

Automatic feature localisation with constrained local models

David Cristinacce*, Tim Cootes

Department of Imaging Science and Biomedical Engineering, University of Manchester, Manchester M13 9PT, UK

Received 5 November 2007; accepted 20 January 2008

Abstract

We present an efficient and robust method of locating a set of feature points in an object of interest. From a training set we construct a joint model of the appearance of each feature together with their relative positions. The model is fitted to an unseen image in an iterative manner by generating templates using the joint model and the current parameter estimates, correlating the templates with the target image to generate response images and optimising the shape parameters so as to maximise the sum of responses. The appearance model is similar to that used in the Active Appearance Models (AAM) [T.F. Cootes, G.J. Edwards, C.J. Taylor, Active appearance models, in: Proceedings of the 5th European Conference on Computer Vision 1998, vol. 2, Freiburg, Germany, 1998.]. However in our approach the appearance model is used to generate likely feature templates, instead of trying to approximate the image pixels directly. We show that when applied to a wide range of data sets, our Constrained Local Model (CLM) algorithm is more robust and more accurate than the AAM search method, which relies on the image reconstruction error to update the model parameters. We demonstrate improved localisation accuracy on photographs of human faces, magnetic resonance (MR) images of the brain and a set of dental panoramic tomograms. We also show improved tracking performance on a challenging set of in car video sequences.

© 2008 Elsevier Ltd. All rights reserved.

Keywords: Shape modelling; Feature detectors; Object detection; Object localisation; Face detection; Active appearance models; Constrained local models

1. Introduction

This paper addresses the problem of automatically finding feature points for a given object in an image. The aim is to build a generic model of a class of object, so that the model can fit to any new instance of the object automatically.

For example with human faces, locating feature points such as the eye pupils and mouth corners is important for many tasks such as face recognition and automatic avatar generation. An accurate local search method is useful to track faces in a video sequence and can be used to aid face behaviour analysis. In medical images, extraction of feature points is critical for making accurate measurements to aid diagnosis and tracking disease progression over time.

A standard approach to this type of problem is to collect a manually labelled training set of images that enable the model to learn the shape and texture variation typically present in

an object class. For example the Pictorial Structure Matching (PSM) approach of Felzenszwalb and Huttenlocher [1] learns detectors for a set of manually labelled points and a tree structure for the spatial relationships between selected pairs of features. This leads to an efficient dynamic programming algorithm for combining feature detection responses, which is useful for global image search and initialisation. However for local search the PSM tree structure is less specific compared to approaches which use the full shape model (e.g. Ref. [2]).

The Active Appearance Model (AAM) [3] is a local search method which combines the full shape model and texture variation learnt from a training set. However the AAM search method relies on predicting model parameters from the residual of the current model and the underlying image. This approach can be prone to local minima, which prevents the model from finding the global optimum in some cases. In the following we introduce the Constrained Local Model (CLM) approach which combines the power of feature detection based approaches (e.g. Ref. [1]), the flexibility of appearance based models [3] and the constraints of a full shape model [2]. The CLM approach

* Corresponding author. Tel.: + 44 161 275 1243.

E-mail address: david.cristinacce@manchester.ac.uk (D. Cristinacce).

learns the variation in appearance of a set of template regions. The template regions are then used as feature detectors in a local search, constrained by the full shape model.

The CLM is matched to new instances of an object using an iterative template generation and shape constrained search technique. Given current image points, the template generation proceeds by fitting the joint model of shape and appearance to regions sampled around each feature point. The current feature templates are then applied to the search image using normalised correlation. This generates a set of response surfaces. The quality of fit of the model is optimised using the Nelder–Mead simplex algorithm [4] to drive the parameters of the shape model in order to maximise the sum of responses at each point. Given a new set of candidate feature locations the templates are regenerated and the search proceeds iteratively.

This CLM approach, summarised in Fig. 2, is shown to be robust, computationally efficient and provide superior tracking performance compared to the AAM matching method [3], when applied to human faces. The CLM was first described in Cristinacce and Cootes [5]. This journal paper describes the technique in more detail, gives a Bayesian interpretation of the matching function, includes displacement experiments and demonstrates performance on several new data sets, including medical images. The CLM is shown to be more accurate and have a wider radius of convergence compared to the AAM when applied to magnetic resonance (MR) brain images, dental panoramic tomograms and human faces.

2. Background

There are many examples of computer vision techniques that combine both shape and texture to build models and match to unseen images [1–3,6–10]. Given an approximate localisation of an object (either segmented manually or found automatically using a global detector) we would like to automatically locate prominent internal features on the object of interest.

There are broadly two different approaches to this problem. The first approach fits a generative model to the region of interest. The best match of the model simultaneously calculates feature point locations. Examples of this approach are the AAM [3] and 3D-Morphable Model [11]. The second approach is to split the object into separate subregions that can be found using feature detection methods, with constraints on the relative configuration of feature points. For example Active Shape Model (ASM) fitting [2] and PSM [1].

A popular example of the generative approach is the AAM algorithm [3], which uses a combined model of shape and texture. A Principal Components Analysis (PCA) model of shape is learnt from a set of manually labelled images and also a model of the triangulated texture variation across the training set. A joint model of the shape and texture parameters allows the generation of new object instances, which resemble the training set. The AAM then searches new images by using the texture residual between the model and the target image to iteratively update model parameters to provide an estimate of the current image.

A related method is the 3D-Morphable Model due to Vetter et al. [11]. This method constructs a 3D model of the whole head from textures and 3D vertices obtained using a laser range finder. PCA is applied to the 3D coordinates and surface texture to build the generative model. The Morphable model is then fitted to new 2D images using a coarse to fine correspondence method based on optical flow. The method requires a few key points to be hand labelled to produce the dense correspondence from 3D to 2D. However recent approaches aim to mitigate the requirement for manual intervention. For example Romdhani et al. [12] use SIFT features [13], an appearance based rejection and a projection constraint to initialise the 3D-Morphable Model automatically.

The generative approach has been expanded in many ways by different researchers. Matthews et al. [14] have developed a more efficient update scheme which utilises the inverse compositional update algorithm [15]. Xiao et al. [16] have developed hybrid 2D–3D versions of the AAM and applied them to human faces. In the medical domain, AAMs have been extended to 3D by Mitchell et al. [8] and applied to cardiac image volumes. van Ginneken et al. [9] compare the AAM and ASM with pixel based segmentation and also investigate hybrid approaches that improve on the normal pixel based AAM. Scott et al. [17] extend the texture sampling part of the algorithm to edge and cornerness values, instead of the normalised pixel values used in the original AAM. Scott et al. demonstrate significantly improved results on faces and spinal images. Therefore the edge/corner formulation of the AAM is used in this paper.

A recent alternative to the AAM algorithm that has been applied to both human face photographs and cardiac images is the RankBoost approach to shape prediction described by Zheng et al. [18]. This method uses Rankboost [19] to rank the possible image warpings from the mean shape to the an unseen image and thus compute feature points. Zheng et al. presents good results compared to the normalised texture AAM on manually cropped images. However the normalised texture version of the AAM is known to be inferior to the edge/corner AAM [17] used in this paper.

An example of the feature based approach to model matching is the ASM algorithm [2]. This method learns a statistical model of shape from manually labelled images and also PCA models of patches around individual feature points. When applied to an unseen image the best local match of each feature is found and the shape model fitted to the updated points. Therefore individual false detections which invalidate the learnt shape configuration are avoided. The search proceeds iteratively until the feature points converge.

An elegant method of combining feature responses and shape constraints is due to Felzenszwalb and Huttenlocher [1]. This PSM approach is very efficient due to the use of pairwise constraints and a tree structure. The method uses the whole response surface of each detector and a dynamic programming search to find the global optimum set of final feature locations. However the PSM is mainly a global search method and does not use a full shape model. Therefore locally it is less accurate compared to methods that use the full

shape model (e.g. ASM). Recently a denser graph matching approach using k -fans has been suggested to provide a stronger shape constraint [20]. The single tree based PSM is; however, very useful to provide initialisation points for local search using AAM or CLM given a rough localisation of the object.

Another method that combines shape and feature detection is the SMAT algorithm described by Dowson and Bowden [7]. The SMAT method tracks an object given an initialisation and generates new templates from a clustered set of templates sampled from previous frames. It uses a shape model to constrain feature configurations, but does not form a combined model of shape and texture. In contrast, the CLM generates appropriate templates using an appearance model, learnt from a fixed training set. Therefore CLM is unable to generate false templates (from false matches) and can be applied to search static images, not just video, but at the expense of requiring a manually labelled training set.

