Spatiotemporal-Chromatic Structure of Natural Scenes

Steven Bergner and Mark S. Drew
School of Computing Science,
Simon Fraser University
Vancouver, British Columbia, Canada V5A 1S6
{sbergner,mark}@cs.sfu.ca
Technical Report SFU-CMPT-TR-2004-11

Sept.1, 2004

Abstract

We investigate the implications of a unified spatiotemporal-chromatic basis for compression and reconstruction of image sequences. Different adaptive methods (PCA and ICA) are applied to generate basis functions. While typically such bases with spatial and temporal extent are investigated in terms of their correspondence to human visual perception, here we are interested in their applicability to multimedia encoding. The performance of the extracted spatiotemporal-chromatic patch bases is evaluated in terms of quality of reconstruction with respect to their potential for data compression. The results discussed here are intended to provide another path towards perceptually-based encoding of visual data by examining the interplay of chromatic features with spatiotemporal ones in data reduction.

1 Introduction

Decorrelation for redundancy reduction has a long history in image processing. In particular, variants of the Principal Component Analysis (PCA) for orthogonal decorrelation have been part of the arsenal of data reduction for many years. The main idea here is to account for most of the variance in the data using the first several principal axes, and then reduce the influence of further terms either by directly omitting these or by adopting a bit allocation scheme to deprecate their influence.

PCA can tell us how Nature processes vision, if we consider natural images. In particular, we expect to see color-opponent channels arise in a natural fashion, simply by automatic inspection of the data. But as well, we hope to glean evidence of how spatial processing operates. And in fact, Ruderman et al. [1] found not only such color-opponent structures but also spatial derivative-like filters, operating similarly and independently in each opponent-color and luminance channel.

The latter result was not surprising (although the decorrelation from color was): Olshausen and Field's seminal work on receptive field properties [2] implied that the receptive fields in mammalian primary visual cortex simple cells are spatially localized, oriented, and spatially bandpass in the sense of being selective to structure at different spatial scales, for non-color luminance inputs. Visually, these fields resemble a 2-dimensional Discrete Cosine Transform (DCT) basis in an $N \times N$ checkerboard structure (see, e.g., [3], and below), but with diagonal as well as rectangular basis images.

For videos, we'd like to ask whether we can bootstrap characteristics of human perception into effective data compression. Therefore we analyze several spatiotemporal-chromatic basis sets to see how combining the three aspects of video — spatial and temporal and chromatic — collude in natural scenes to promote compression. We address the fundamental question, and leave motion compensation as well as video transitions aside in this study. Results produced are very promising.

2 Related Work

As opposed to an orthogonal PCA basis, some workers have also considered an Independent Component Analysis (ICA) of natural images [4]. ICA proceeds by producing a minimally redundant set of basis functions. To do so, a set of

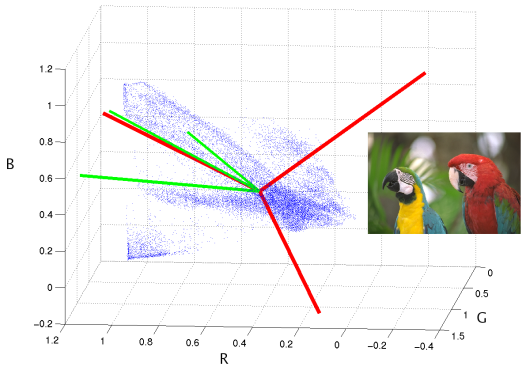


Figure 1: Basis vectors for a given color distribution from right image as found by PCA (red) and ICA (green).

maximally statistically *independent* basis vectors is found. The process of finding these vectors is based on the Central Limit Theorem, which states that a sum of non-Gaussian random variables is more like a Gaussian than are its individual components. But the independent sources sought can be written as sums of the observed data. Thus, we can move toward independent sources by trying to find a sum of the observed data over vectors which have maximum non-Gaussianity. This property can be measured in terms of higher order statistics, e.g. kurtosis or negentropy. For a more detailed discussion of ICA the interested reader is referred to [5]. The method we use here is the FastICA algorithm [6], which converges in cubic time.

