

The Effect of Image Compression on Classification and Storage Requirements in a High-Throughput Crystallization System

Ian Berry¹, Julie Wilson², Chris Mayo¹, Jon Diprose¹, and Robert Esnouf¹

¹ Division of Structural Biology and the Oxford Protein Production Facility
University of Oxford, The Henry Wellcome Building for Genomic Medicine
Roosevelt Drive, Oxford, OX3 7BN, UK
{ian,mayo,jon,robert}@strubi.ox.ac.uk
<http://www.strubi.ox.ac.uk>

² York Structural Biology Laboratory, Department of Chemistry, University of York
Heslington, York, YO10 5DD, UK
julie@ysbl.york.ac.uk

Abstract. High-throughput crystallization and imaging facilities can require a huge amount of disk space to keep images on-line. Although compressed images can look very similar to the human eye, the effect on the performance of crystal detection software needs to be analysed. This paper tests the use of common lossy and lossless compression algorithms on image file size and on the performance of the York University image analysis software by comparison of compressed Oxford images with their native, uncompressed bitmap images. This study shows that significant (approximately 4-fold) space savings can be gained with only a moderate effect on classification capability.

1 Introduction

The Oxford Protein Production Facility (OPPF) is a facility within the Division of Structural Biology at the University of Oxford that is funded by the MRC to develop technologies for the high-throughput production and crystallization of proteins. It is seeking to develop a protein production pipeline by automating, parallelizing and miniaturizing all stages of the process. Target proteins are human proteins and those of human pathogens, selected for their direct biomedical relevance. The OPPF is the first stage in a structural genomics programme for the UK and represents an essential stepping stone toward the practical exploitation of the wealth of information coming from the human genome sequencing projects.

The key stages in the OPPF pipeline are target selection, cloning, expression, purification and crystallization. It is the crystallization phase that is relevant to this paper. Crystallization experiments are performed in 96-well Greiner crystallization plates using the sitting drop vapour diffusion method and a Cartesian Technologies Microsys MIC400 (Genomic Solutions, Huntingdon, UK) to dispense 100nL drops [1-2].

Once the plates are created, they are sealed and placed in a temperature-controlled storage vault (The Automation Partnership, Royston, UK) that has a capacity for 10000 crystallization plates. An automated Oasis 1700 imaging system (Veeco, Cambridge, UK) has been integrated with the storage vault and images of the crystallization droplets are taken automatically at regular intervals. Images are classified automatically using York University image analysis software and presented in a web interface for manual inspection.

At this time, this system has taken over nine million images and is now generating up to seventy-five thousand images per day. Native images are 1 megabyte bitmap (BMP) images and after cropping by Veeco software, they are reduced to 520 kilobytes. This causes a distinct storage problem and while it is necessary to keep all the images taken on-line, ready for access through a web interface by the crystallographers, there is also a need to minimise storage required.

In order to make all images available on-line for browsing, the file size was reduced as much as possible without affecting the visual state of the image. This led to the choice of jpeg with a quality setting of 60% as the image file type for the permanent on-line storage. The original bitmap images are currently stored online for about a month before being migrated to tape.

As improvements are made to the crystal classification software and as new techniques such as time-course analysis are introduced, more and more uncompressed images are needed on-line. As a result, there is a need to determine the best trade off between image file size and the ability to repeat the original classification.

2 Image Classification

The crystal image classification software [3-7], used to detect the presence of crystals in drops, is still under development and its accuracy is still such that manual classification is also required. The software, described elsewhere [3-4], currently makes classifications based on single images, as they are generated, to determine a simple numerical classification score (Table 1).

Table 1. Classification output categories from the image classifier provide the simple score output from the classifier

Class Number	Class Description
-1	Unable to classify drop
0	Empty Well / Drop
1	Rubbish
2	Precipitate
3	Interesting or granular precipitate
4	Small crystals or something else interesting
5	Some Crystals
6	Good Crystals

3 Jpeg Compression

Jpeg compression [8-9] is a four stage process. The first stage separates the image into 8×8 pixel tiles and converts the image colour space from red-green-blue (RGB) into a luminance / chrominance colour space such as YUV (where Y is luminance [brightness] and UV is the chrominance [colour]).

For the second stage, each block is then passed through a Discrete Cosine Transform (DCT) which calculates and stores the value of the pixel relative to the average for the block.

The third stage is the where the compression occurs – according to the quality setting chosen when encoding the file. It determines the number of quantisation levels that are applied to the array of DCTs. This is achieved by generating two tables, one each for luminance and chrominance, of quantisation levels between the minimum and maximum DCT values and split according to the quality value. The DCT coefficients are then quantised using these tables.

The final stage involves encoding the reduced DCT coefficients using a Huffman encoding scheme which compresses the data (but adds no extra compression to the image information itself).

