Determining Facial Expressions in Real Time

Yael Moses*, David Reynard and Andrew Blake
Dept. of Engineering Science
University of Oxford
Oxford OX1 3PJ.

Abstract

In this paper we suggest an approach to describing and tracking the deformation of facial features. We concentrate on the mouth since its shape is important in detecting emotion, however we believe that our system could be extended to deal with other facial features. In our system, the mouth is described by a valley contour which is based between the lips. This contour is shown to exist independently of illumination, viewpoint, identity, and expression. We will present a real-time mouth tracking system that follows this valley. It will be shown to be robust to changes in identity, illumination and viewpoint. A simple classification algorithm was found to be sufficient to discriminate between 5 different mouth shapes, with a 100% recognition rate.

1 Introduction

Faces are dynamic objects that can undergo a vast number of non-rigid transformations. Facial images contain features that can be used to help understand different expressions and facial states. For instance, the shape of the mouth can be used to detect the emotive state of a subject, such as happiness or anger. When sequences of data are available, then the mouth may be used to help in speech recognition.

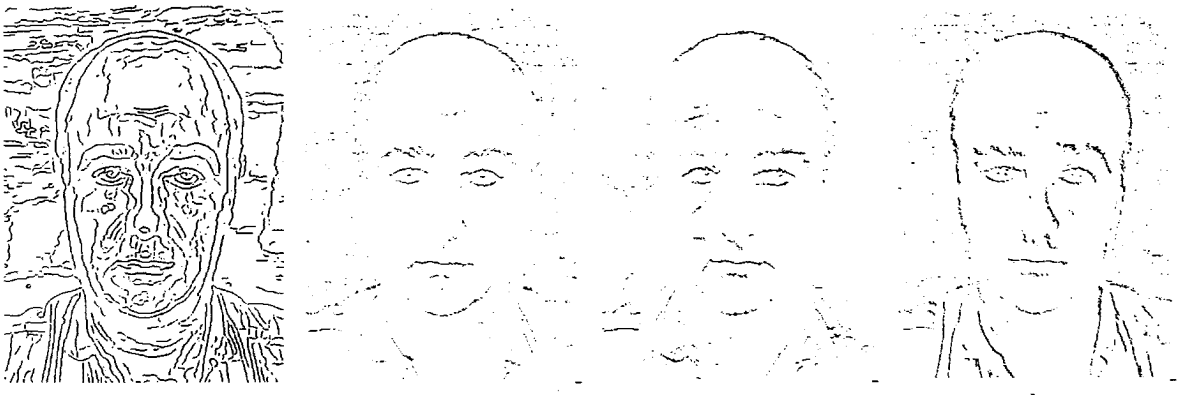
In order to identify and analyse facial expressions a well defined description of the deformations is required. It is relatively straightforward to identify a description of rigid transformations such as rotation and translation, but a simple, natural description of non-rigid deformations does not exist. One method for creating a framework in which to study facial deformations is to construct a model of the facial muscles and skin tissue based on anatomical descriptions [14]. Such an approach is limited by the quality of existing biological models and by its complexity. Another approach only models the parts of the face that are relevant for understanding motion. Typically, only the eyes and mouth are studied, since these are considered to be the most important features [11]. On static images, expressions have been represented as the deviation of up to 100 special points from their neutral position [10]. Detecting the special points automatically is a non-trivial problem and therefore features were often extracted manually. Automatic modelling of facial features from sequences of images was studied by [3]. Real time tracking of special points on the face was implemented by Gee and Cipolla (1994) to determine the face's gaze direction. Essa *et al* (1994) suggest the use of image sequences in expression understanding. In their work, optical flow was used.

In this paper we address the problem of constructing a framework that will allow the description of facial expressions and movement based on image features. Furthermore, we have studied the detection and tracking problem in real-time. Although we have concentrated on tracking the mouth, we believe that our techniques will be applicable to other features. We present a tracker that follows the valley in pixel intensity that lies between the two lips. The valley contour is tracked using 'dynamic contours' which consist of an estimator that uses both prediction and measurement to follow the movement of the contour. The prediction uses an *a priori* model of the contour dynamics perturbed by noise. In this case the estimator is a Kalman filter. The advantage of this approach is that the dynamics may be 'tuned' using a motion learning algorithm. It is also relatively straightforward to implement the full algorithm on modest hardware in real time, typically running at 50 Hz.