An extension of the ASM approach is the Shape Optimised Search (SOS) method due to Cristinacce et al. [21]. The SOS computes the response surface around each individual feature point, instead of merely computing the best response (as in the ASM). The model is then fitted to this set of response surfaces by allowing the shape parameters to vary whilst optimising the sum of feature responses (using the Nelder–Mead simplex optimiser [4]). Limits on the shape parameters enforce the relative shape constraint and utilising the whole response surface provides more robustness compared to the ASM approach [21].

The contribution of this paper is a method known as the CLM which combines the generative and feature based approaches described above. The CLM learns a model of shape and texture variation from a labelled training set (similar to the AAM). However, the texture is sampled in patches around individual feature points. Given a set of feature locations the CLM generates a set of feature detectors, which resemble the underlying image, but are constrained to be similar in form to the training set. The feature detectors are then applied to the underlying image, to compute local response surfaces (similar to the SOS method) and the shape parameters optimised to update the feature locations, subject to the global shape constraint. In Ref. [6] a similar method is described; however, the feature templates are updated by selecting example templates from the training set using a nearest neighbour approach, which is not as general as the statistical model used by the CLM.

We compare the local search accuracy of the CLM and AAM on three different data sets using displacement experiments. This provides localisation accuracy results independent of global search (see Section 5). In the case of faces we test fully automatic localisation of facial features using the Viola–Jones face detector [22] to find the face in the image. Within the detected face region we apply smaller Viola and Jones feature detectors constrained using the PSM algorithm [1], to compute initial feature points. We then refine these feature points using CLM local search and compare the results with the AAM [3] algorithm.

3. Algorithm

3.1. Constrained local appearance models

A joint shape and texture model is built from a training set of manually labelled images (see Fig. 3 for examples) using the method of Cootes et al. [2]. This is similar to the AAM; however, the texture sampling method is different. A training patch is sampled around each feature and normalised such that the pixel values have zero mean and unit variance.¹ The texture patches from a given training image are then concatenated to form a single intensity vector. The set of intensity training vectors and normalised shape co-ordinates are used to construct linear models, as follows:

$$\mathbf{x} = \bar{\mathbf{x}} + \mathbf{P}_s \mathbf{b}_s, \quad \mathbf{g} = \bar{\mathbf{g}} + \mathbf{P}_g \mathbf{b}_g, \quad (1)$$

where $\bar{\mathbf{x}}$ is the mean shape, \mathbf{P}_s is a set of orthogonal modes of variation and \mathbf{b}_s is a set of shape parameters. Similarly $\bar{\mathbf{g}}$ is the mean normalised intensity vector, \mathbf{P}_g is a set of orthogonal modes of variation and \mathbf{b}_g is a set of texture model parameters. The shape and template texture models are combined using a further PCA to produce one joint model. The joint model has the following form:

$$\mathbf{b} = \mathbf{P}_c \mathbf{c},$$

where

$$\mathbf{P}_c = \begin{pmatrix} \mathbf{P}_{cs} \\ \mathbf{P}_{cg} \end{pmatrix} \quad \text{and} \quad \mathbf{b} = \begin{pmatrix} \mathbf{W}_s \mathbf{b}_s \\ \mathbf{b}_g \end{pmatrix}. \quad (2)$$

Here \mathbf{b} is the concatenated shape and texture parameter vector, with a suitable weighting \mathbf{W}_s to account for the difference between shape and texture units (see Ref. [3]). \mathbf{c} is a set of joint appearance parameters. \mathbf{P}_c is the orthogonal matrix computed using PCA, which partitions into two separate matrices \mathbf{P}_{cs} and \mathbf{P}_{cg} which together compute the shape and texture parameters given a joint parameter vector \mathbf{c} .

By varying the first two parameters of \mathbf{c} the first two modes of variation for the joint appearance model can be computed, as shown in Fig. 1. In the rest of this section the PCA models and matrices in Eq. (2), learnt from the training set are collectively referred to as joint model Θ .

3.2. Template generation

Suppose we have a set of initial feature locations, an image \mathbf{I} and the joint model Θ learnt from the training set (see Section 3.1). Let (X_i, Y_i) be the position of feature point i . The positions can be concatenated into a vector \mathbf{X} ,

$$\mathbf{X} = (X_1, \dots, X_n, Y_1, \dots, Y_n)^T. \quad (3)$$

¹ The regions from the training images are re-sampled to a fixed sized rectangle to allow for scale changes.



Fig. 1. PCA modes of combined shape and texture variation for CLM face model ($\pm 3\text{std}$).

The feature locations \mathbf{X} can be approximated by fitting the shape model (see Eq. (1)) using the iterative procedure described by Cootes et al. [2]. This gives an approximation to the current feature points in terms of a similarity transform S_t and shape parameters \mathbf{b}_s as follows:

$$\mathbf{X} \approx S_t(\bar{\mathbf{x}} + \mathbf{P}_s \mathbf{b}_s). \quad (4)$$

The parameters of the similarity transform, S_t and, shape parameters \mathbf{b}_s are concatenated into a single vector $\mathbf{s} = (\mathbf{t}^T | \mathbf{b}_s^T)^T$.

The current templates \mathbf{T}_i can be generated from the joint model θ using the texture patches from the image \mathbf{I} at points \mathbf{X}_i to estimate the best fitting parameters \mathbf{b}_g (see Ref. [3]). The joint parameters \mathbf{c} can also be estimated. Then the model shape and texture are generated using Eqs. (1) and (2). The texture templates \mathbf{T}_i are recorded for use in the search update step, see Section 3.3.

3.3. Shape constrained search update

Given the current set of shape parameters \mathbf{s} the search update step relies on optimising an objective function $f(\mathbf{s})$ which takes into account the current feature templates \mathbf{T}_i , the current image \mathbf{I} and the joint model θ described in Section 3.1. The optimal transformation \mathbf{s} then defines the new updated points \mathbf{X} .

A Bayesian approach to deriving this objective function $f(\mathbf{s})$ is to define $f(\mathbf{s}) = p(\mathbf{X}|\mathbf{I}, \theta)$ the probability of finding the object at location \mathbf{X} given image \mathbf{I} and model θ . Therefore following Bayes rule:

$$p(\mathbf{X}|\mathbf{I}, \theta) \propto p(\mathbf{I}|\mathbf{X}, \theta)p(\mathbf{X}|\theta), \quad (5)$$

where $p(\mathbf{I}|\mathbf{X}, \theta)$ is the probability of the image \mathbf{I} given a location \mathbf{X} and model θ . $p(\mathbf{X}|\theta)$ is the probability of a configuration of the object \mathbf{X} given model θ . The aim is to find the optimum location \mathbf{X}' which maximises Eq. (5).

In our approach $p(\mathbf{I}|\mathbf{X}, \theta)$ is modelled using a matching score between the current set of region templates \mathbf{T} and the current image \mathbf{I} . The probability of a set of image subregions matching the current region templates \mathbf{T}_i can be estimated using a Gibbs–Boltzmann distribution

$$p(\mathbf{I}|\mathbf{X}, \theta) \propto \prod_{i=1}^n e^{-\alpha q_i}, \quad (6)$$

where n is the number of templates, q_i is the match quality between template \mathbf{T}_i and image \mathbf{I} at point (X_i, Y_i) and α is a normalising constant.

$p(\mathbf{X}|\theta)$ is the probability of a given shape as estimated by the statistical shape model in Eq. (1). We follow the work of Dryden et al. [23] and assume that the b_j shape parameters in Eq. (1) are independent and Gaussian distributed. Therefore the probability of a given shape is

$$p(\mathbf{X}|\theta) \propto \prod_{j=1}^k e^{-b_j^2/\lambda_j}, \quad (7)$$

where k is the total number of shape parameters, b_j are the elements of \mathbf{b}_s and λ_j are the corresponding eigenvalues of the shape model. Inserting Eqs. (6) and (7) into Eq. (5) and taking logs gives

$$\log p(\mathbf{X}|\mathbf{I}, \theta) = \sum_{i=1}^n -\alpha q_i + \sum_{j=1}^k \frac{-b_j^2}{\lambda_j} + \text{const.} \quad (8)$$

In the experiments below, the matching scores q_i in Eq. (8) are pre-computed by correlating the current templates T_i with the search image I . Let (X_i, Y_i) be the position of feature point i and $q_i = -R_i(X_i, Y_i)$ is the normalised correlation response of the i th feature template at that point. Then Eq. (8) can be re-written as an objective function $f(\mathbf{s})$ which is dependent on the shape parameter \mathbf{s} as follows:

$$f(\mathbf{s}) = \alpha \sum_{i=1}^n R_i(X_i, Y_i) - \sum_{j=1}^k \frac{b_j^2}{\lambda_j}. \quad (9)$$

Therefore when updating the points, the current feature templates are correlated with the image in a local neighbourhood around each feature point i to produce a set of response images $R_i(X_i, Y_i)$, one for each feature. Then the objective function $f(\mathbf{s})$ is optimised by allowing the transformation parameters \mathbf{s} to vary. This procedure searches for peaks in the response surfaces $R_i(X_i, Y_i)$, whilst taking into account the likely shape of the object learnt from the training set, due to the second shape penalty term in Eq. (9).