ICA has been found useful for data reduction by 'sparse' coding, i.e., finding underlying sources such that any given image is naturally represented in terms of just a small number of these: ICA [4] reproduces results for optimizing sparseness [2]. Typically, the technique is used for the extraction of hidden sources generating observed data. For example, consider the image in Fig. 1. The RGB values in this image form clusters, as in shown in the left of Fig. 1, with orthogonal PCA axes shown in red. In contrast, the ICA axes (in green) show that the image is actually comprised of just a few independent sources. It should be noted that ICA is data adaptive: we would like to develop a set of basis vectors that is specific for a certain type of video contents. We could also target at developing a universal basis from a training set, but an adaptive model is bound to be more expressive since, as we see from Fig. 1, the hidden characteristics of content are extracted.

For grayscale imagery [2, 4], PCA indicates that a mutually orthogonal spa-

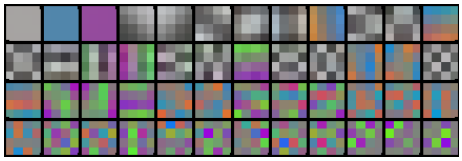


Figure 2: Spatio-chromatic basis obtained from PCA on $4 \times 4 \times RGB$ image patches of the example in Fig. 1.

tial basis for imagery consists of bandpass filters similar to 2-dimensional DCT basis images, but with some non-rectangular orientation present (cf. Fig. 2). For grayscale, how one creates such an image is by randomly selecting N-pixel by N-pixel square patches from an image set, vectorizing these N^2 values, and identifying the basis as the eigenvectors of the mean-subtracted covariance matrix. In contrast, ICA of grayscale imagery produces basis functions that are again bandpass, but are more obviously oriented and are similar to Gabor functions — Gaussian-windowed sine waves[2, 4].

2.1 ICA basis functions for natural images

A survey of applications of ICA to the processing of different media (image/video, multimodal brain data, audio, text, and combined data) is provided by [7]. However, while ICA has been widely used for classification, implications of ICA for multimedia compression have not been much studied, and usually have been discussed in simple terms of dimensional reduction. Studies including a bit allocation scheme have so far considered only audio [8], and grayscale imagery [9, 10], with inclusion of color in still imagery only in [11]. Here we extend results in [11] to color video.

When color is included, our patch vectors become $N \times N \times 3$ structures, for RGB images. PCA proceeds as stated above, but for these longer vectors, and the resulting structure can be visualized as in Fig. 2 (shown for $N = 4$). Note that the PCA basis is adaptive to the image (but in fact does not change much, for natural imagery we have tried).

In [1], Ruderman et al. first extended color PCA, as in the red vectors in Fig. 1, to a spatial patch domain by using 3×3 patches of 3-vector color data. In comparison, in a sense Fig. 1 shows results for 1×1 spatial patches. Ruderman et al. conclude that for foliage images, PCA of log long-medium-short (LMS) visual color channel data tends to decorrelate spatial processes from chromatic ones,

leading to 9 spatial features times 3 color ones. The latter are, in order, luminance, blue-yellow, and red-green. This result was extended by Wachtler et al. [12, 13] by replacing PCA with ICA, again for LMS data but now using 7×7 patches.

Color images and stereo vision have been investigated by [14] showing that the derived independent components also yield a separation of basis vectors into luminance and opponent colors.

Besides allowing for conclusions regarding human visual perception, these chromatic bases with spatial extent are very interesting from an image compression point of view. In the following we take a closer look at the implications of encoding visual data with respect to these bases.

2.2 Video processing

To include the temporal dimension, van Hateren and RudermanS [15] examined 12×12 patches of grayscale video, using ICA. They recovered in this case spatiotemporal bases that were that localized in both space and time, and bandpass in spatial and temporal frequency, appearing as oriented moving edges or bars. Hyvärinen et al. [16] introduced a unifying framework that combines sparseness with ICA, temporal coherence, and topography to consider complex cell pooling in a single model. Again, these authors are primarily concerned with relating ICA to human vision, whereas here we aim at understanding the potential for compression.

3 Video basis functions

The goal of our analysis is to compare the suitability of different data-adaptive basis functions for compressing color video. In the following we will consider two data-adaptive sets of bases — *viz.* PCA and ICA — first for still imagery and then including temporal dependence.

Fig. 3 shows the result of ICA performed on a standard image set, for a particular patch size. While the results of DCT and PCA can be interpreted as a frequency decomposition of the data, the functions obtained by ICA exhibit a combined localization in space and frequency [14, 11]. Apparently, this basis seems to again treat opponent color separately from luminance — similar to the observation that is already illustrated in Fig. 2 for the PCA case.

