Deformable Object Modelling and Matching

Tim F. Cootes

The University of Manchester, UK

Abstract. Statistical models of the shape and appearance of deformable objects have become widely used in Computer Vision and Medical Image Analysis. Here we give an overview of such models and of two efficient algorithms for matching such models to new images (Active Shape Models and Active Appearance Models). We also describe recent work on automatically constructing such models from minimally labelled training images.

1 Statistical Shape Models

Many objects of interest in computer vision can be considered to be some deformed version of an "average" shape. For instance, most human faces have two eyes, a nose and a mouth in similar relative positions, and good approximations to each face can be generated by modest distortions of a standard template. Similarly many anatomical structures (such as human bones, or the heart) have broadly similar shapes across a population.

Statistical shape models seek to represent such objects. Since their introduction Point Distribution Models [1], which represent shapes as a linear combination of modes of variation about the mean, have found wide application.

These represent a shape using a set of points $\{x_i, y_i\}$, $(i = 1..n)$, which define particular positions on the object of interest. They are placed at consistent positions on every example in a training set (see Figure 1).

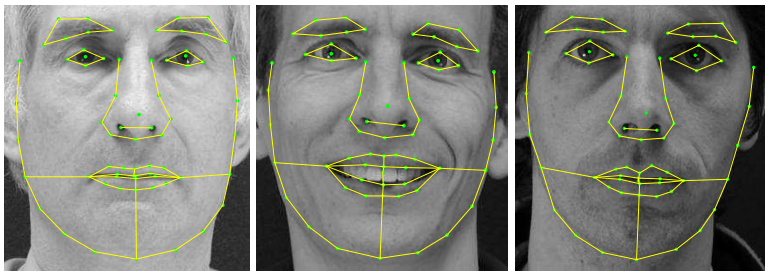


Fig. 1. Examples of faces with 68 points annotated on each, defining correspondences across the set.

By applying Procrustes Analysis [2] the examples can be aligned into a common co-ordinate frame. Principal Component Analysis is then used to build a linear model of the variation over the set as

$$\boldsymbol{x} = \hat{\boldsymbol{x}} + \boldsymbol{P}\boldsymbol{b} \tag{1}$$

where $\boldsymbol{x} = (x_1, y_1, ..., x_n, y_n)^T$, $\hat{\boldsymbol{x}}$ is the mean shape, $\boldsymbol{P}$ is a set of eigenvectors of the covariance matrix describing the modes of variation, and $\boldsymbol{b}$ is a vector of shape parameters (see [1] for details). For instance, Figure 2 shows the effect of varying the first three parameters of a face shape model [1].

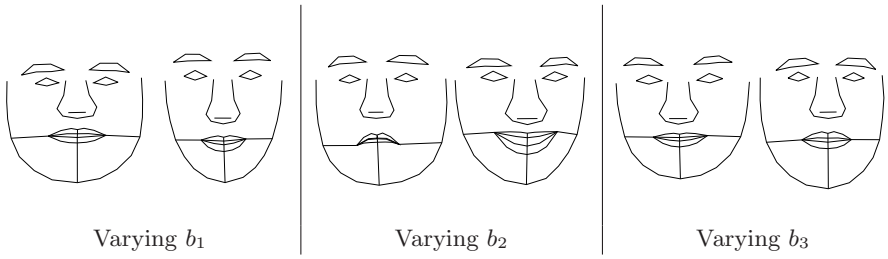


Fig. 2. First three modes of shape variation of a face model.

Such models typically have far fewer parameters (elements of $\boldsymbol{b}$) than points, as they take advantage of the correlation between points on the shape. For instance, nearby points on a boundary are usually correlated, and symmetries of the shape reduce the degrees of freedom even further.

Statistical shape models can be used to analyse the differences in shapes between populations, or can be used to help locate structures in new images.