A suitable value α in Eq. (9) can be determined by computing the ratio of $\sum_{i=1}^n R_i(X_i, Y_i)$ and $\sum_{j=1}^k \frac{b_j^2}{\lambda_j}$ when applied to a verification set with human labelled ground truth, as part of the training stage.

The optimisation of $f(\mathbf{s})$ is performed using the Nelder–Mead simplex algorithm [4]. It starts from the initial feature points and terminates when the parameter changes are less than a small positive constant. The Nelder–Mead simplex was chosen because of its robustness to local minima; however, any non-linear optimiser could be used (for example gradient descent). Multiple starting locations could also be evaluated. However in the interests of efficiency a single Nelder–Mead optimisation is used in this paper.

3.4. Search algorithm

The CLM search algorithm (see Fig. 2) combines the methods described in Sections 3.1–3.3 and proceeds as follows:

- (i) Input an initial set of feature points.
- (ii) Repeat:
 - (a) Fit the joint model to the current set of feature points (see Section 3.1).
 - (b) Generate a set of templates (see Section 3.2).
 - (c) Use the shape constrained search method to predict a new set of feature points (see Section 3.3).
- (ii) until converged.

When tracking the initial points are propagated from the previous frame. On a new sequence (or if tracking fails) a global search can be used.

3.5. Differences between CLM, TST, SOS and AAM methods

The search procedure described in Section 3.3 is fundamentally different from the AAM [3]. However the joint model of

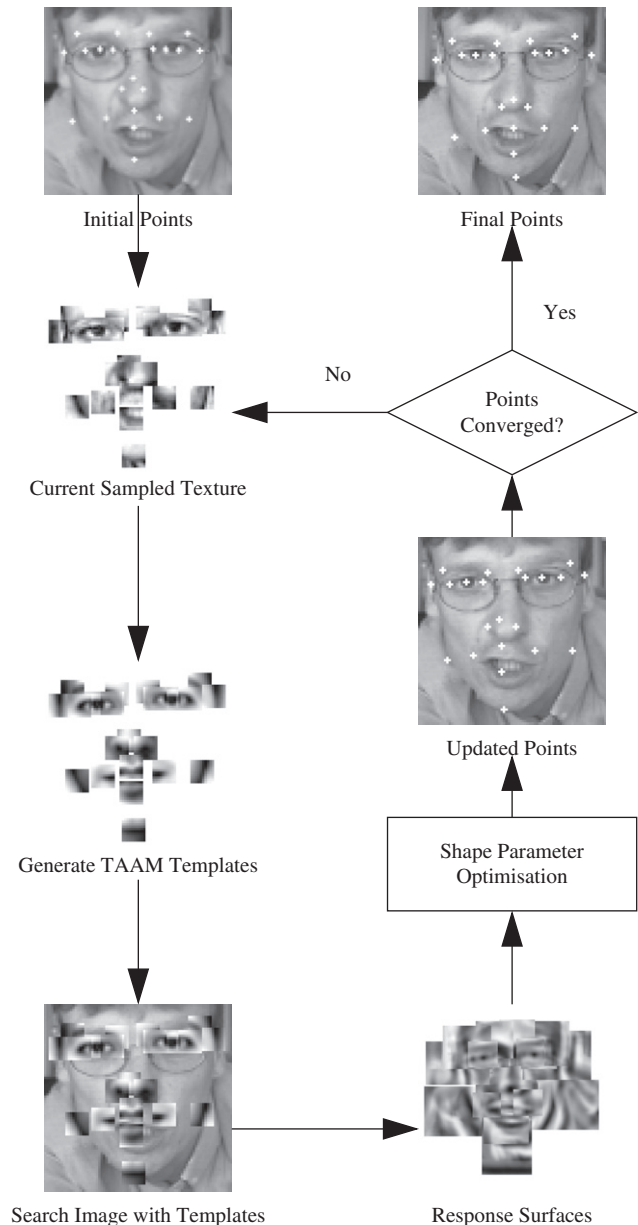


Fig. 2. Constrained Local Model (CLM) search algorithm.

appearance and texture (see Section 3.1) has the same form as the AAM. The CLM appearance model takes the form of rectangular regions around each feature, which allows the non-linear shape constrained search described in Section 3.3.

The CLM local model search is the same as described in the author's previous work [21,6]. However the Template Selection Tracker (TST) method [6] uses nearest neighbour matching to select the most appropriate templates. The CLM uses a joint model fitted to the current feature points to generate templates. In contrast the SOS method [21] uses fixed templates, which do not change during search.

In Ref. [5] the CLM is shown to be a superior version of the TST and SOS. Therefore in this paper, only the CLM algorithm is compared with the AAM.

4. Data sets

We build AAM and CLM models for three different types of image data (see Fig. 3). The localisation accuracy of the algorithms are compared using displacement experiments (see Sections 5.3–5.5).

The three different types of model are built from three different types of image data, as follows:

4.1. Annotated data

- MR (magnetic resonance) brain slices—A 123 point markup scheme labelling the ventricles, caudate nucleus and lentiform nucleus, see Fig. 3(a). The data set consists of 69 images, which are split into 35 training and 34 test images. The image data set was originally described in Ref. [3].
- Dental Panoramic Tomograms of the Human Jaw—A 78 point markup scheme marking the lower jaw (or mandible), see Fig. 3(b). The data set consists of 134 images, which are split into 67 training and 67 test images (described in Ref. [24]).
- Photographs of human faces—A 22 point markup scheme as shown in Fig. 3(c). About 1052 face images were collected in our lab and used for training. The model was tested on the publicly available BIODID [25] and XM2VTS [26] data sets. The images shown in Fig. 3 are all manually labelled by

human operators (qualified clinicians in the case of the brain and dental data). The manual points are used for model building during training and provide ground truth during testing.

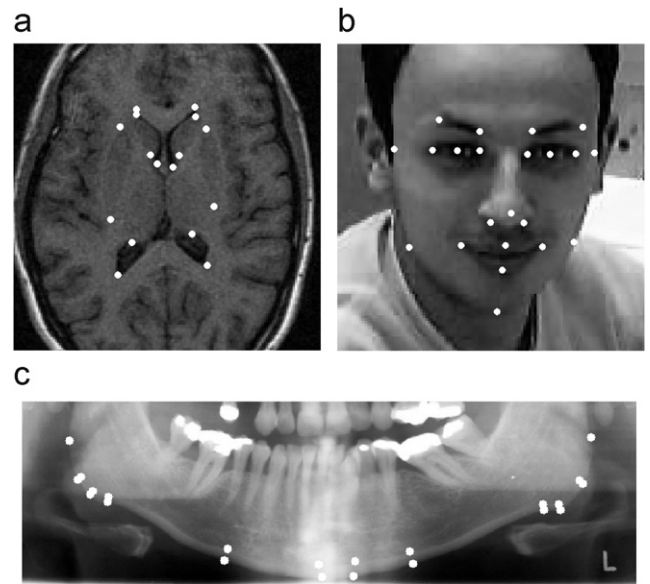


Fig. 4. Subset of marked up points used to train CLM models: (a) Brain images—CLM 16 points, (b) face images—CLM 22 points and (c) dental images—CLM 22 points.