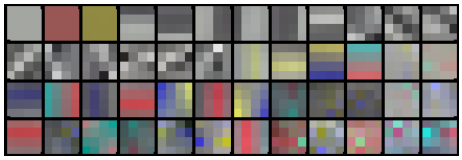


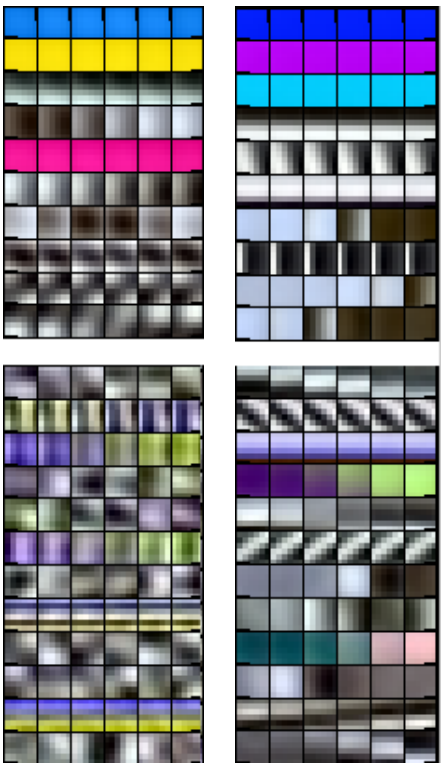
Figure 3: Basis patches for ICA of spatio-chromatic $4 \times 4 \times RGB$ patches. Sorted in order of decreasing variance.

Besides deciding on a method of basis generation, we have to make a choice about the size of the patches we will operate on. To create a basis, the analysis is performed on squared (or cubed, for video) pixel neighborhoods. We randomly sampled a total of 50000 patches over the frames of a set of videos. Also, for ICA we have restricted the analysis to the first 250 basis vectors. This saves computational resources and helps to suppress influence of higher frequency noise. The resulting basis functions then reflect the statistical properties of the presented data.

The method applied to 2D image patches above is now extended to 3D spatiotemporal color patches. Fig. 4 compares patches that result from analyses using PCA and ICA. The ICA-generated patches are seen to more resemble localized frequency properties of the data. In comparison, PCA extracts frequency information over the entire patch. This corresponds to observations for luminance-only image sequences [15]. Furthermore, we can state that the separation of basis vectors for color and luminance is still retained with motion information included; also, the color bases are restricted to opponent color pairs. Besides this visual comparison, these two video basis sets are compared below based on their suitability for data compression.

Note that for the statistical analysis, all videos have been normalized to the same scaling along the spatial and the temporal axes, which amounts to a fixed resolution and frame rate. For variation in scale of natural objects it is possible to argue that due to their fractal character they have scale invariant frequency characteristics. Nevertheless, for more general classes of objects, scale (or richness of detail in the scene) will probably influence the formation of the basis.

To use the basis for reconstruction, video frames are regularly tiled spatially and temporally into an arrangement of non-overlapping patches. The coefficients for each spatiotemporal patch of the video can be obtained by a linear transform using the filter patches. These are essentially the inverse of the (non-orthogonal) basis patches. Respectively, going back from the coefficients to the actual image data is done by transforming the coefficients in a linear combination of the basis



(a) (b)

Figure 4: Spatiotemporal color basis patches of size $6 \times 6 \times 6 \times RGB$ obtained through (a) PCA and (b) ICA. Rows are sorted in order of decreasing variance-accounted-for from top to bottom. Time increases from left to right over 6 frames. The gap in the middle represents 10 omitted vectors, to show the increasing importance of color further down in the sequence.

patches.

3.1 Quantization and entropy encoding of coefficients

After having projected the video data to the new basis the resulting coefficients have to be reduced in some way. If no reduction takes place no compression will apply. The method we have implemented performs a variance-based quantization. The discretized output can then be efficiently compressed using entropy-based compression.

The coefficients are given as continuous (floating point) numbers. To apply entropy coding, we have to quantize them first. In our tests we have decided for the method of assigning each basis vector a number of bits proportional to its standard deviation [17]. The proportionality factor is chosen to fulfill a given overall contingent of bits. Given the number of bits, each channel of coefficients is uniformly quantized from its minimum to its maximum occurring value.