Once tracking can be achieved in real time, the results can be used for classifying the shape of the mouth. We will present a classifier that can discriminate between five mouth shapes: a neutral expression, a down-turned expression, an open mouth, a smile with close mouth or with the teeth showing, and an expression with pursed lips. The conclusions suggest ways in which

*Currently at Dept. of Applied Mathematics and Computer Science, The Weizmann Institute of Science, Rehovot 76100, Israel

0-8186-7042-8/95 $4.00



a b c d

Figure 1: All images show the same face under different conditions. (a) shows the canny edges, (b) shows the horizontal valleys, (c) shows the horizontal valleys when the illumination is from the left, not from the center, (d) shows the effect of rotating the valley direction through 45°.

this technique could be extended to track the mouth more fully. Additionally, further applications of our approach are considered, some of which are already under construction.

2 The Use of the Valley Contour as a Feature

The lips around the mouth are usually a different colour from the surrounding skin. Therefore edges have often been used as image features by face recognition or tracking systems. Unfortunately the number of edges near the mouth is quite large as can be seen in Fig. 1(a). This is not only due to the parameter choice, but due to changes in the albedo together with the shadows in the mouth area. It is, therefore, hard to detect which edge corresponds to which lip. Another disadvantage in using the edges directly is that the contrast of the edges depends on the illumination. It is therefore necessary to change the parameters of the edge detector when the illumination changes. The net result of this is that existing methods use artificial techniques for ensuring that tracking systems remain stable. One method uses make-up to colour parts of the face [13]. Another approach tracked the lips from the side, where the contrast would be much greater [3].

In this paper we suggest the use of a valley contour that lies between the two lips as a feature. Valley contours have previously been used by [9, 15, 6]. Fig. 1(b) shows that there are fewer valleys than edges around the mouth. A simple directional valley detector was implemented. Given a direction on the image $u = (\cos(\theta), \sin(\theta))^T$ where the intensity of the image is I then a valley point, in this case, occurs at a point where:

$$\text{and} \begin{cases} u \cdot \nabla I &= 0 \\ u^T(\nabla\nabla I)u &> 0 \end{cases}$$

The valley contour is then a set of connected valley points. These expressions are approximated so that, in practice, the presence of a valley at a point P is determined by sampling the pixel intensity at the three points P, $P+\mu u$ and $P-\mu u$ where μ is a scalar referred to as the "valley scale" parameter. This type of valley is robust to the direction parameter, θ. A change of 45° to the valley direction (Fig. 1 (b) and (d)) only produces minor changes to the valleys around the mouth The valleys are also robust to the valley size. The robustness of the valley to illumination is demonstrated in Fig. 1 (b) and (c), where the same parameters were used to produce images of the same face with illumination from the left and the center. The existence of the valley between the lips is also independent of changes in facial expressions.

The disadvantage of using only valleys is that they lack some information about the lips' position contained in the edges. However, in future work we intend to use the valley to help in the detection of the lip edges and to determine the visibility of the teeth and tongue. This will give a full description of the mouth shape.

3 The Tracking Framework

The tracker described in this section consists of a Kalman filter designed to model the dynamics of a moving contour. The contour, a quadratic B-spline can be described in terms of a set of control points. Measurements are made along the length of the spline which relate the contour to the tracked feature. The estimate