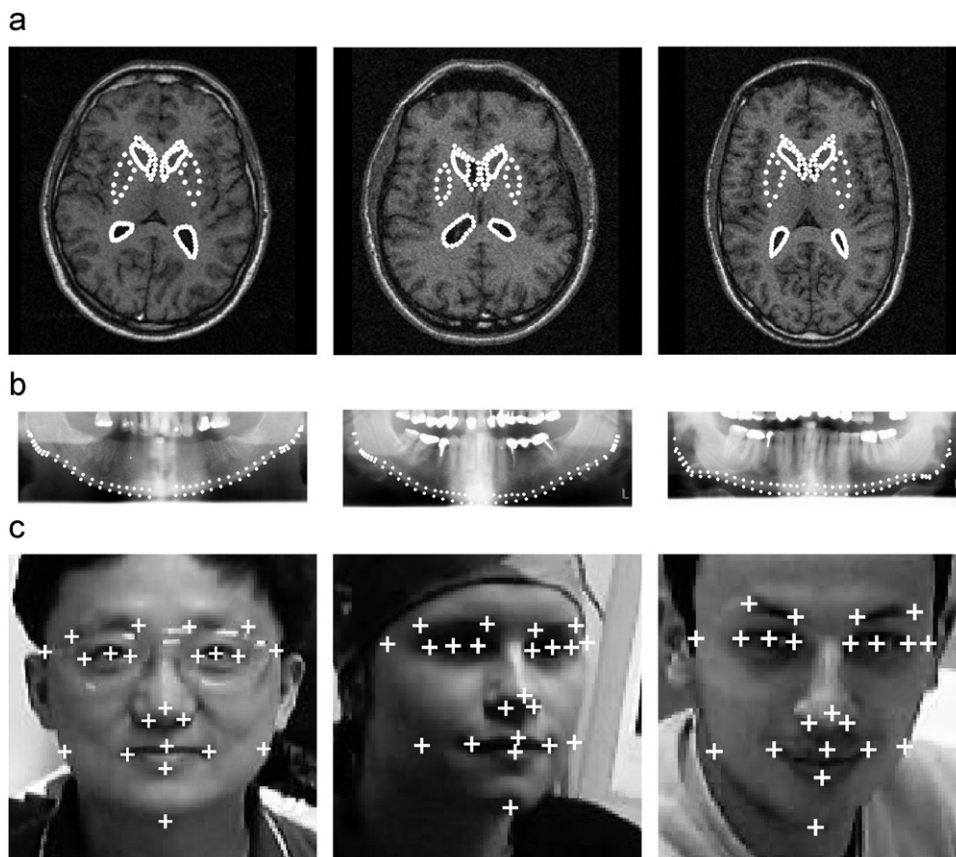


Fig. 3. Examples from three manually labelled training sets: (a) Brain Images—123 points, (b) dental images—78 points and (c) face images—22 points.

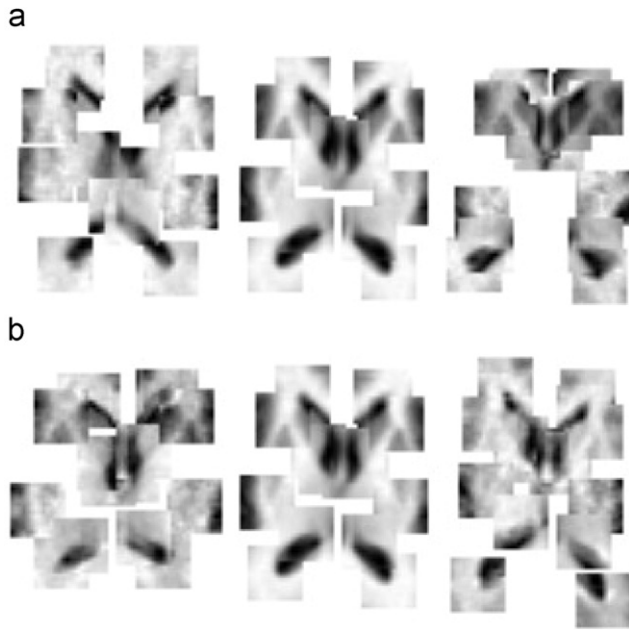


Fig. 5. First two PCA modes of joint shape and texture variation for CLM brain model (± 3 std).

The training and test sets are independent for all three data types.

4.2. Model specifications

The original markups for the brain and dental images contain many feature points which are required for the AAM to model the shape variation effectively (see Fig. 3). However, this large number of points is not required for the CLM. Therefore a subset of features are selected to train the CLM as shown in Fig. 4 and the localisation accuracy of both algorithms evaluated on this subset of points during testing. The full set of points is used for AAM training.

The region sizes around each feature point were chosen so that the set of patches roughly covered the object. See Fig. 1 for the modes of the face CLM model and Figs. 5 and 7 for the medical image models. The resolutions of the CLM and AAM models were chosen such that the total number of pixels in each model was similar, around 3000 pixels for the models described here.

5. Experiments

5.1. Distance error measure

The criteria for success is the distance of the points computed using automated methods compared to manually labelled ground truth. The average error is given as

$$m_e = \frac{1}{ns} \sum_{i=1}^{i=n} d_i. \quad (10)$$

Here d_i are the Euclidean point to point errors for each individual feature location and s is the distance between a pair of reference feature points which determine the scale of the object in the image. The scale reference separations for each of the data sets are as follows:

- Brains—The most frontal points of the left and right lateral ventricle horns (markup points at top of Fig. 4(a)).
- Dental—The right and left ends of the lower jaw (see top left and top right points in Fig. 4(c)).
- Faces—The centre of the left and right eye pupils (see Fig. 4(b)).

Therefore the distance error measure is invariant to the variation in size of each individual face, brain or jaw image over each test set, which allows scaled comparison of point to point errors between test images. The eye separation varies from 34 to 104 pixels over the BIODID test set. The variation in reference separation for the brain test images is 27–36 pixels and for the higher resolution dental images the variation is 800–1114 pixels.

To remain consistent with [5] only the 17 reference points internal to the face are used in the face results. This is due to the five points on the exterior of the face being somewhat ambiguous when the head is rotated.

5.2. Design of experiments

The AAM and CLM algorithms were compared by fitting each model to the ground truth points of the test set. The shape and texture of each model were reset to the mean values learned from the training set. The model was then systematically displaced from the true locations in each of the eight possible directions and the search algorithm applied to the image. The cumulative distribution of the distance measure (see Eq. (10)) was then used to compare the AAM and CLM methods on the various image data sets, see Sections 5.3–5.5.

The face models were also compared by applying the Viola and Jones face detector [22] and PSM method due to Felzenszwalb and Huttenlocher [1] to provide starting locations for the CLM and AAM search, see Section 5.6. The full face system was also tested by applying a face detect/track system in Section 5.7. In all cases the distance measure specified by Eq. (10) was used to calculate search accuracy relative to manually labelled ground truth points.

5.3. MR brain image displacement experiments

Fig. 5 shows the first two modes of variation of the CLM brain model, built from the manually labelled points shown in Fig. 4(a).

Fig. 6(a) shows the cumulative distribution of the search accuracy when the CLM and AAM models are displaced by 20% of the separation between the left and right ventricle horns, eight times on each of the 34 brain test images (see Fig. 3a).