Now, having expressed the coefficients as discrete numbers, we can apply an entropy-based encoding, e.g. Huffman coding, or other variable length coding (VLC). The VLC compression indicated in the graphs below is a theoretical limit that can be computed from the sum of entropies for all channels by exponentiation to the base 2. The rate of compression, then, is the factor by which the estimated encoded data is smaller than the original data, which has been stored with 8 bits per channel.

3.2 Quality of reconstruction

Besides considering at the size of the data after compression, we most importantly have to look at the quality of the reconstruction obtained from the reduced representations. Here we simply asses this quality using the peak signal to noise ratio (PSNR), in decibels (dB)

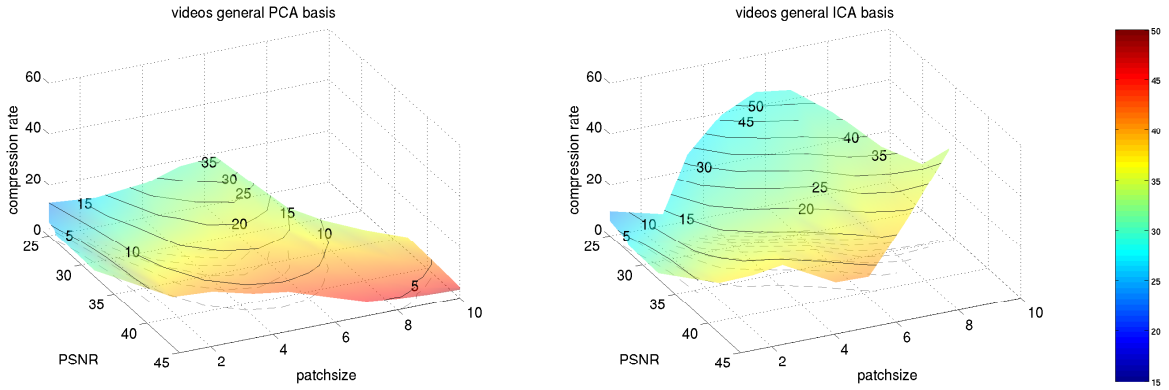
$$PSNR = 10 \log_{10} \left(\frac{G^2}{MSE} \right) \tag{1}$$

$$MSE = \frac{\sum_i^N \sum_j^M \sum_{k \in RGB} (p_k(i,j) - o_k(i,j))^2}{3MN} \tag{2}$$

where G is the maximum representable value (e.g. 255) and MSE is the mean squared error of the reconstructed picture, $p_k(i,j)$, with respect to the original, $o_k(i,j)$. A drawback of the PSNR is that it does not actually reflect the distortion as *perceived* by a human observer. Nevertheless, it is a convenient measure to illustrate the quality of reconstruction in a physical sense, and shall be sufficient for our subsequent evaluation.

4 Evaluation of different bases

The efficiency of a basis is understood as the relation between image quality retained for an achievable rate of compression (or vice versa). Thus, we have conducted a number of tests for different sets of video bases (ICA, PCA, and DCT). Each basis is generated and applied separately over range of cubic patch sizes



(a) (b) (c)

Figure 5: Entropy compression using (a) PCA generated and (b) ICA basis functions; color indicates PSNR as per (c).

from $1 \times 1 \times 1$ to $10 \times 10 \times 10$ with each dimension $\times$ RGB. The first case is similar to just interpreting the pixel colors. As patch size increases, the influence of neighbors in space and time is included more and more. Another variable in the comparison is the compression parameter. This is the ratio to the overall maximum number of bits for the stream. Note that the achievable compression entirely depends on the entropy of the data.

4.1 Compression vs. quality using spatiotemporal-chromatic bases

Fig. 5(a,b) provides a comparison of the entropy-based variable length coding of the spatiotemporal-chromatic coefficients of ICA vs. PCA (DCT gives very similar results to PCA). The edge length of the cubic basis patches is given in pixels. The compression rate is the factor by which the encoded data is smaller than the original video. Here the property of ICA to result in sparsely coded coefficients becomes apparent. The lower entropy of the quantized data results in significantly higher compression rates. Both plots show a significant improvement of the compression/error tradeoff as the size of basis patches increases.

5 Summary

The computation of individual bases for similar images sequences is interesting from a vision and a multimedia point of view. While the first point has been subject of previous work targeting analogies to human perception, we have tried to illuminate the latter. The results indicate a significant difference comparing the compressibility of coefficients from ICA and PCA. The sparse coding property of ICA bases has been shown to have a noticeable impact on the efficiency of subsequent entropy compression.