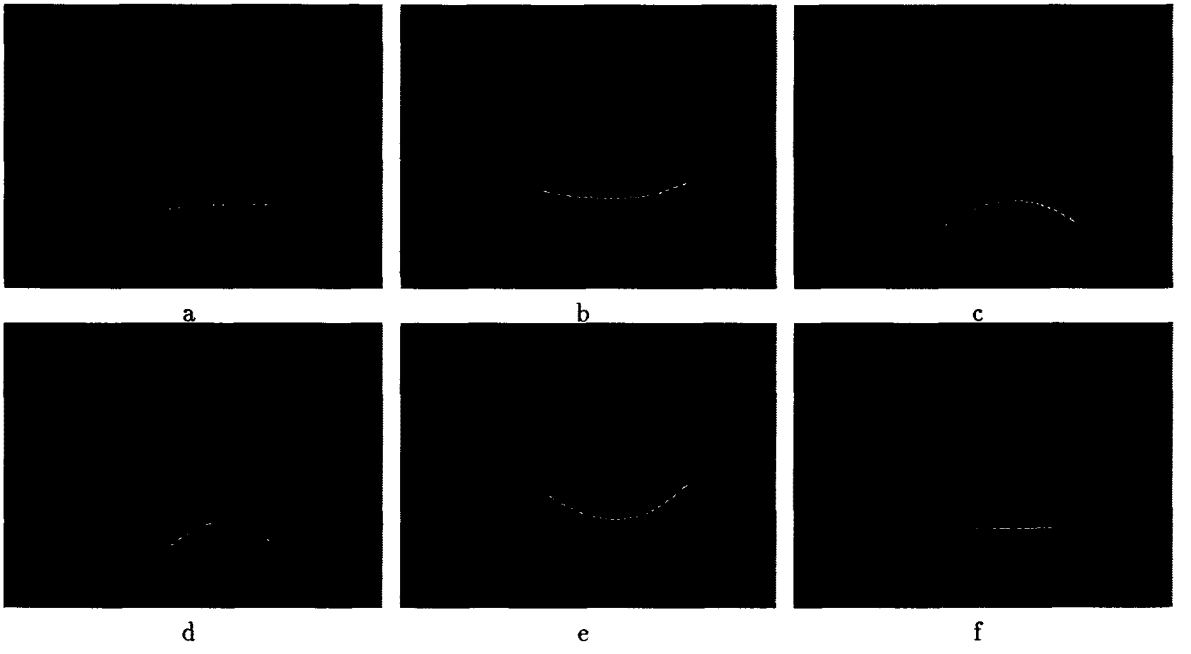


Figure 2: The tracked spline position for six different expressions: (a) *neutral*, (b) *smile*, (c) *sad*, (d) *open*, (e) *ee*, (f) *oo*.

of the spline's current position is calculated from the measurements and a prediction of the position based on the previous history of the contour. The basic tracker framework is that of Blake *et al* (1993) and the particular implementation used is briefly described here.

The curve to be tracked is represented by a set of parametric B-splines with N spans. In this implementation an open quadratic spline was used. The B-spline representation allows the curve to be characterised by a number of control points, N_c, where $N_c = N + 2$. The curve is related to the control points by equations of the type:

$$x(s,t) = B(s)X(t) \text{ and } y(s,t) = B(s)Y(t)$$

where t represents motion through time, s parameterises the B-spline curve and $B(s)$ is the usual B-spline matrix. It is assumed for tracking purposes that the image of the object to be tracked may be adequately described by the B-spline and its control points, $X(t)$ and $Y(t)$. A template is also required to enable successful tracking. This is equivalent to the steady state position of the contour, and is written as $\bar{X}$ and $\bar{Y}$.

To increase the stability of the tracker and improve the speed of the algorithm, we used the six dimensional affine subspace of the template to describe the possible contour deformations. It is clear that the range of deformations of the mouth do not belong to the affine subspace. However, practical experience indicates that the affine space is sufficient to represent all of the deformations that the contour makes provided that the original template is not a straight line.

The 6 dimensional affine vector Q can be related to the control point space by the matrices W and M in the following way

$$Q = M \begin{pmatrix} X \\ Y \end{pmatrix} \text{ and } \begin{pmatrix} X \\ Y \end{pmatrix} = WQ$$

The motion of the spline is represented by a second order stochastic differential equation. This is discretised to give an expression for the motion of the form.

$$\mathcal{X}_{n+1} - \bar{\mathcal{X}} = A(\mathcal{X}_n - \bar{\mathcal{X}}) + \begin{pmatrix} 0 \\ \omega_n \end{pmatrix}$$

Where the noise term ω_n is Gaussian. The vector $\mathcal{X}$ is defined in the following way:

$$\mathcal{X} = \begin{pmatrix} Q \\ \dot{Q} \end{pmatrix}$$

and A is derived by discretising a matrix of the form:

$$F = \begin{pmatrix} 0 & I \\ F_0 & F_1 \end{pmatrix}$$

The measurement process is similar to that used by Blake *et al* (1993). Measurements are made along rays

perpendicular to the spline, and a validation gate mechanism of the type proposed by Bar-Shalom (1988) is implemented. This limits the length of each ray to be proportional to the error covariance projected along the perpendicular to the spline. However, rather than detecting edges, the measurement process tests for the presence of a valley along each perpendicular. Since most of the spline is itself horizontal, this gives very little information about the horizontal positioning of the spline — the aperture problem [7]. To counteract this an "end of valley" detector was used at the ends of the spline.

