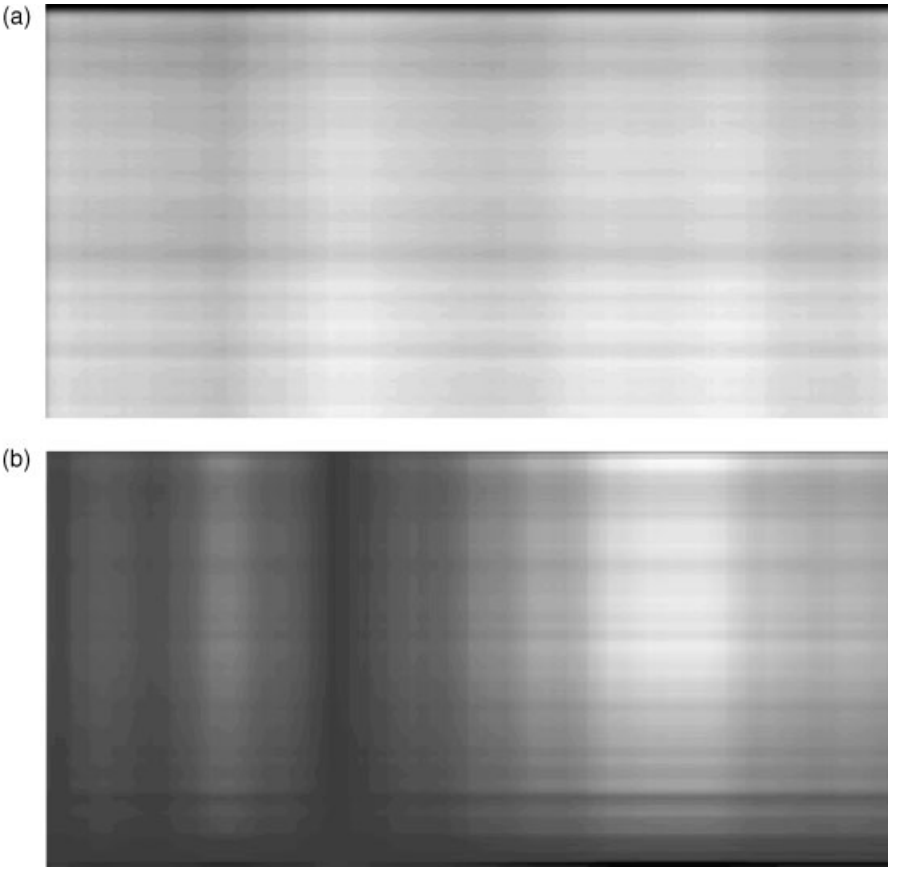

Figure 13. Reconstructed images for (a) component 1 and (b) component 2 from a PARAFAC of the movie.

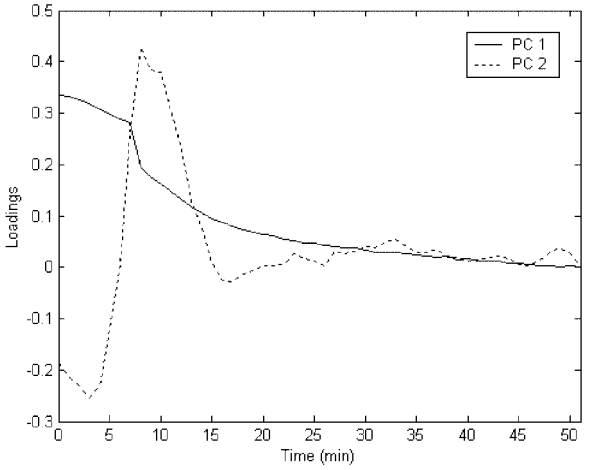

Figure 14. Loadings describing the time dimension from a PCA of the movie.

Figures 14 and 15 respectively. Whilst the same features—and interpretation—are found, in this particular case the clarity of interpretation is slightly diminished owing to the need for the model to satisfy an orthogonality constraint. This (a) forces the second component describing the time dimension (broken line in Figure 14) to dip below zero for the first 6 min and (b) causes a slight mixing of the two main features (i.e. the CO_2 layer and the 'hole') between the two score images. One possible solution to this problem is to employ a bilinear decomposition which uses alternative parameter constraints to identify the model [42–45].