The final jpeg image will also have header information (including details of the encoding process) added to it to complete the file.

4 Method

In order to ascertain the properties of the different image encoding, a selection of around 1100 images was chosen that showed most of the different types of features that can occur in crystallization experiments. The original bitmap images were converted using ImageMagick [10] into the various different formats on test. Code was also added to the classifier to allow it to parse the different image types. For this study, the image formats chosen were limited to jpeg (with various quality factors – Fig. 1), Graphic Interchange Format (GIF) and Run Length Encoded Bitmap (RLE BMP).

The purpose of this study is not to evaluate the accuracy of the image classification software; rather it is to compare the consistency of classification of reduced-size images with those obtained using the original bitmap images. This allows two useful simplifications: First, the classification program does not need to be trained individually for each image type. Second, it gives an obvious metric for assessing consistency as the fraction of images classified the same way for the reduced-size and native images.

Image classification was performed using a C version of the image classifier running under Red Hat Linux 9.0 on a PC containing two 2.66GHz Intel Pentium 4 Processors. Average classification time per image on this system was 2 – 4 seconds depending on the complexity of the image and the level of compression.

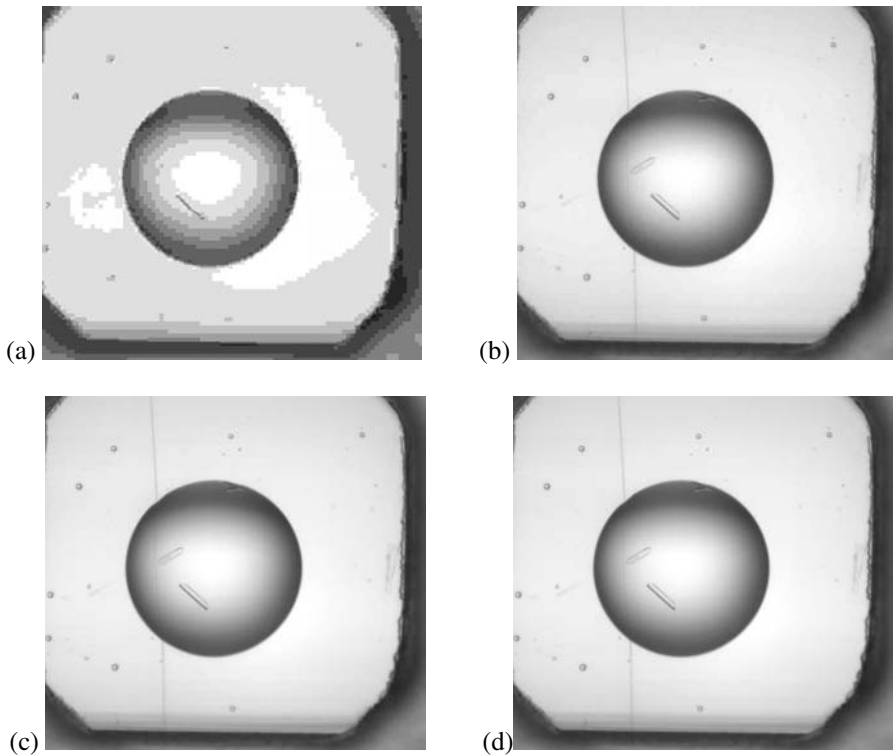


Fig. 1. A sample image compressed with various jpeg quality settings (a) 1%, (b) 30%, (c) 60% and (d) 100%

5 Results

The bitmap drop images typically cover the full 8-bit range of the grey scale and there is significant variation in background intensity over both the background and drop (e.g. Fig. 1(d)). Unsurprisingly, therefore, the two lossless ‘compression’ algorithms failed to reduce the image size (Table 2) and so were not tested further against the classifier. GIF images use an indexed colour palette and so the use of the full dynamic range in the images renders this ineffective. In the case of the RLE bitmap, the bloated size is due to the noisiness of the images – in a low noise image, RLE can provide significant file size savings as it compresses lines of pixels of the same value together (first pixel is the number of pixels, next pixel is the colour) so where there a few pixels that are the same, each pixel is replaced by two pixels, hence expanding the file size.

For the jpeg images, the resulting file size is critically dependent on the quality setting (Fig. 2), but as a guide 90% quality setting reduces the file size by approximately 90%.