The Kalman filter evolves in two stages, prediction and measurement. For each stage the means and covariance of the state estimates are updated. The sequential version of the Kalman filter was used [1]. The basic idea is to process each of the measurements sequentially, and update the covariance matrices accordingly. If a measurement fails, then no update is made.

Initially the dynamics were chosen in a similar way to those in Blake *et al* (1993). This produced a tracker that worked reasonably well. The data produced by this tracker was then used to bootstrap a learning process to generate better dynamics using Maximum Likelihood Estimation [3, 4]. Clearly better constraints on the dynamics will result in more robust tracking since unnatural deformations and dynamics of the contour will be avoided. Since we are interested in tracking the changing shape of the mouth as well as its location in the image, the training should include the dynamics of all the possible transformations of the lips. We collected sequences for the expression changes, as well as rotation about the vertical axis, rotation about a horizontal axis (nodding) and translation.

4 Robust Tracking

It is important for any tracking system to be robust to illumination, identification and viewing position as well as being able to track the typical expressions of an individual. This section demonstrates how successful our algorithm is. Firstly, it should be noted that all the tracking experiments were run at 50 Hz on a standard SUN SPARC 2 with a Datacell S2200 frameboard.

Tests sets were created out of the following six expressions. A *neutral* face (Fig. 2a). A closed mouth *smile* with the corners of the lips drawn upwards (Fig 2b). A *sad* face with the corners of a closed mouth turned down as far as possible (Fig. 2c). An *open* mouth with the lower jaw dropped as far as possible (Fig. 2d). Open smile, *ee*, similar to the smile, but with the teeth visible (Fig. 2e). Pursed lips, *oo*, an exaggerated version of the "oo" sound (Fig 2f). A typical set would contain about 1000 frames and would consist of the subject moving from one expression to another and then holding each expression for about 100 frames. Test sets of increasing difficulty could be created by changing expression more rapidly.

The initial test sequences were created with a single stationary subject under standard office lighting from a viewpoint directly in front of the subject. The tracking of the 6 different mouth shapes is shown in Fig. 2(a-f). The system was shown to be robust to changes in illumination condition without any parameter tuning (Fig. 3(a-c)). The system needed no extra training to cope with a change of viewing position as long as the mouth location in the image was relatively fixed (Fig3(d)). When the mouth location changed on the image, additional learning was required. The learning was based on two sequences of purely translational and purely rotational motion. The effect of translational motion tracking is shown in Fig. 3(e). Finally it is important for a tracker to work successfully with a range of subjects. Since the template is so simple, it can be used, unchanged, on a variety of subjects (Fig. 3(f)).

5 Classification of the Valley Shape

A system that will track and analyse the shape of an individual's mouth can be used to support speech recognition [12], and to analyse facial expressions. In this section we investigate the amount of information that the valley contains.

Classification algorithms were tested on the same expressions used in the previous section. However, an initial examination of the data indicated that the difference between the *ee* and the *smile* were negligible, and these two expressions were joined together for the purposes of classification. This highlights one obvious failing of a single contour; subtleties of expression will not be distinguished when the differences are due to the mouth being open or the teeth being visible.

Based on the 4 parameters of the affine transformation (no translational parameters were considered), we classified the shape of the valley using a linear approximation to the Bayesian classifier [8]. The basic postulate is that each class has a fixed mean and covariance in the 4-dimensional space. In the linear approximation to the Bayesian classifier, it may be assumed that the covariance of each class (*smile, sad, etc*) is identical. A measurement is classified by determining the class that it lies closest to.