2 Active Shape Models

An Active Shape Model (ASM) is a method of matching a statistical shape model of the form described above to a new image. The positions of the points in the image are given by the equation

$$\boldsymbol{X} = T_\theta(\hat{\boldsymbol{x}} + \boldsymbol{P}\boldsymbol{b}) \tag{2}$$

where $T_\theta(\boldsymbol{x})$ applies a transformation (typically a similarity transformation) to the set of points encoded in the vector $\boldsymbol{x}$, with parameters θ (for instance rotation, scaling and translation). T_θ thus defines the mapping from the reference frame to the target image frame, giving the global pose of the object.

For an ASM we require a method of locating a good candidate position for each model point in a region. Typically this involves building a statistical model

[1] C++ source code for building shape models will be made available in VXL (vxl.sourceforge.net) in the `contrib/mul/msm` library

of the image patch about each point from the training set, then searching the region for the best match using this model [1, 3].

The Active Shape Model matches using a simple alternating algorithm:

1. Search around each current point position $\boldsymbol{X}_i$ with the associated local model to find a better position, $\boldsymbol{Z}_i$
2. Find the shape and pose parameters $\{\boldsymbol{b}, \theta\}$ to best fit the model to the found points $\{\boldsymbol{Z}_i\}$

Though in early work the local search was along profiles normal to the model boundaries [1], this is naturally generalised to searching regions around each point [3]. The local models can either be simple Gaussian models of the image patch [1, 4], or more sophisticated classifiers [5, 6]. Alternatively regression techniques can be used to directly predict the movement of each point [7].

3 Active Appearance Models

Rather than search for each point independently, as the ASM does, the Active Appearance Model (AAM) approach is to predict an update to the model parameters directly from samples of the image. This takes into account correlations between the image patterns across the shape.

The original formulation [8, 9] was designed to efficiently fit a statistical appearance model to a new image. Such a model combines a statistical shape model with a model of image texture in a normalised reference frame. It is a generative model, in that it can create synthetic images of the objects on which it has been trained [10, 11]. A texture model is used to generate the intensity pattern in the reference frame, which is then warped to the target image frame using a deformation defined by the shape model parameters. Figure 3 shows the first three appearance modes of a face model, demonstrating how shape and texture vary together.

Fig. 3. First three modes of appearance variation of a face model.

Matching such a model to a new image is a difficult optimisation problem. If $\boldsymbol{p}$ represents all the model parameters, $\boldsymbol{t}(\boldsymbol{p})$ the texture generated by the current model, and $\boldsymbol{s}(\boldsymbol{p})$ a vector of image samples taken at the current position, then a simple approach is to seek the parameters which minimise

$$E(\boldsymbol{p}) = |\boldsymbol{s}(\boldsymbol{p}) - \boldsymbol{t}(\boldsymbol{p})|^2 \tag{3}$$

the sum of squared errors between model and image.

By differentiating this equation and making certain approximations, it is possible to show [12] that a good estimate of the optimal update to the parameters is given by a simple linear equation,

$$\delta \boldsymbol{p} = \boldsymbol{R}(\boldsymbol{s}(\boldsymbol{p}) - \boldsymbol{t}(\boldsymbol{p})) \tag{4}$$

where the update matrix $\boldsymbol{R}$ is the pseudo-inverse of a Jacobian which can be estimated from training data.

The AAM matching algorithm then simply repeats this process, updating the model parameters based on the difference between the model and the image samples. Usually a coarse-to-fine approach is used, in which low resolution models are used during the early parts of the search, and later refined with more detailed models.

Although the derivation of the update matrix $\boldsymbol{R}$ as the pseudo-inverse of the Jacobian is elegant, it turns out that often it is better to treat the task as a regression problem, in which we seek the matrix which gives the best parameter updates. If we generate many random parameter displacements $\delta \boldsymbol{p}$ on a training set, and for each compute the residual error $\boldsymbol{r} = \boldsymbol{s}(\boldsymbol{p} + \delta \boldsymbol{p}) - \boldsymbol{t}(\boldsymbol{p} + \delta \boldsymbol{p})$, we can then use regression methods to estimate the matrix $\boldsymbol{R}$ [9, 13]. This has been shown to lead to more accurate results than using the inverse of the Jacobian [13, 14].