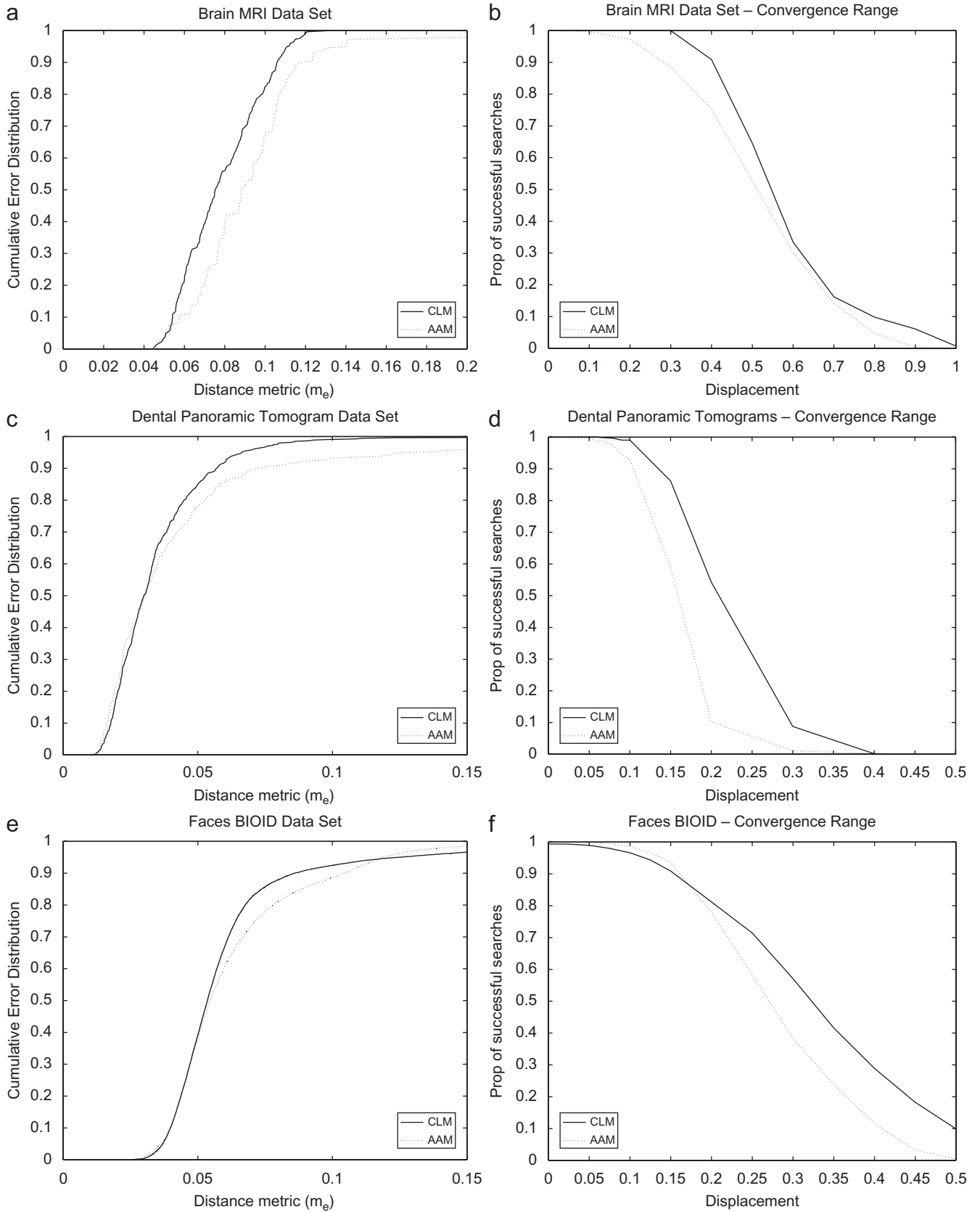


Fig. 6. Displacement results for CLM and AAM models when applied to test set. (a) Search accuracy after displacement of 20% on brain images. (b) Range of convergence region after various displacements on brain images (success if $m_e < 0.15$). (c) Search accuracy after displacement of 10% on dental images. (d) Range of convergence region after various displacements on dental images (success if $m_e < 0.10$). (e) Search accuracy after displacement of 10% on face images. (f) Range of convergence region after various displacements on face images (success if $m_e < 0.15$).

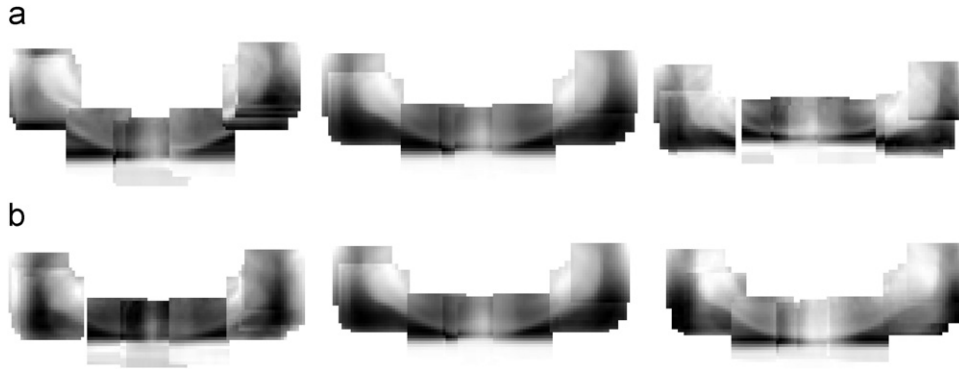


Fig. 7. First two PCA modes of shape and texture variation for CLM dental model ($\pm 3\text{std}$).

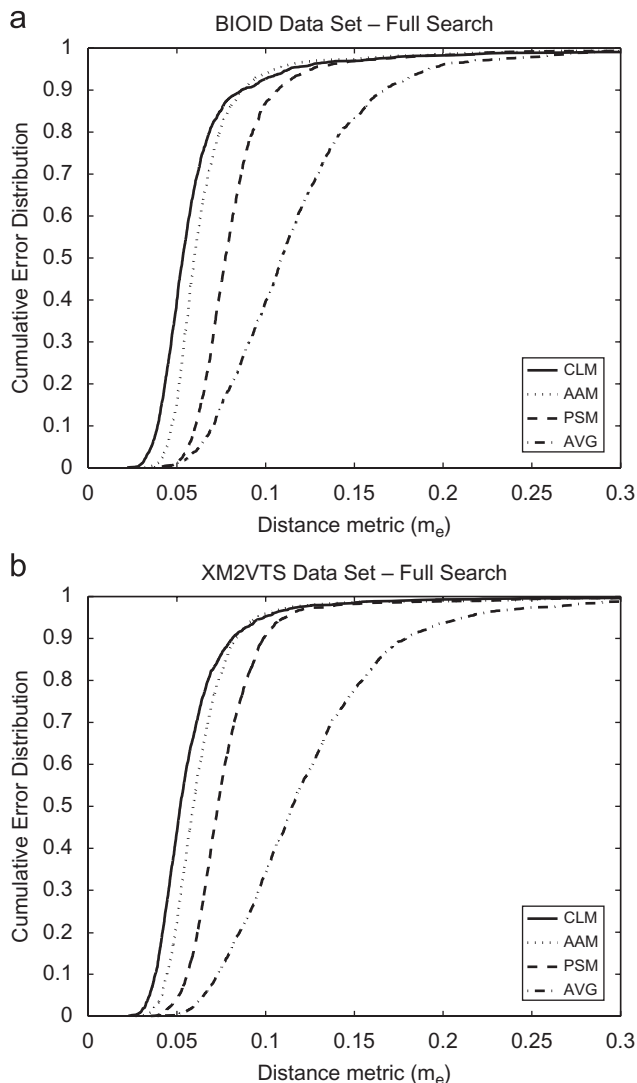


Fig. 8. Cumulative distribution of point to point measure (m_e) on BIOID and XM2VTS data sets. (a) BIOID results—CLM vs PSM vs AVG. (b) XM2VTS results—CLM vs AAM vs PSM vs AVG.

Fig. 6(a) shows that given a 20% displacement the CLM is more accurate than the AAM finding facial feature points with an accuracy of $m_e < 0.12$ in 100% of cases compared to 90% of cases for the AAM. The CLM also achieves an accuracy of

$m_e < 0.08$ in 55% of the brain test images compared to 40% of with the AAM. Therefore the CLM is more accurate given a starting displacement of 20%.

Fig. 6(b) shows the proportion of successful searches at different size displacements from the ground truth. For example the CLM finds 100% of brains accurately (i.e. $m_e < 0.15$) at a displacement of 30% compared to just 89% of brains using the AAM and hence the radius of convergence is larger for the CLM brain model.

5.4. Dental tomogram displacement experiments

Fig. 7 shows the first two modes of variation of the CLM dental model, built from the manually labelled points shown in Fig. 4(c).

Fig. 6(c) shows the cumulative distribution of the search accuracy when the CLM and AAM models are displaced by 10% of the jaw width, eight times on each of the 67 dental test images (see Fig. 3(b)). The CLM is more stable than the AAM because the point to point accuracy (m_e) is improved in 99% of cases using the CLM, whereas the AAM only reduces the point to point accuracy in 90% of cases given an initial displacement of 10%.

Fig. 6(d) shows that the range of convergence of the dental CLM model is also greater than the AAM. For example, given a displacement of 15% the CLM achieves a final point to point error of $m_e < 0.15$ in 86% of searches compared to 55% using the AAM.

5.5. Face image displacement experiments

Fig. 1 shows the first two modes of variation of the CLM face model, built from the manually labelled points shown in Fig. 4(b).

Fig. 6(e) shows the cumulative distribution of the search accuracy when the CLM and AAM models are displaced by 10% of the inter-ocular eye separation, eight times on each of the 1521 BIOID test images [25]. The CLM face model is more stable than the AAM because the point to point accuracy (m_e) is improved in 90% of cases using the CLM, whereas the AAM only reduces the point to point accuracy in 85% of cases given an initial displacement of 10%.