8. CONCLUSIONS

In this paper the application of PARAFAC has been shown to be a useful approach for the analysis of multivariate images, offering a latent variable decomposition different from, and in some cases complementary to, the alternative PCA approach. One fundamental point in question is whether a linear decomposition of the spatial dimensions of an image, in addition to the wavelength and/or time dimensions, is beneficial. The answer is obviously highly dependent on the images being analyzed. Complex spatial forms do not lend themselves well to low-dimensional, linear decompositions, and the PCA approach leaves the spatial recognition part of model interpretation to the human (who is usually good at this). However, in cases where the important factors in the images can be described by simple decompositions, such as the CO_2 surface layer or the smooth (and not immediately perceivable) background effect found in this study, a discrimination between these spatial factors using different components can lead to an advantage in interpretation. Three-dimensional images, i.e. *height* × *width* × *depth*, are already common in other fields (e.g. brain scans), and a model which retains the spatial correlation information could prove to have some advantages over methodology in which spatial (auto)correlation is ignored.

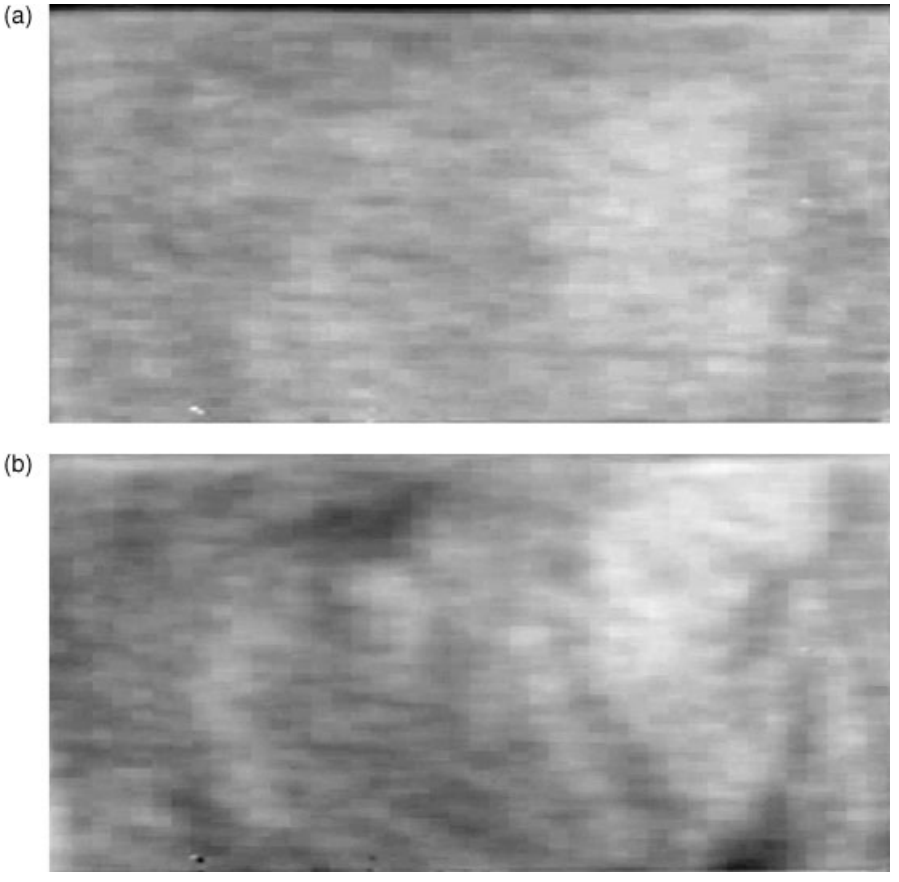

Figure 15. Reconstructed score images for (a) component 1 and (b) component 2 from a PCA of the movie.

A general approach for the analysis of time-resolved image data has been presented. The multivariate modeling of movies has been successfully used to identify features with different dynamic profiles. It can easily be envisaged that by fitting kinetic models to the profiles found, as has already been done within standard spectrometry [44,45], imaging could be used as a means of identifying chemical constants such as diffusion rates. Application of this methodology is not limited to video imaging, but is relevant to any form of chemical imaging in which measurement at regular time intervals is possible.