This general approach has proved very effective, and has been developed in many directions including

- a more elegant compositional update scheme [15]
- the inclusion of constraints [16]
- the use of other image features for improved robustness [17, 18]
- methods of dealing with occlusion [19]
- combining 2D and 3D models for face tracking [20]
- more sophisticated update schemes [21, 22, 14]

amongst many others.

Though often used for face model matching, the AAM has been widely applied in medical image interpretation. For instance, modelling the heart [23], hand [24], brain [25] or knee [26].

4 Automatic Model Construction

The shape and appearance models used in the above methods rely on training sets containing points annotated on each of a representative set of images. Manually annotating such data is time consuming, and is particularly difficult for three dimensional volume images, widely used in medical imaging.

Thus there has been a long history of work attempting to automate the model building process. The points on each image define the correspondences between the objects viewed. If such correspondences can be estimated automatically, a model can be constructed with minimal human intervention.

If we have the boundary of each 2D object (or surface of each 3D object), effective correspondence can be found using techniques which optimise the ability of the resulting model to encode the shapes - Minimum Description Length methods [27, 28].

Where we only have images, with no annotation, the problem is more challenging. A common approach is to use non-rigid registration or optical flow methods to find the correspondences between each image and a reference image. Some of the earliest work in this vein was by Vetter, Jones and Poggio [29, 30] who used a combination of model fitting and optical flow to estimate dense vector fields across sets of objects, to build 'morphable models' - statistical models of shape and appearance.

Registering images to build an average 'atlas' is a widely used technique in medical image analysis. For instance, Guimond *et al.* [31] used Thirion's 'Demons' algorithm [32] to register sets of images, describing how to iteratively update the group mean. Frangi *et al.* used non-rigid B-spine registration [33] to correspond 3D images and build statistical shape models [34]. Joshi *et al.* [35, 36] demonstrate how to simultaneously estimate the reference shape and image in the case of large scale diffeomorphisms, where linear approximations to averages break down.

Our early work in this field focused on estimating diffeomorphisms (smooth invertable mappings) from a reference frame to each target image in a set. We assume that each image in a set should contain the same structures, and hence there should be a unique and invertible one-to-one correspondence between all points on each pair of images - a diffeomorphism (see [37]).

For any two diffeomorphisms $f(\boldsymbol{x})$, $g(\boldsymbol{x})$, their composition $(f \circ g)(\boldsymbol{x}) \equiv f\left(g(\boldsymbol{x})\right)$ is also a diffeomorphism. We can thus construct a wide class of diffeomorphic functions by repeated compositions of a basis set of simple diffeomorphisms. In [38] we describe a coarse-to-fine algorithm for estimating the correspondences across a group using such compositions.

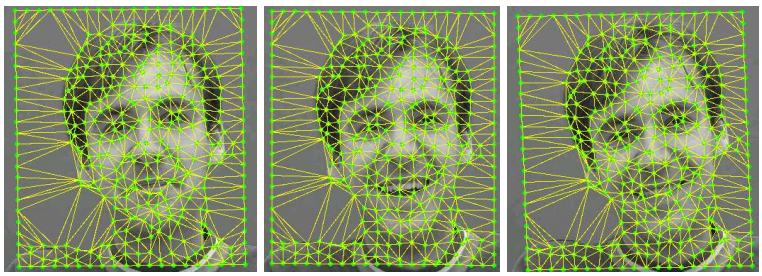


Fig. 4. Examples from a training set with resulting control points [39]

However, although the compositional approach can generate deformation fields which are guaranteed to be invertable, actually computing the inverse can be difficult, as it usually involves a non-linear optimisation. A more prag-