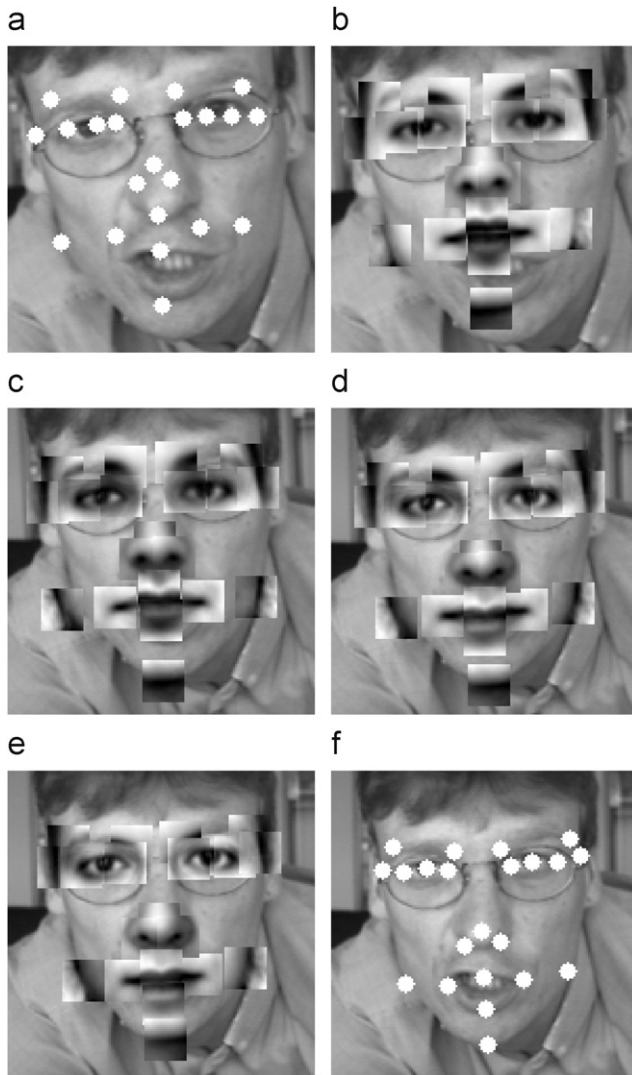


Fig. 9. Evolution of CLM templates when searching a static image. (a) Start points, (b) start templates, (c) after iteration 1, (d) after iteration 2, (e) after iteration 3 and (f) final points.

Fig. 6(f) shows the proportion of successful searches at different size displacements from the ground truth. For example the CLM finds 65% of faces accurately (i.e. $m_e < 0.15$) at a displacement of 30% compared to just 39% of faces using the AAM. This shows again that the range of convergence is larger for the CLM model.

Note that the point to point accuracy of the face models is lower for both the AAM and CLM compared to the medical models. Indicating that the photographic face images represent a more challenging data set compared to the brain and dental images. This is probably due to the confounding effects of lighting variation and the large variability in appearance of the human face.

5.6. Facial feature localisation experiments

The fully automatic localisation accuracy of the CLM and AAM algorithms was tested by applying the methods to the

publicly available BIOID [25] and XM2VTS [26] data sets. Note that these images are completely independent of the training images which contain different people imaged under different conditions (see Fig. 3(c)).

Our procedure for finding initial facial feature locations in a static image is to apply our implementation of the Viola and Jones face detector [22], then apply similar smaller region detectors within the face candidate region, which are constrained using the PSM approach due to Felzenswalb [1]. This method produces a set of points from which to initialise the CLM or AAM algorithms. Four different procedures were evaluated as follows:

- AVG—Average points within the global Viola and Jones face detector (dot-dash line).
- PSM—Pictorial Structure Matching points found within the Viola and Jones candidate face region (dashed line).
- AAM—Active Appearance Model algorithm initialised with PSM points² (dotted line).
- CLM—Constrained Local Model initialised with the PSM points (solid line).

Results of applying these methods to the BIOID and XM2VTS data sets are shown in Fig. 8, which shows that the least successful method is simply using the average points from the global face detector with no local search (AVG dot-dash line). However, the global face detector alone is reasonably successful finding 95% of facial feature points within 20% of the inter-ocular separation on the BIOID data set and 92% on the XM2VTS image database. Given the detected face region the feature localisation accuracy is improved on both data sets by applying smaller feature detectors and using the PSM [1] constraint method (see the dashed line).

Figs. 8(a) and (b) also show a large improvement when using the CLM approach (solid line) or AAM method (dotted line) to refine the PSM starting points (dashed line). The CLM gives more accurate localisation on both the BIOID and XM2VTS data sets (solid line vs dotted line). This result agrees with the displacement experiments presented in Fig. 6(e) and (f) with the CLM out performing the AAM on the BIOID data set.

Fig. 9 shows an example of the CLM search converging to a successful search solution on one example from the BIOID data set. The templates steadily change to resemble the image being searched.

5.7. Facial feature tracking experiments

The CLM algorithm automatically adjusts the feature templates to match the current image. Therefore it is a natural tracking method in the sense that the templates learn to match the image, but are also constrained by joint shape and texture model to remain plausible feature templates.

² Note that the AAM formulation we use is actually the edge/corner AAM due to Scott et al. [17] which has been shown to be more effective than the basic texture method.

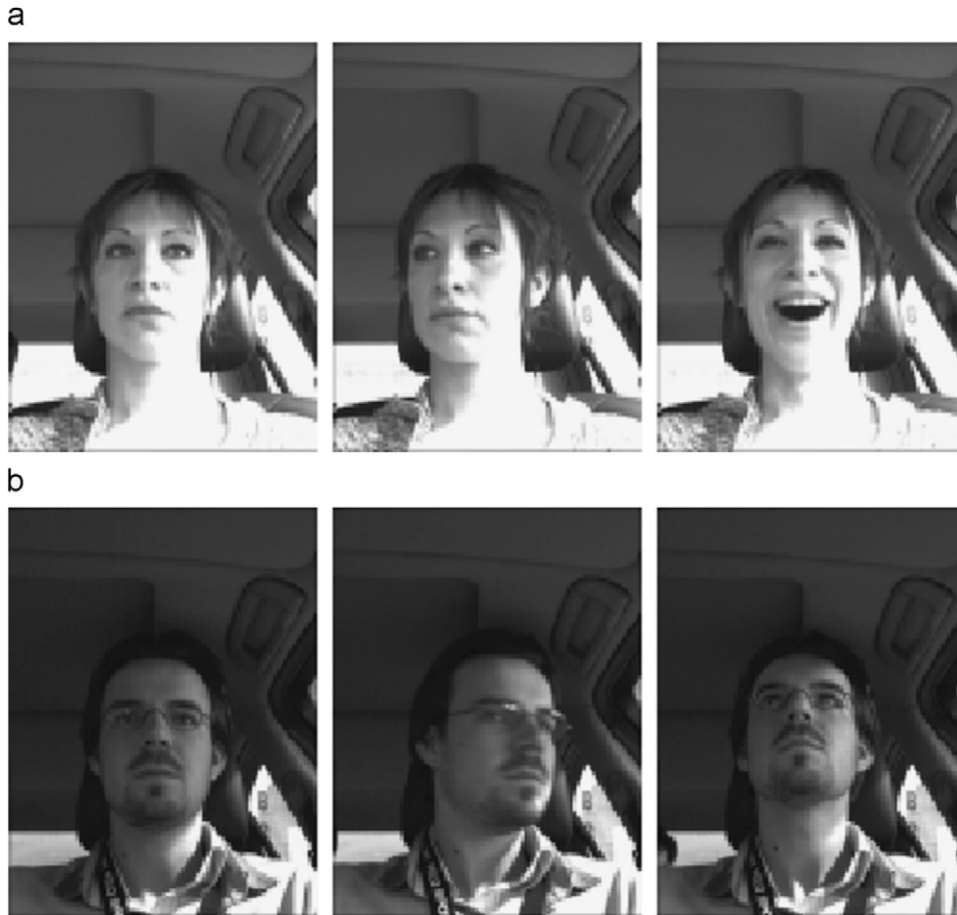


Fig. 10. Examples from car driver sequences. (a) Sequence 1 and (b) Sequence 2.

We tested the CLM method by applying it to two sequences of people driving in cars. The test sequence involves a large amount of lighting variation and head movement and is thus a challenging data set. Each sequence contains approximately 1000 frames (taken at 10 fps). See Fig. 10 for example frames from the two test sequences.