Table 2. Overall file size information for various image formats

File Type	Average File Size (bytes)
BMP	526078
RLE BMP	1576454
GIF	596089
Jpeg	Min: 2903 Max: 190863

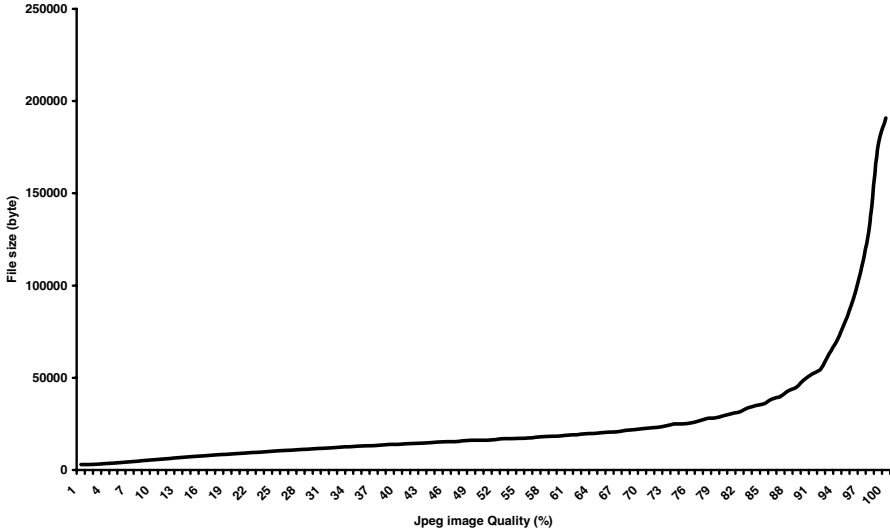


Fig. 2. The variation of image size against image quality

The ensemble of images were converted to jpegs with quality settings ranging from 1% to 100%. These were tested using the classifier and the results compared with those obtained from the original bitmaps. As a control, the results were also compared to a “dumb classifier” that made all the images equal to one class only (the sum of the percentage accuracies for these dumb classifiers being 100%). The results can be seen in Fig. 3 as percentage agreement with classifications for the original bitmap images.

The results show a striking dependence of classification quality on jpeg quality setting. Even with a quality setting of 95%, less than 50% matched those with the uncompressed bitmap. For a jpeg image quality of 100%, the agreement is still only 80%. This demonstrates the approximations inherent in the jpeg algorithm (which could be attributed the colour space transform and DCT calculations) and the sensitivity of the classification algorithm to the fine details of the image. For low quality jpeg images (< 90% quality), the classifications are little better than random, and indeed a fixed classification of “All Class 6” (all drops contain a crystal) would give a comparable classification consistency with this ensemble! It should be noted that ‘interesting’ drops are heavily over-represented in this test set – for our complete

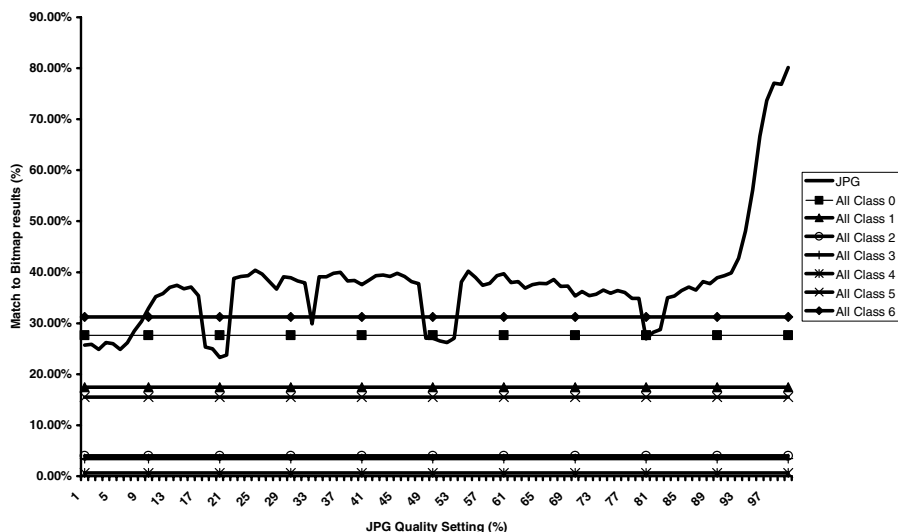


Fig. 3. Comparison of different image types and classifier outputs for different jpeg image quality settings

‘interesting’ drops are heavily over-represented in this test set – for our complete database, only 1-2% of drops contain crystals, whereas 50% are empty drops.

In the low quality region, there are several places where the consistency drops significantly further (e.g. 20%, 50% and 80% quality). It appears that these may be explained by a combination of the artifacts caused by the compression algorithm, particularly the tiling, and the techniques used in the classification.

At the low quality limit, the losses in the image are so severe that even the drop detection algorithm begins to fail and the consistency tends to the “All Class 0” classification (i.e. no drop detected).

The main purpose of this study is to relate image storage requirements to image classification. Fig. 4 summarizes this as the correlation between classification matches and file size. For a reasonable (> 75%) consistency with bitmap classifications, the smallest file size is around 120 kilobytes. This is a good reduction from the 512kb that is required for the uncompressed bitmap files, but would result in a significant misclassification of images. At this compression level and with the current image acquisition rate, our database would grow by approximately 8 gigabytes per day.