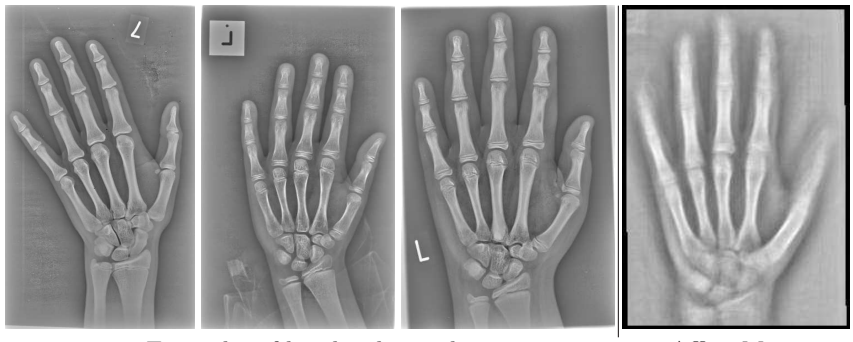
matic alternative is to represent deformation fields using a triangulated mesh (in 2D) or a tetrahedral mesh (in 3D) guided by the control points at the nodes. Affine interpolation can be used to compute the transformation inside the mesh elements, leading to a piece-wise affine representation of the warp. As long as the triangles/tetrahedra do not 'flip', the transformation is invertable - one simply swaps the source and destination control points [39]. Although not smooth, as there are discontinuities in the derivative at the element boundaries, such representations are accurate enough for many applications (Figure 4).

In [39, 12] we describe a groupwise image registration algorithm in which such meshes are used in a Minimum Description Length optimisation framework. A related method is described by Baker *et al.* [40].

Figure 5 shows the first three appearance modes of a model constructed from 300 face images of different people, with the correspondences computed automatically using the groupwise algorithm from [41].

Mode 1 Mode 2 Mode 3

Fig. 5. First three modes of appearance variation of a face model constructed from automatically computed correspondences.



Examples of hand radiographs Affine Mean

Fig. 6. Examples of hand radiographs, which display considerable shape variation, and the resulting affine mean (from [42]).

4.1 Initialising Groupwise Registration

The methods described above work well, given a good enough initialisation. As they involve local optimisation, if poorly initialised, they fall into local minima. Typically an affine transformation is estimated as the initialisation. However, for objects with significant shape variation this may not be sufficient. For example when registering hand radigraphs (Figure 6), the resulting affine mean is poor, and further registration does not significantly improve the result as it has fallen into a local minima.

To overcome this, more accurate initialisation is required. In [43] we show that by manually annotating a small number of key points on a single image we can construct a parts+geometry model from the whole training set, which can accurately locate those points on all the images. Such points give a sparse corresondence, which can be used to initialise a denser groupwise registration.

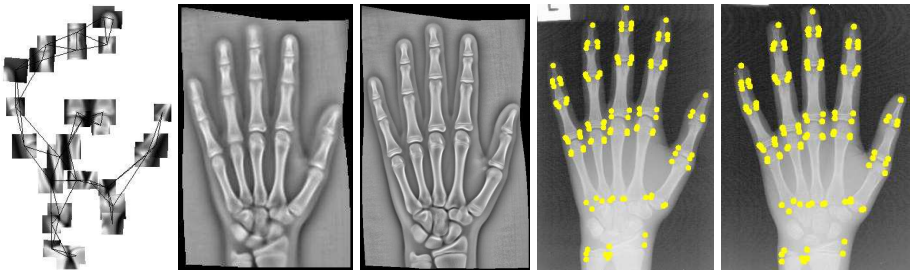


Fig. 7. Automatically generated parts+geometry models and results of dense groupwise registration [42]

In [42] this approach is extended to automatically select a sparse set of points which can be accurately located across the whole set. We use a variant of the Genetic Algorithm to select good subsets from a large pool of candidate part models. For instance, Figure 7 shows the application of this approach to hand radiographs.

5 Conclusions

Statistical models of shape and appearance are powerful tools for interpretting images. They can be matched to new images using the Active Shape Model or Active Appearance Model algorithms or their variants. Such models are built from annotated training images. Though significant progress has been made in developing algorithms to automatically estimate suitable correspondences, more work is required to make such algorithms sufficiently robust on challenging data.

Acknowledgements. Thanks to all to collaborators within ISBE and beyond who have contributed to this work.