The face rotates out of plane at some point in both driver sequences. Therefore we used a quality of fit measure to test when the face has been lost and reinitialise by searching subsequent frames with the global face detector. The quality of fit measure used for the CLM method is the shape constrained response score (see Eq. (9)). The AAM fit quality is the sum of residuals of the texture model fitted to the image.

To provide ground truth for our experiments every 10th frame (i.e. once a second) was labelled by a human operator, providing all the facial features are visible. The distance measure for each labelled frame was the point to point error (see Eq. (10)) unless the labelled face is undetected in the image, when the distance is recorded as infinite. The results of applying this detect/track scheme to the driver sequences are shown in Fig. 11.

The graphs in Fig. 11 show that the CLM (solid line) generally gives better tracking performance than the AAM search algorithm (dotted line) on both video sequences. This is

probably due to the shape constrained search being more robust to local minima compared to the AAM. Tracking results are also dependent on the ability of the models to detect tracking failure and re-initialise using the global search when required.

An example of the CLM templates used to track the face in Sequence 1 are shown in Fig. 12. The short series shows a period in the video where the driver is blinking. Note that the CLM successfully tracks the sequence during this period (see the white crosses on the left column of Fig. 12) and the template patches for the eyes are able to model the blinking texture of the underlying image, whilst the remaining templates remain unchanged (see right column of Fig. 12).

5.8. Timings

The computational speed of the face models used in this paper are summarised by Table 1. The convergence time of the CLM and AAM local search method varies with different face images; however, in our implementation they are approximately equivalent. The CLM requires ~ 40 ms per iteration. When detecting two or three iterations are usually required. However when tracking, for most frames only one iteration is usually required which only takes 40ms using

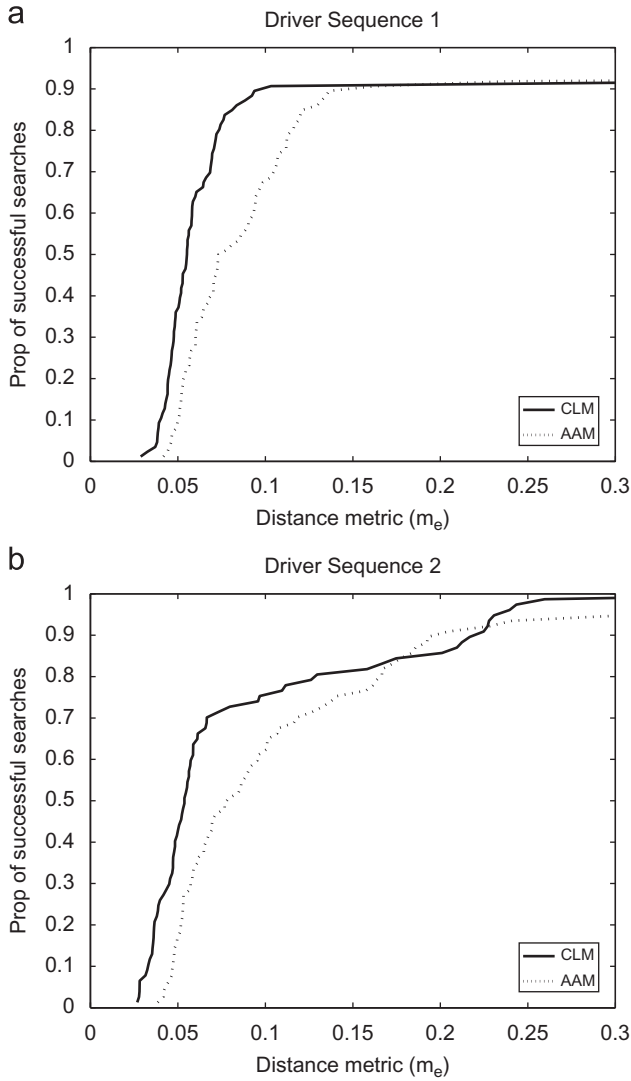


Fig. 11. Cumulative distribution of point to point error measure. (a) Sequence 1 and (b) Sequence 2.

the CLM. Similarly the AAM usually converges faster when tracking.

Therefore when detecting the full search time for a static image from the BIOID data set (384×286 pixels) is about 250 ms or 4 fps using a P4 3 GHz Processor. When tracking with the CLM the search time drops to 40 ms or 25 fps. Therefore the CLM achieves close to real-time performance on human faces.

On other data, such as the jaw and brain images (see Fig. 3) the convergence time depends on the size of the region and number of regions patches used by the CLM and the number of pixels incorporated into the AAM. The models described in this paper have broadly similar computational complexity.

6. Summary and conclusions

We have presented a novel algorithm to model a deformable object, which we refer to as the Constrained Local

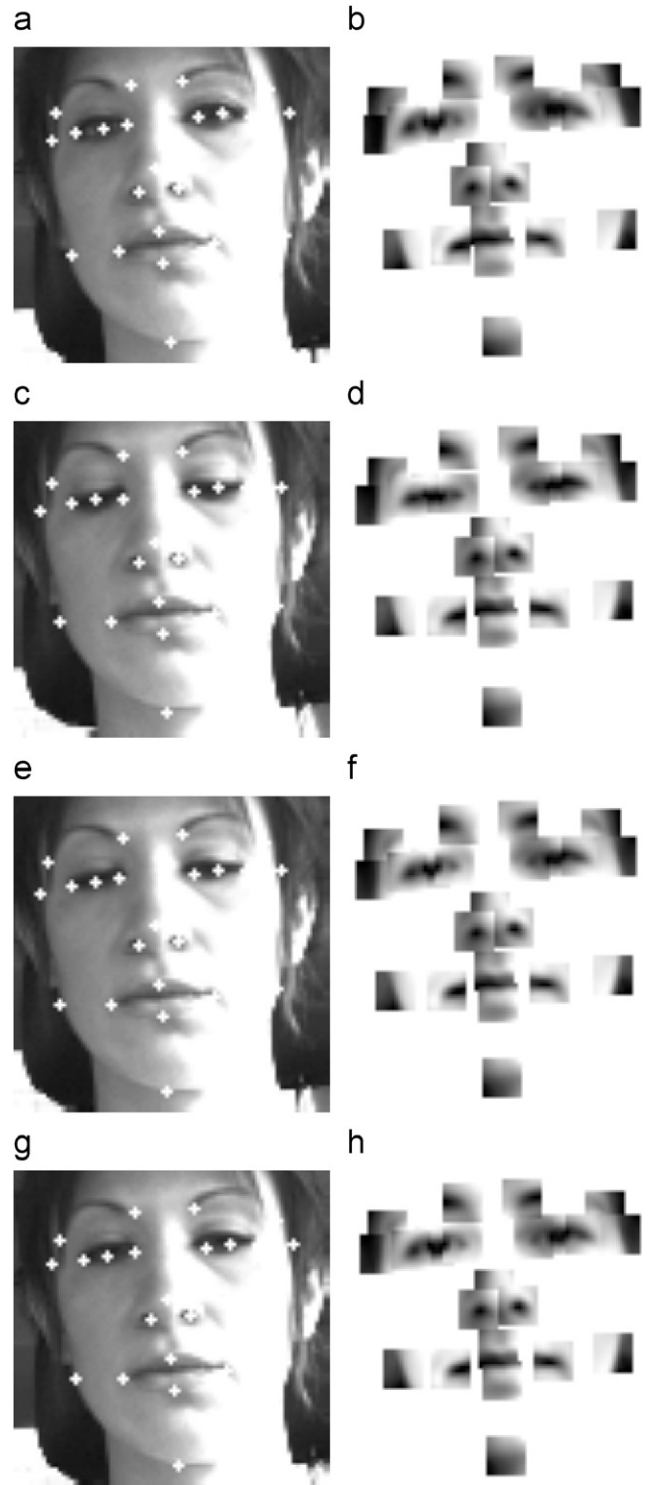


Fig. 12. CLM templates during blinking: (a) Frame 1, (b) Template 1, (c) Frame 2, (d) Template 2, (e) Frame 3, (f) Template 3, (g) Frame 4 and (h) Template 4.

Model (CLM) search. The method of building the CLM model is similar to the AAM [3] approach, but instead of modelling the whole object region we model a set of local feature templates. The feature templates are then matched to the image using an efficient shape constrained search of the template

Table 1
Time to search a single BIOID image using a P4 3GHz processor, using CLM or AAM methods

Event	CLM (ms)	AAM (ms)
Global search	~ 80	~ 80
Local feature detection (PSM)	~ 50	~ 50
Local search	~ 40–120	~ 50–100
Total	~ 170–250	~ 180–230

response surfaces. We show that when applied to faces the CLM is more accurate and has a wider radius of convergence compared to the AAM search.