6 Conclusion

As crystallization storage and imaging systems become more affordable and widespread, many labs will be faced with the twin problems of image classification and

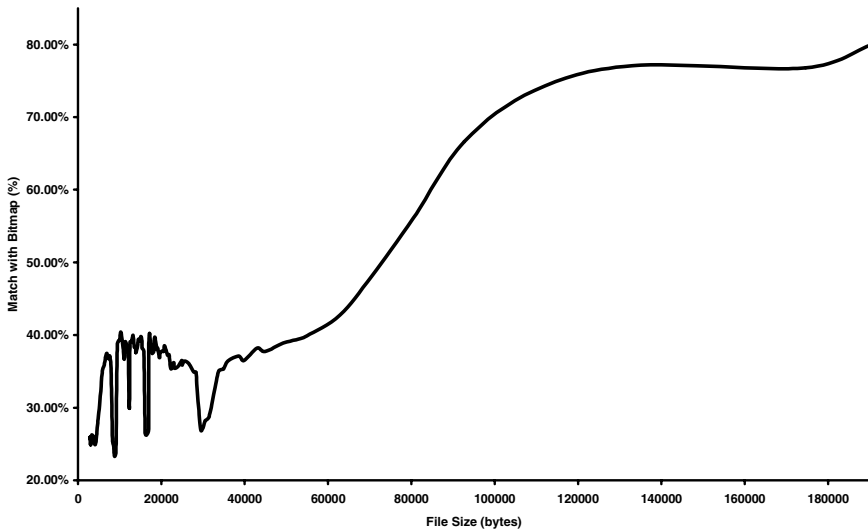


Fig. 4. Comparison of file size against similarity to bitmap results for different bitmap quality settings

image storage. Software developments have focussed on the analysis of bitmap images, but these can pose data storage problems – even at its current rate, the OPPF is acquiring about 3 terabytes annually. If disk and data redundancy are included to ensure data integrity, then the problem is even bigger. Whilst a development facility such as the OPPF may be able to provide sufficient resources, the cost of keeping image libraries may become prohibitive for some sites and ways to reduce this requirement need to be investigated.

Lossless compression algorithms may provide some saving with no cost to classification accuracy, but savings are likely to be modest – and in some cases even increase the file size.

We have investigated lossy jpeg compression as giving a potential for far greater savings that have to be offset against loss of image analysis accuracy. For our test data set (heavily biased in favour of ‘interesting’ images compared to our full database) a compression ratio of approximately 4:1 (95% image quality) yielded a classification consistency of 75% that may still be sufficient.

As the image classification software develops, two trends can be foreseen. First the method becomes more robust and therefore becomes less affected by image compression. Second, increasingly subtle characteristics of the image may be used to improve classification accuracy and therefore the method may become increasingly sensitive to loss of image detail. More radical changes may also be foreseen such as the use of multiple imaging or time course data, both of which will require substantially more images to be analysed and stored.

Acknowledgements

The OPPF is funded by the UK Medical Research Council, Ian Berry is supported by the Wellcome Trust and Julie Wilson is supported by a Royal Society University Research Fellowship.

References

1. Walter, T, et al: A procedure for setting up high-throughput, nanolitre crystallization experiments. I: Protocol design and validation. *J Appl Cryst* 36 (2003) 308-314
2. Brown, J., Walter, T., et al: A procedure for setting up high-throughput, nanolitre crystallization experiments. II: Crystallization results. *J Appl Cryst* 36 (2003) 315-318
3. Wilson, J.: Towards the automated evaluation of crystallization trials. *Acta Cryst. D58* part 11 (2002) 1907-1914
4. Wilson, J.: Automated evaluation of crystallisation experiments. *Cryst. Rev.*, Vol. 10, No.1 (2004) 73-84
5. Spraggon, G., Lesley, S.A., Kreusch, A. and Priestle, J.P.: Computational analysis of crystallization trials. *Acta Crystallographica, D58* part 11 (2002) 1915-1923
6. Cumbaa, C.A., Lauricella, A., Fehrman, N., Veatch, et al: Automatic classification of sub-microlitre protein-crystallization trials in 1536-well plates. *Acta Cryst. D59* (2003) 1619-1627
7. Bern, M., Goldberg, D., Stevens, R.C. and Kuhn, P.: Automatic classification of protein crystallization images using a curve-tracking algorithm. *J. Appl. Cryst.*, 37 (2004) 279-287
8. Joint Photographic Experts Group (JPEG): <http://www.jpeg.org>
9. JPEG Image Compression FAQ: <http://www.faqs.org/faqs/jpeg-faq/>
10. ImageMagick: <http://www.imagemagick.org>