References

1. Cootes, T.F., Taylor, C.J., Cooper, D., Graham, J.: Active shape models - their training and application. Computer Vision and Image Understanding **61** (1995) 38–59
2. Goodall, C.: Procrustes methods in the statistical analysis of shape. Journal of the Royal Statistical Society B **53** (1991) 285–339
3. Milborrow, S., Nicolls, F.: Locating facial features with an extended active shape model. In: ECCV. (2008) http://www.milbo.users.sonic.net/stasm.
4. Cootes, T.F., Page, G., Jackson, C., Taylor, C.J.: Statistical grey-level models for object location and identification. Image and Vision Computing **14** (1996) 533–540
5. van Ginneken, B., A.F.Frangi, J.J.Stall, ter Haar Romeny, B.: Active shape model segmentation with optimal features. IEEE Trans. Medical Imaging **21** (2002) 924–933
6. M.Wimmer, F.Stulp, S.J.Tschechne, B.Radig: Learning robust objective functions for model fitting in image understanding applications. In: Proc. British Machine Vision Conference. Volume 3. (2006) 1159–168
7. Cristinacce, D., Cootes, T.: Boosted active shape models. In: Proc. British Machine Vision Conference. Volume 2. (2007) 880–889
8. Edwards, G., Taylor, C.J., Cootes, T.F.: Interpreting face images using active appearance models. In: 3^{rd} International Conference on Automatic Face and Gesture Recognition 1998, Japan (1998) 300–305
9. Cootes, T.F., Edwards, G.J., Taylor, C.J.: Active appearance models. In H.Burkhardt, Neumann, B., eds.: 5^{th} European Conference on Computer Vision. Volume 2., Springer, Berlin (1998) 484–498
10. Lanitis, A., Taylor, C.J., Cootes, T.F.: Automatic interpretation and coding of face images using flexible models. IEEE Trans. on Pattern Analysis and Machine Intelligence **19** (1997) 743–756
11. Edwards, G.J., Taylor, C.J., Cootes, T.F.: Learning to identify and track faces in image sequences. In: 8^{th} British Machine Vision Conference, Colchester, UK (1997) 130–139
12. Cootes, T.F., Edwards, G.J., Taylor, C.J.: Active appearance models. IEEE Trans. Pattern Analysis and Machine Intelligence **23** (2001) 681–685
13. Donner, R., Reiter, M., Langs, G., Peloschek, P., Bischof, H.: Fast active appearance model search using canonical correlation analysis. IEEE Trans. Pattern Analysis and Machine Intelligence **28** (2006) 1690–4
14. Tresadern, P., Sauer, P., Cootes, T.: Additive update predictors in active appearance models. In: British Machine Vision Conference, BMVA Press (2010)
15. Matthews, I., Baker, S.: Active appearance models revisited. International Journal of Computer Vision **60** (2004) 135 – 164
16. Cootes, T.F., Taylor, C.J.: Constrained active appearance models. In: 8^{th} International Conference on Computer Vision. Volume 1., IEEE Computer Society Press (2001) 748–754
17. T.F.Cootes, C.J.Taylor: On representing edge structure for model matching. In: Computer Vision and Pattern Recognition. Volume 1. (2001) 1114–1119
18. I.M.Scott, T.F.Cootes, C.J.Taylor: Improving appearance model matching using local image structure. In: Information Processing in Medical Imaging, Springer-Verlag (2003) 258–269
19. Gross, R., Matthews, I., Baker, S.: Constructing and fitting active appearance models with occlusion. In: Proceedings of the IEEE Workshop on Face Processing in Video. (2004)