We have shown that the CLM outperforms the AAM method using displacement experiments on medical image data and also when applied to the BIOID and XM2VTS static data sets as part of a fully automatic global search system (see Section 5.6). We have also shown that the CLM is more robust than the AAM when used to track faces in a set of in-car driver sequences (see Section 5.7). The CLM is found to have similar computational efficiency to the AAM, able to track faces at approximately 25 fps (see Section 5.8).³

Future work will involve extending our approach to model gradients rather than normalised pixel values, as this has been shown to improve the AAM search [17]. We may also investigate automatic model building methods, as presently the set of features and template region sizes are picked by hand, which may well be sub-optimal. Additionally the CLM local texture models and shape constraint search method may easily be extended to 3D for use in high dimensional medical data. A further option is building multi-resolution CLM models in a similar way to the multi-resolution AAM so that the CLM search can be made more efficient and potentially more accurate.

In conclusion the CLM method is a simple, efficient and robust alternative to the AAM algorithm, which models the appearance of a set of feature templates, instead of the image pixel values. We demonstrate that the new CLM approach outperforms the AAM when applied to human faces, MR brain images and dental panoramic tomograms.

Acknowledgements

We would like to thank Ryuji Funayama and Gabriel Othmezzouri at Toyota Motor Europe for providing funding and collaborating on the work described in this paper and providing video sequences from the Toyota test vehicle.

We would also like to thank Keith Horner and Hugh Devlin at the University of Manchester Dentist school and Danny Allen of the Imaging Science and Biomedical Engineering Department at University of Manchester for allowing us to reuse the panoramic tomograph images and markup previously described in Ref. [24].

³ Note a live webcam demo of the face tracker described in this paper is now available at the following web address: http://mimban.smb.man.ac.uk/downloads/face_demo.php.

Appendix A. Supplementary data

Supplementary data associated with this article can be found in the online version, at doi:10.1016/j.jedc.2004.03.007

References

- [1] P. Felzenszwalb, D. Huttenlocher, Pictorial structures for object recognition, *Int. J. Comput. Vision* 61 (2005) 55–79.
- [2] T.F. Cootes, C.J. Taylor, Active shape models, in: *Proceedings of the 3rd British Machine Vision Conference 1992*.
- [3] T.F. Cootes, G.J. Edwards, C.J. Taylor, Active appearance models, in: *Proceedings of the 5th European Conference on Computer Vision 1998*, vol. 2, Freiburg, Germany, 1998.
- [4] J.A. Nelder, R. Mead, A simplex method for function minimization, *Comput. J.* 7 (1965) 308–313.
- [5] D. Cristinacce, T. Cootes, Detection and tracking with constrained local models, in: *Proceedings of the 17th British Machine Vision Conference 2006*, Edinburgh, Scotland, 2006.
- [6] D. Cristinacce, T. Cootes, Facial feature detection and tracking with automatic template selection, in: *Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition 2006*, Southampton, UK, 2006.
- [7] N. Dowson, R. Bowden, Simultaneous modeling and tracking (smat) of feature sets, in: *Proceedings of the 23rd Computer Vision and Pattern Recognition Conference 2005*, San Diego, USA, 2005.
- [8] S. Mitchell, B. Lelieveldt, J. Bosch, R. van der Geest, J. Reiber, M. Sonka, Segmentation of cardiac MR volume data using 3D active appearance models, in: M. Sonka, J. Fitzpatrick (Eds.), *Proceedings of SPIE, Medical Imaging 2002: Image Processing*, vol. 4684, p. 433–443.
- [9] B. van Ginneken, M. Stegmann, M. Loog, Segmentation of anatomical structures in chest radiographs using supervised methods: a comparative study on a public database, *Med. Image Anal.* 10 (2006) 19–40.
- [10] V. Blanz, T. Vetter, Face recognition based on fitting a 3d morphable model, *IEEE Trans. Pattern Anal. Mach. Intell.* 25 (9) (2003) 1063–1074.
- [11] T. Vetter, V. Blanz, Estimating coloured 3d face models from single images: an example based approach, in: H. Burkhardt, B. Neumann (Eds.), *Proceedings of the 5th European Conference on Computer Vision 1998*, vol. 2, Freiburg, Germany, Springer, Berlin, 1998.
- [12] S. Romdhani, T. Vetter, 3d probabilistic feature point model for object detection and recognition, in: *Proceedings of the 25th Computer Vision and Pattern Recognition Conference 2007*, Minneapolis, USA, 2007.
- [13] D.G. Lowe, Distinctive image features from scale-invariant keypoints, *Int. J. Comput. Vision* 60 (2) (2004) 91–110.
- [14] I. Matthews, S. Baker, Active appearance models revisited, *Int. J. Comput. Vision* 60 (2) (2004) 135–164.
- [15] S. Baker, I. Matthews, Lucas-kanade 20 years on: A unifying framework, *Int. J. Comput. Vision* 54 (2004) 221–255.
- [16] J. Xiao, S. Baker, I. Matthews, T. Kanade, Real-time combined 2d+3d active appearance models, in: *Proceedings of the 22nd Computer Vision and Pattern Recognition Conference 2004*.
- [17] I.M. Scott, T.F. Cootes, C.J. Taylor, Improving appearance model matching using local image structure, in: *Information Processing in Medical Imaging*, 18th International Conference, 2003.
- [18] Y. Zheng, X.S. Zhou, B. Georgescu, S. Zhou, D. Comaniciu, Example based non-rigid shape detection, in: *Proceedings of the 9th European Conference on Computer Vision 2006*, Graz, Austria, 2006.
- [19] Y. Freund, R. Iyer, R.E. Schapire, Y. Singer, An efficient boosting algorithm for combining preferences, *J. Mach. Learn. Res.* 4 (2003) 933–969.
- [20] P. Felzenszwalb, D.C.D. Huttenlocher, Spatial priors for part-based recognition using statistical models, in: *Proceedings of the 23rd Computer Vision and Pattern Recognition Conference 2005*, vol. 1, San Diego, USA, 2005.
- [21] D. Cristinacce, T. Cootes, A comparison of shape constrained facial feature detectors, in: *Proceedings of the 6th International Conference on Automatic Face and Gesture Recognition 2004*, Seoul, Korea, 2004.

- [22] P. Viola, M. Jones, Rapid object detection using a boosted cascade of simple features, in: *Proceedings of the 19th Computer Vision and Pattern Recognition Conference 2001*, vol. 1, Hawaii, USA, Kauai, Hawaii, 2001.
- [23] I. Dryden, K.V. Mardia, *Statistical Shape Analysis*, Wiley, London, 1998.
- [24] K. Karayianni, K. Horner, A. Mitsea, L. Berkas, M. Mastoris, R. Jacobs, C. Lindh, P.F. van der Stelt, E. Harrison, J.E. Adams, S. Pavitt, H. Devlin, Accuracy in osteoporosis diagnosis of a combination of mandibular cortical width measurements—the osteodent project, *Bone* 40 (1) (2007) 223–229.
- [25] O. Jesorsky, K.J. Kirchberg, R.W. Frischholz, Robust face detection using the hausdorff distance, in: *Proceedings of the 3rd International Conference on Audio- and Video-Based Biometric Person Authentication 2001*, Halmstad, Sweden, 2001.
- [26] K. Messer, J. Matas, J. Kittler, J. Luettin, G. Maitre, Xm2vtsdb: The extended m2vts database, in: *Proceedings of the 2nd International Conference on Audio- and Video-Based Biometric Person Authentication 1999*, Washington DC, USA, 1999.

About the Author—DAVID CRISTINACCE received a B.A. degree in Mathematics from Cambridge University, England, in 1997, and an M.Sc. in Cognitive Science in 2000 and a Ph.D. in 2004 from the University of Manchester. He is currently employed as a Post Doctoral Researcher within the department of Imaging Science and Biomedical Engineering at the University of Manchester. His research interests include feature detection, object localisation, shape modelling and computer vision applied to human faces.

About the Author—TIMOTHY F. COOTES received a B.Sc. degree in Mathematics and Physics from Exeter University, England, in 1986, and a Ph.D. in Engineering from Sheffield City Polytechnic, in 1991. He obtained a postdoctoral fellowship from SERC in 1993, and an advanced fellowship from EPSRC in 1995. He became a Reader at the University of Manchester in 2004 and a Professor in 2006. His research interests include statistical models of shape and appearance variation, and their applications to industrial and medical computer vision problems.