Data was collected in bursts lasting around 20 seconds (around 1000 frames). The expressions were captured in the following order: *neutral, smile, neutral,*

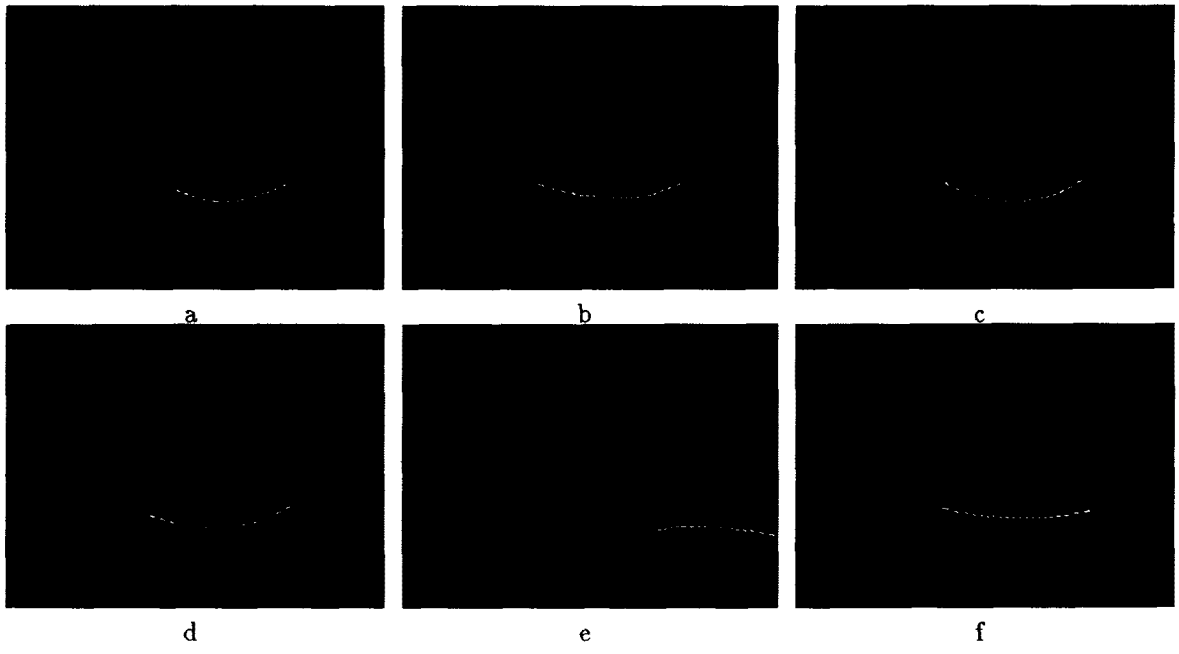


Figure 3: This figure demonstrates the robustness of the tracker to varying illumination (a-c), viewpoint change (d-e) and identity (f).

		Predicted Class				
		neu	smi	sad	open	oo
actual class	neu	1138	0	2	2	9
	smi	56	346	0	0	0
	sad	26	0	232	0	0
	open	26	0	0	212	0
	oo	52	0	0	0	255

Table 1: Confusion Matrix for Linear Discriminant. Note that the classes can effectively be distinguished perfectly, since expressions are only ever confused with the neutral expression.

sad, neutral, open, neutral, ee, neutral, oo, neutral. [1]. Two independent samples were collected. This meant that one set could be used for training purposes, and one for testing. It was assumed that the expression should be independent of translation and consequently only the four affine parameters were used for classification. The classifier worked remarkably well on the test data. Since the data was trained on one set and tested on another, two sets of results may be obtained, one for each training set. Table 1 presents the sum of the two confusion matrices. It shows that classification is perfect if the confusion with the neutral expression is disregarded.

This type of classifier shows that it is possible to distinguish between different features using our approach. Although the training must be done off-line, and presumably independently from one subject to another, the technique could be used as an on-line recognition system without any additional hardware.

So far, we have only studied the classification problem from a single viewpoint — facing forward. We also investigated the problem of determining a changing viewpoint as well as expression simultaneously. Preliminary studies indicate that using a similar Bayesian classifier, it is possible to classify both the valley shape and the viewpoint, in the case where just three viewpoints are considered — from the left, the right and straight ahead. However, to improve robustness we intend to study a system that will classify the face position first and then classify the mouth shape given an approximate viewpoint. Since the tracking performance has been shown to be independent of identity, it may be possible to construct a classifier that is also identity independent.