20. Xiao, J., S.Baker, Matthews, I., Kanade, T.: Real-time combined 2D+3D active appearance models. In: Computer Vision and Pattern Recognition. Volume 2., IEEE (2004) 535–542
21. Saragih, J., Goecke, R.: Iterative error bound minimisation for AAM alignment. In: Proc.ICPR. Volume 2. (2006) 1192–1195
22. Saragih, J., Goecke, R.: A non-linear discriminative approach to AAM fitting. In: Proc.ICCV. (2007)
23. Mitchell, S., B.Lelieveldt, van der Geest, R., Schaap, J., Reiber, J., Sonka, M.: Segmentation of cardiac MR images: An active appearance model approach. In: SPIE Medical Imaging. (2000)
24. H.H.Thodberg: Hands-on experience with active appearance models. In: SPIE Medical Imaging. (2002)
25. Babalola, K., T.Cootes, C.Twining, V.Petrovic, C.Taylor: 3D brain segmentation using active appearance models and local regressors. In: Proc. MICCAI. Volume 1. (2008) 401–408
26. Vincent, G., Wolstenholme, C., Scott, I., Bowes, M.: Fully automatic segmentation of the knee joint using active appearance models. In: Medical Image Analysis for the Clinic: A Grand Challenge. (2010)
27. Davies, R., C.Twining, Cootes, T., Taylor, C.: A minimum description length approach to statistical shape modelling. IEEE Trans. on Medical Imaging **21** (2002) 525–537
28. Davies, R., C.Twining, Cootes, T., Waterton, J., Taylor, C.: 3D statistical shape models using direct optimisation of description length. In: European Conference on Computer Vision. Volume 3., Springer (2002) 3–20
29. Vetter, T., Jones, M., Poggio, T.: A bootstrapping algorithm for learning linear models of object classes. In: Computer Vision and Pattern Recognition Conference 1997. (1997) 40–46
30. Jones, M.J., Poggio, T.: Multidimensional morphable models. In: 6^{th} International Conference on Computer Vision. (1998) 683–688
31. Guimond, A., Meunier, J., Thirion, J.P.: Automatic computation of average brain models. In: Proc. MICCAI. (1998) 631–640
32. Thirion, J.P.: Image matching as a diffusion process: an analogy with Maxwell's demons. Medical Image Analysis **2** (1998) 243–260
33. Rueckert, D., Sonoda, L.I., Hayes, C., Hill, D.L.G., Leach, M.O., Hawkes, D.J.: Non-rigid registration using free-form deformations: Application to breast MR images. IEEE Trans. Medical Imaging **18** (1999) 712–721
34. Frangi, A., Rueckert, D., Schnabel, J., Niessen, W.: Automatic 3D ASM construction via atlas-based landmarking and volumetric elastic registration. In: 17^{th} Conference on Information Processing in Medical Imaging. (2001) 78–91
35. Joshi, S., B.Davis, M.Jomier, G.Gerig: Unbiased diffeomorphic atlas construction for computational anatomy. NeuroImage **23** (2004) S151–S160
36. Lorenzen, P., Davis, B., Joshi, S.: Unbiased atlas formation via large deformations metric mapping. In: Proc. MICCAI. (2005) 411–418
37. Twining, C., Marsland, S., Taylor, C.: Measuring geodesic distances on the space of bounded diffeomorphisms. In P.L.Rosin, Marshall, D., eds.: 13^{th} British Machine Vision Conference. Volume 2., BMVA Press (2002) 847–856
38. Cootes, T., Marsland, S., Twining, C., Smith, K., Taylor, C.: Groupwise diffeomorphic non-rigid registration for automatic model building. In: 8^{th} European Conference on Computer Vision. Volume 4., Springer (2004) 316–327

39. Cootes, T., Twining, C., V.Petrović, R.Schestowitz, Taylor, C.: Groupwise construction of appearance models using piece-wise affine deformations. In: 16th British Machine Vision Conference. Volume 2. (2005) 879–888
40. Baker, S., Matthews, I., J.Schneider: Automatic construction of active appearance models as an image coding problem. IEEE Trans. on Pattern Analysis and Machine Intelligence **26** (2004) 1380–84
41. Cootes, T.F., C.J.Twining, V.S.Petrović, K.O.Babalola, Taylor, C.J.: Computing accurate correspondences across groups of images. IEEE Trans. Pattern Analysis and Machine Intelligence **32** (2010) (To Appear)
42. Zhang, P., T.F.Cootes: Learning sparse correspondences for initialising groupwise registration. In: MICCAI. Volume 2. (2010) 635–642
43. Adeshina, S., T.F.Cootes: Constructing part-based models for groupwise registration. In: Proc. ISBI. (2010)