As has already been documented, one of the major problems in image analysis is the computational burden levied by the huge data arrays being produced. Whilst multivariate modeling techniques such as the subspace decompositions described here can play an important role, it is also useful to consider how image analysis and processing techniques—of which the histogram and mapping tools illustrated here are only a small part—can be used to extract the relevant chemical information from the data. One challenge in chemical imaging will be the combination of the multivariate statistical analysis methods already proving successful in chemistry with the vast range of image analysis tools already available in the image analysis sciences.

Acknowledgements

The authors gratefully acknowledge financial support from the State of São Paulo Research Foundation (FAPESP) and the Brazilian National Research Council (CNPq).

REFERENCES

1. Bhargava R, Levin IR. Fourier transform infrared imaging: theory and practice. *Anal. Chem.* 2001; **73**: 5157–5167.
2. Tran CD. Development and analytical applications of multispectral imaging techniques: an overview. *Fresentus J. Anal. Chem.* 2001; **369**: 313–319.
3. Tran CD. Visualising chemical composition and reaction kinetics by the near infrared multispectral imaging technique. *J. Near Infrared Spectrosc.* 2000; **8**: 87–99.
4. Macpherson JV, Jones CE, Barker AL, Unwin PR. Electrochemical imaging of diffusion through single nanoscale pores. *Anal. Chem.* 2002; **74**: 1841–1848.
5. Bharati MH, MacGregor JF. Multivariate image analysis for real-time process monitoring and control. *Ind. Eng. Chem. Res.* 1998; **37**: 4715–4724.
6. van Espen P, Janssens G, Vanhoolst W, Geladi P. Imaging and image processing in analytical chemistry. *Analusis* 1992; **20**: 81–90.
7. Geladi P, Isaksson H, Lindqvist L, Wold S, Esbensen K. Principal component analysis of multivariate images. *Chemometrics Intell. Lab. Syst.* 1989; **5**: 209–220.
8. Esbensen K, Geladi P. Strategy of multivariate image analysis (MIA). *Chemometrics Intell. Lab. Syst.* 1989; **7**: 67–86.
9. Lied TT, Geladi P, Esbensen KH. Multivariate image regression (MIR): implementation of image PLSR—first forays. *J. Chemometrics* 2000; **14**: 585–598.
10. Huang J, Esbensen KH. Applications of angle measure technique (AMT) in image analysis. Part I. A new methodology for *in situ* powder characterization. *Chemometrics Intell. Lab. Syst.* 2000; **54**: 1–19.
11. Indahl UG, Næs T. Evaluation of alternative spectral feature extraction methods of textual images for multivariate modeling. *J. Chemometrics* 1998; **12**: 261–278.
12. Gouti N, van Espen P, Feinberg MH. Quantitative reconstruction of objects from spatially correlated image sequences. *Chemometrics Intell. Lab. Syst.* 1999; **47**: 21–31.
13. Artyushkova K, Fulghum JE. Multivariate image analysis methods applied to XPS imaging data sets. *Surf. Interface Anal.* 2002; **33**: 185–195.
14. Courcoux P, Devaux M-F, Bouchet B. Simultaneous decomposition of multivariate images using three-way data analysis. Application to the comparison of cereal grains by confocal laser scanning microscopy. *Chemometrics Intell. Lab. Syst.* 2002; **62**: 103–113.
15. Lage EM. Fluxo de CO_2 na interface água–ar. *PhD Thesis*, Universidade Estadual de Campinas (UNICAMP), 2002.
16. Esbensen KH, Wold S, Geladi P. Relationships between higher-order data array configurations and problem formulations in multivariate data analysis. *J. Chemometrics* 1988; **3**: 33–48.
17. Geladi P, Grahn H. *Multivariate Image Analysis.* Wiley: Chichester, 1996.
18. Wold S, Geladi P, Esbensen K, Öhman J. Multi-way principal components- and PLS-analysis. *J. Chemometrics* 1987; **1**: 41–56.
19. Geladi P, Grahn H, Esbensen K, Bengtsson E. Image analysis in chemistry. II. Multivariate image analysis. *Trends Anal. Chem.* 1992; **11**: 121–130.
20. Harshman RA, Lundy ME. PARAFAC—parallel factor analysis. *Comput. Statist. Data Anal.* 1997; **18**: 39–72.
21. Bro R. PARAFAC. Tutorial and applications. *Chemometrics Intell. Lab. Syst.* 1997; **38**: 149–171.
22. Smilde AK. 3-way analyses—problems and prospects. *Chemometrics Intell. Lab. Syst.* 1992; **15**: 143–157.
23. Gurden SP, Westerhuis JA, Bro R, Smilde AK. A comparison of multiway regression and scaling methods. *Chemometrics Intell. Lab. Syst.* 2001; **59**: 121–136.
24. Sarmiento JL, Murnane R, le Quéré C. Air–sea CO_2 transfer and the carbon budget of the North Atlantic. *Philos. Trans. R. Soc. Lond. B* 1995; **348**: 211–219.
25. Sarmiento JL, Toggweiler JR, Najjar R. Ocean carbon-cycle dynamics and atmospheric pCO_2. *Philos. Trans. R. Soc. Lond. A* 1988; **325**: 3–21.
26. Jähne B, Haussecker H. Air–water gas exchange. *Ann. Rev. Fluid Mech.* 1998; **30**: 443–468.
27. Phillips LF. Experimental demonstration of coupling of heat and matter fluxes at a gas–water interface. *J. Geophys. Res.* 1994; **99**: 18577–18584.
28. Wanninkhof R, Asher W, Weppernig R, Chen H, Schlosser P, Langdon C, Sambrotto R. Gas transfer experiment on Georges Bank using two volatile deliberate tracers. *J. Geophys. Res.* 1993; **98**: 20237–20248.
29. Erickson III DJ. A stability dependent theory for air–sea gas exchange. *J. Geophys. Res.* 1993; **98**: 8471–8488.
30. Jähne B, Münnich KO, Bösinger R, Dutzi A, Huber W, Libner P. On the parameters influencing air–water gas exchange. *J. Geophys. Res.* 1987; **92**: 1937–1949.
31. Wanninkhof R. Relationship between wind speed and gas exchange over the ocean. *J. Geophys. Res.* 1992; **97**: 7373–7382.
32. Phillips LF. Steady-state thermodynamics of transfer through a gas–liquid interface, treated as a limiting case of thermo-osmosis. *Chem. Phys. Lett.* 1994; **228**: 533–538.
33. Andersson CA, Bro R. *Chemometrics Intell. Lab. Syst.* 2000; **52**: 1–4.
34. *The N-way Toolbox for MATLAB* [Online]. Available: http://www.models.kvl.dk/source/nwaytoolbox/ [6 February 2003].
35. Vogt F, Tacke M. Fast principal component analysis of large data sets. *Chemometrics Intell. Lab. Syst.* 2001; **59**: 1–18.