6 Conclusions and Further Work

In this paper we have suggested a framework to describe and track the deformations of the shape of the

[1] Other sequences with different orders were also examined and shown to produce similar results.

mouth which may be described by a valley contour lying between the lips. The contour was shown to exist independently of illumination, viewpoint, identity and expression. Additionally, the shape of the valley was shown to contain extractable information about the shape of the mouth. A 'learning' algorithm was implemented to improve the performance of the tracker, and it has been demonstrated that its effect is significant. The tracker was implemented in software and needs no specialised hardware beyond a camera and framegrabber which are becoming standard on several systems. The system will run in real time and is robust to lighting conditions, viewing angle and identity.

The valley contour contains useful information about the shape of the mouth. We have used the tracker to construct classifiers and demonstrate that the mouth shape can be classified in real time. We intend to introduce additional splines to improve classification. The first stage will be to add in extra splines to model the upper and lower lips. Later we propose to model more of the head to enable more accurate recovery of the mouth location and the viewing angle. Once these features have been implemented we will attempt to construct an improved classifier able to analyse a wide range of expressions.

Tracking in this manner has numerous applications such as identifying expressions and supporting speech recognition systems. When combined with an appropriate animation system it should be possible to compress the expression data for video communication. In the same way applications exist in the field of real time animation, currently only possible with costly specialised hardware. Finally, it might be possible to use information from this type of tracking to analyse the variability that occurs within the same face due to viewing conditions, and hence construct systems that recognise people. We believe that our system is a significant step towards a more generalised approach that will support such applications.

References

[1] B.D.O. Anderson and J. B. Moore. *Optimal Filtering.* Electrical Engineering Series. Prentice Hall, 1979.

[2] Y. Bar-Shalom and T.E. Fortmann. *Tracking and Data Association.* Academic Press, 1988.

[3] A. Blake and M. Isard. 3d position, attitude and shape input using video tracking of hands and lips. In *Proc. ACM SigGraph Conference*, pages 185–192, Orlando, USA, 1994.

[4] A. Blake, M. Isard, and D. Reynard. Learning to track the visual motion of contours. *J. Artificial Intelligence, in press.*, 1995.

[5] I.A. Essa, T Darrell, and A. Pentland. Modeling and ineractive animation of facial expressions using vision. Technical Report No. 256, M.I.T Media Laboratory Perceptual Computing Section, 1994.

[6] A. Gee and R. Cipolla. Determining the gaze of faces in images. *Image and Vision in Computing*, 12:639–648, 1994.

[7] B.K.P. Horn and M.J. Brooks. The varaitional approach to shape from shading. *Computer Vision, Graphics and Image Processing*, 33:174–208, 1986.

[8] M. James. *Classification Algorithms.* Collins, 1985.

[9] D.E. Pearson, E. Hanna, and K. Martinez. Computer-generated cartoons. In H. Barlow, C. Blackemore, and M. Weston-Smith, editors, *Images and Understanding.* Cambridge Univ. Press, NY, 1990.

[10] I. Pilowsky, M. Thornton, and B.B. Stokes. Towards the quantification of facial expressions with the use of a mathematics model of the face. In H.D. Ellis, M.a. Jeeves, F. Newcombe, and A. Young, editors, *Aspects of Face Processing*, pages 340–348. 1986.

[11] J. Shepherd, G. Davies, and H. Ellis. Studies of cue saliency. In G. M. Davies, H.D. Ellis, and J. Shepherd, editors, *Perceiving and Recognizing Faces*, pages 104–131. Academic Press, 1981.

[12] Quentin Summerfield. Some preliminaries to a comprehensive account of audio-visual speech perception. In Barbara Dodd and Ruth Campbell, editors, *Hearing by Eye: The Psychology of Lip Reading.* Lawrence Erlbaum, 1987.

[13] D. Terzopoulos and K. Waters. Analysis and synthesis of facial images using physical and anatomical models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15:569–579, 1993.

[14] K. Waters. A muscle model for animating three-dimensional facial expression. *Computer Graphics*, 21:17–24, 1987.

[15] A. L. Yuille, D.C. Cohen, and P.W. Hallian. Feature extraction from faces using deformable templates. In *Proc. CVPR-89*, San Diego, CA, 1989.

Acknowledgements Thanks to Rupert Curwen, Nicola Ferrier, Andrew Wildenberg, Simon Rowe, Nick Pillow and Sue Johnson for their willingness to participate in our experiments and assistance in constructing our software. We acknowledge the financial support of the BBSRC, the Clore Foundation and the British Council.