A problem inherent in the approach of adaptive bases is that they first have to be generated in a computationally expensive preprocessing. Furthermore, a basis specific to one data set would have to be stored along with the coefficients to allow for decoding. This would certainly add overhead to the compressed data. Nevertheless, in a constrained domain it is possible to prepare basis functions that can be reused.

Because of the proximity of the outcome of independent component analysis to receptive fields of simple cells in the V1 visual cortex, it could be possible to derive a more perceptually-based error metric for evaluation of the quality of visual representations. Advances of research in the human perceptual system may lead the way to an error metric that more closely corresponds to the assessment by a human observer.

Another interesting property of the ICA basis is that it resembles expressive features of the data. This property also hints at the relationship between ICA filters and wavelet analysis. Taking this into account, it seems worthwhile to consider the compressed coefficients as a higher-level feature description of the visual data. In terms of video analysis these features might be useful for object tracking.

References

[1] Daniel L. Ruderman, Thomas W. Cronin, and Chuan-Chin Chiao, "Statistics of cone responses to natural images: Implications for visual coding," *J. Opt. Soc. Am.*, vol. 15, no. 8, pp. 2036–2045, August 1998.

[2] B.A. Olshausen and D.J. Field, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," *Nature*, vol. 381, pp. 607–609, 1996.

[3] Z.N. Li and M.S. Drew, *Fundamentals of Multimedia*, Prentice-Hall, 2004.

[4] J.H. van Hateren and A. van der Schaaf, "Independent component filters of

natural images compared with simple cells in primary visual cortex," *Proc. Roy. Soc. B*, vol. 265, pp. 359–366, 1998.

[5] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, John Wiley and Sons, Inc., New York, 2001.

[6] E. Oja, "Convergence of the symmetrical FastICA algorithm," in *9th Int. Conf. on Neural Information Processing*, 2002.

[7] J. Larsen, L. K. Hansen, T. Kolenda, and F. AA. Nielsen, "Independent component analysis in multimedia modeling," in *Fourth Intl. Symposion on ICA and BSS*, Nara, Japan, apr 2003, pp. 687–696.

[8] A. Ben-Shalom, M. Werman, and S. Dubnov, "Improved low bit-rate audio compression using reduced rank ICA instead of psychoacoustic modeling," in *ICASSP 2003*, 2003.

[9] A.J. Ferreira and M.A.T. Figueiredo, "Class-adapted image compression using independent component analysis," in *ICIP 2003*, 2003, pp. I: 625–628.

[10] A.T. Puga and A.P. Alves, "An experiment on comparing PCA and ICA in classical transform image coding," in *ICA 98*, 1998, pp. 105–108.

[11] M.S. Drew and S. Bergner, "Analysis of spatio-chromatic decorrelation for colour image reconstruction," in *12th Color Imaging Conference: Color, Science, Systems and Applications.* Society for Imaging Science & Technology (IS&T)/Society for Information Display (SID) joint conference, 2004.

[12] T. Wachtler, T.-W. Lee, and T.J. Sejnowski, "Chromatic structure of natural scenes," *J. Opt. Soc. Am. A*, vol. 18, pp. 65–77, 2001.

[13] T.-W. Lee, T. Wachtler, and T.J. Sejnowski, "Color opponency is an efficient representation of spectral properties in natural scenes," *Vis. Res.*, vol. 42, pp. 2095–2103, 2002.

[14] P.O. Hoyer and A. Hyvrinen, "Independent component analysis applied to feature extraction from colour and stereo images," *Network: Computation in Neural Systems*, vol. 11, no. 3, pp. 191–210, 2000.

[15] J.H. van Hateren and D.L. Ruderman, "Independent component analysis of natural image sequences yields spatio-temporal filters similar to simple cells in primary visual cortex," *Proc. Roy. Soc. B*, vol. 265, pp. 2315–2320, 1998.

[16] A. Hyvärinen, J. Hurri, and J. Väyrynen, "Bubbles: a unifying framework for low-level statistical properties of natural image sequences," *JoSA A*, vol. 20, no. 7, pp. 1237–1252, July 2003.

[17] Pamela C. Cosman, Robert M. Gray, and Martin Vetterli, "Vector quantization of image subbands: A survey," *IEEE Transactions on Image Processing*, vol. 5, no. 2, pp. 202–225, 1996.