36. Gonzalez R, Woods R. *Digital Image Processing*. Addison-Wesley: Boston, MA, 1992; 414–428.
37. *MATLAB Image Processing Toolbox User's Guide (Version 2)*. The Mathworks: Natick, MA, 1997; 6–19.
38. Bro R. Multiway analysis in the food industry. *PhD Thesis*, University of Amsterdam, 1998; 113.
39. Martens H, Naes T. *Multivariate Calibration* (2nd edn), vol. 1. Wiley: Chichester, 1989.
40. Harshman RA, Lundy ME. Data preprocessing and the extended PARAFAC model. In *Research Methods for Multimode Data Analysis*, Law HG, Snyder CW, Hattie JA, McDonald RP (eds). Praeger: New York, 1984.
41. Bro R, Smilde AK. Centering and scaling in component analysis. *J. Chemometrics* 2003; **17**: 16–33.
42. Lawton WE, Sylvestre EA. Mathematical determination of the pure spectra of two components in a two-component mixture. *Technometrics* 1971; **13**: 617–633.
43. Tauler R. Calculation of maximum and minimum band boundaries of feasible solutions for species profiles obtained by multivariate curve resolution. *J. Chemometrics* 2001; **15**: 627–646.
44. Bezemer E, Rutan SC. Multivariate curve resolution with non-linear fitting of kinetic profiles. *Chemometrics Intell. Lab. Syst.* 2001; **59**: 19–31.
45. Bijlsma S, Boelens HFM, Hoefsloot HCJ, Smilde AK. Constrained least squares methods for estimating reaction rate constants from spectroscopic data. *J. Chemometrics* 2002; **16**: 28–40.

