Combining Local and Global Shape Models for Deformable Object Matching

Philip A Tresadern
philip.tresadern@manchester.ac.uk
Harish Bhaskar
harish.bhaskar@manchester.ac.uk
Steve A Adeshina
steve.adeshina@postgrad.manchester.ac.uk
Chris J Taylor
chris.taylor@manchester.ac.uk
Tim F Cootes
tim.cootes@manchester.ac.uk

Imaging Sciences and Biomedical Engineering,
School of Cancer and Imaging Sciences,
University of Manchester,
Manchester M13 9PT,
United Kingdom

A common problem in vision is that of localizing and tracking non-rigid, deformable objects. One approach represents the object as a constellation of patches located at N 'landmark' points whose relative positions and orientations are constrained by spring-like forces. A dense network of constraints between parts (a *global* model) results in algorithms such as the Active Shape Model [1] and Constrained Local Model [3] that are prone to local minima. In contrast, a sparse network of constraints (*i.e.* a Markov Random Field or *local* model) permits a number of candidates to be considered for each part. As a result, this is potentially more robust as demonstrated by the Pictorial Structures Model (PSM [4]) but does not capture all correlations between feature locations. Intermediate levels of sparsity may therefore be employed to maintain efficiency without sacrificing too much accuracy [2].

In this paper we combine a global shape model with a local model of displacements (see also [5]). The local models efficiently select good candidate points, giving robustness to false matches, while the global model regularizes the result to ensure a plausible final shape. We denote this method a *cascade of Combined Shape Models* (c-CSM) and demonstrate it on images of faces and hands. One novelty of the method is that it uses a two-stage *cascade* implementation for robustness, using each trained level of the cascade to update the predicted feature locations before training the following level.

Representing the model as a set of N points, $\mathbf{X} = \{\mathbf{x}_i = (u_i, v_i)\}$, our aim is then to find the optimal set, $\mathbf{X}^*$, that maximizes the posterior,

$$p(\mathbf{X}|\mathbf{I}) \propto p(\mathbf{I}|\mathbf{X})p(\mathbf{X}), \tag{1}$$

when given a query image, $\mathbf{I}$. Starting with an initial estimate of feature locations, $\mathbf{X}_0^*$, the first step in this process is to select the set of most promising candidates:

$$\mathbf{Y}_t = \arg\max_{\mathbf{Y}} p(\mathbf{I}|\mathbf{Y})p(\mathbf{Y}|\mathbf{X}_{t-1}^*) \tag{2}$$

where: $p(\mathbf{I}|\mathbf{Y})$ is the *likelihood* of a given set of candidates and is approximated via normalized cross correlation of the image gradients in a region around the current position for each feature with a learned template; $p(\mathbf{Y}|\mathbf{X}_{t-1}^*)$ is the *prior* probability of a set of feature locations based on sparse constraints between their relative displacements from current estimates. Having chosen a set of candidates, we then regularize them in order to update our estimate of feature locations:

$$\mathbf{X}_t^* = \arg\max_{\mathbf{X}} p(\mathbf{X}|\mathbf{Y}_t) \tag{3}$$

where $p(\mathbf{X}|\mathbf{Y}_t)$ is the probability of a more densely constrained set of feature locations, given selected candidate locations. We maximize this quantity by projecting the candidate locations onto a learned subspace:

$$\mathbf{X}_t^* = S_t(\overline{\mathbf{X}} + \mathbf{P}\mathbf{P}^{\mathrm{T}}(S_{t-1}^{-1}\mathbf{Y}_t - \overline{\mathbf{X}})) \tag{4}$$

where S_t is the estimated pose (position, orientation and scale) of the object at time t (estimated from $\mathbf{X}_t^*$), $\overline{\mathbf{X}}$ is the mean shape over the training data (in a normalized reference co-ordinate frame) and $\mathbf{P}$ is a set of eigenvectors that span the shape subspace. These two steps (candidate selection and regularization) are then iterated until $p(\mathbf{X}|\mathbf{I})$ is maximized.

We apply this search algorithm in a hierarchical fashion, first localizing a small subset of highly salient points which are then used as an initialization for a more complex model with more points and greater flexibility in the global shape model. Importantly, search parameters and the properties of the MRF are re-learned at each level such that the correct distribution parameters are employed rather than (incorrectly) assuming constant values for all levels. This maximizes efficiency whilst maintaining performance.

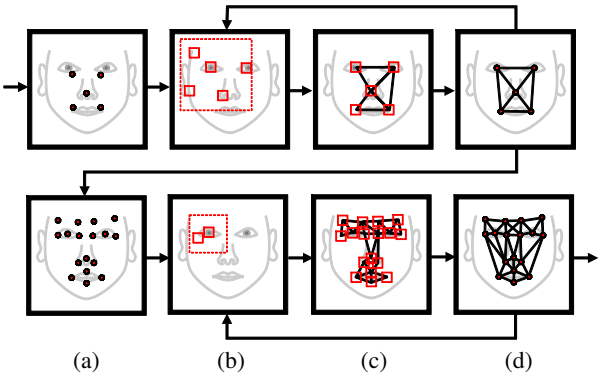

Figure 1: Schematic of a two-level cascade search: (a) initialize; (b) find candidate points; (c) select best set of candidates using MRF; (d) regularize using global model and either iterate, go to next level or finish.

For face datasets, we train our model with 1052 images collected from a number of sources under varying conditions. The algorithm was used to locate 17 landmark points on faces in the XM2VTS and BioID datasets. Performance on the XM2VTS dataset was considerably improved using our algorithm than the PSM or CLM methods. For the BioID dataset there is some improvement in accuracy though a small increase in the number of failures (*e.g.* due to missed detections). With regard to computation, the PSM and CLM approaches are more efficient ($\sim$200ms per BioID image compared to $\sim$300ms for the c-CSM) though one should note that the c-CSM also captures rotation in the plane (requiring repeated sub-sampling of the image) which our implementations of the PSM/CLM do not. We also tested the method for a dataset of hand radiographs, giving a median error of approximately 1.5% when normalized to the length of the fifth metacarpal.

In conclusion, PCA-based global shape models can be combined with MRF-based local models, using a hierarchical matching approach where lower levels of the cascade give a coarse alignment that is later refined by higher levels with a more flexible shape model. Estimating system parameters from training data rather than specifying them empirically results in a method that outperforms similar recent methods on standard datasets.

[1] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham. Active shape models – Their training and application. *Comput. Vis. Image Und.*, 61(1):266–275, January 1995.

[2] D. Crandall, P. Felzenszwalb, and D. Huttenlocher. Spatial priors for part-based recognition using statistical models. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, volume 1, 2005.

[3] D. Cristinacce and T. F. Cootes. Automatic feature localisation with constrained local models. *Pattern Recogn.*, 41:3054–3067, 2008.

[4] P. Felzenszwalb and D. Huttenlocher. Pictorial structures for object recognition. *Int. J. Comput. Vis.*, 61(1):55–79, January 2005.

[5] L. Liang, F. Wen, X. Tang, and Y.-Q. Xu. An integrated model for accurate shape alignment. In *Proc. European Conf. on Computer Vision*, 2006.