Statistical Foreground Modelling for Object Localisation

Josephine Sullivan[1], Andrew Blake[2], and Jens Rittscher[1]

[1] Dept. of Engineering, Oxford University, Oxford OX2 3PJ, UK.
{sullivan, jens}@robots.ox.ac.uk,
WWW home page: http://www.robots.ox.ac.uk/~vdg
[2] Microsoft Research Ld., 1 Guildhall Street, Cambridge CB2 3NH, UK.
ablake@microsoft.com

Abstract. A Bayesian approach to object localisation is feasible given suitable likelihood models for image observations. Such a likelihood involves statistical modelling — and learning — both of the object foreground and of the scene background. Statistical background models are already quite well understood. Here we propose a "conditioned likelihood" model for the foreground, conditioned on variations both in object appearance and illumination. Its effectiveness in localising a variety of objects is demonstrated.

1 Introduction

Following "pattern theory" [15,21], we regard an image of an object as a function $I(\mathbf{x}),\ \mathbf{x} \in \mathcal{D} \subset \mathcal{R}^2$, generated from a template image $\bar{I}(\mathbf{x})$ over a support $\bar{S}$ that has undergone certain distortions. Much of the distortion is accounted for as a warp of the template $\bar{I}(\mathbf{x})$ into the image by a warp mapping T_X:

$$\bar{I}(\mathbf{x}) = I(T_X(\mathbf{x})),\ \mathbf{x} \in \bar{S}, \tag{1}$$

where T_X is parameterised by $X \in \mathcal{X}$ over some configuration space $\mathcal{X}$, for instance planar affine warps. We adopt the convention that $X = 0$ is the template configuration so that T_X is the identity map when $X = 0$.

Using the warp framework, "analysis by synthesis" can be applied to generate the posterior distribution for X. Given a prior distribution $p_0(X)$ for the configuration X, and an observation likelihood $L(X) = p(Z|X)$ where $Z \equiv Z(I)$ is some finite-dimensional representation of the image I, then the posterior density for X is given by

$$p(X|Z) \propto p_0(X)p(Z|X). \tag{2}$$

This can be done very effectively by *factored sampling* [16] which produces a weighted "particle-set" $\{(s^{(1)}, \pi_1), \ldots, (s^{(N)}, \pi_N)\}$, of size N that approximates the posterior [7]. From this approximation of the distribution fusion of inference about X from different sensors, over time and across scales. It also allows a structured way of incorporating prior knowledge to the algorithm.

Much of the challenge with the pattern theory approach is in constructing a suitable matching score. Examples of non-Bayesian approaches include correlation scores [9,

D. Vernon (Ed.): ECCV 2000, LNCS 1843, pp. 307–323, 2000.

4,10,17] and mutual information [27]. But factored sampling, calls for a Bayesian approach in which both the foreground and background image statistics are modelled [14]. In particular, modelling a likelihood $p(Z|X)$ in terms of the foreground/background statistics of receptive field outputs is employed in *Bayesian Correlation* [26]. Although background statistics for Bayesian Correlation, and their independence properties, are quite well understood [11,22,1,28,6,25] foreground statistics are more complex.

Foreground statistics should be characterised by the response of a receptive field *conditioned* on its location relative to the object and on the object's pose. This can be achieved by performing template subtraction. This increases the specifity and selectivity between background and foreground over the method of adhoc foreground "partitioning" implemented in [26]. The weakness of the latter approach is demonstrated in figure 1. Even when receptive fields are mutually independent over the background, independence need not necessarily hold over the foreground. It was hoped that the new foreground measurements would also be decorrelated and/or independent. However, it turns out that the statistical dependencies between measurements are not greatly affected by the template subtraction. This paper proposes a more acutely tuned foreground likelihood,

Fig. 1. Simple foreground partitioning gives poor selectivity. An *decoy* object produces an alternative likelihood peak of sufficient strength that the mean configuration (black contour) is substantially displaced from the true location of the head. (white contours represent the posterior distribution; wider contours indicate higher likelihood for the face object.)

conditioned explicitly on variability of pose and illumination, that pays greater respect to the deterministic properties of the object's geometric layout.

2 Modelling Image Observations

In the framework presented here, image intensities are observed via a bank of filters, isotropic ones in the examples shown here, though steerable, oriented filters [23] would also be eminently suitable. The likelihood of such observations depends both on foreground and background statistics [26] and this approach is reviewed below, before looking more carefully at foreground models in the following section.

2.1 Filter Bank

The observation $Z = Z(I)$ is to be a fixed, finite dimensional representation of the image I, consisting of a vector $Z = (z_1, \ldots, z_K)$ whose components

$$z_k = \int_{S_k} W_{x_k}(\mathbf{x}) I(\mathbf{x}) \mathrm{d}\mathbf{x}, \tag{3}$$

are an inner product of the image with a filter function W_{x_k}, over a finite support S_k. In [26] it was argued that a suitable choice of filter function is a Laplacian of Gaussian $W_{\mathbf{x}}$, centred at $\mathbf{x}$:

$$W_{\mathbf{x}}(\mathbf{x}') = \nabla^2 G_\sigma(\mathbf{x}' - \mathbf{x})$$

with hexagonally tesselated, overlapping supports as in figure 2. The scale parameter of

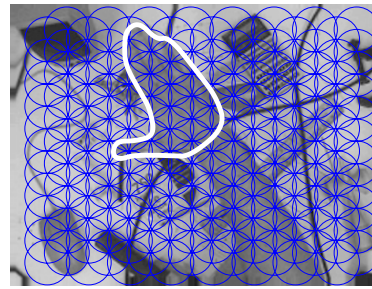


Fig. 2. Tessellation of filter supports. Filters are arranged in a hexagonal tessellation, as shown, with substantial overlap (support radius $r = 40$ pixels illustrated).

the Gaussian is σ and it is adequate to truncate the Gaussian to a finite support of radius $r = 3\sigma$. The tessellation scheme was arrived at [26] by requiring the densest packing of supports while maintaining statistical de-correlation between filters over background scene texture. In practice, at that separation, filter responses are not only decorrelated but also, to a good approximation, independent over the background.

2.2 Probabilistic Modelling of Observations

The observation (ie output value) z from an individual filter is generated by integration over a support-set S such as the circular one in figure 3, which is generally composed of both a background component $B(X)$, and a foreground component $F(X)$:

$$z|X = \underbrace{\int_{B(X)} W(\mathbf{x}) I(\mathbf{x})\, \mathrm{d}\mathbf{x}}_{\text{MAIN NOISE SOURCE}} + \int_{F(X)} W(\mathbf{x}) I(\mathbf{x})\, \mathrm{d}\mathbf{x}. \tag{4}$$

Densities $p^{\mathcal{B}}(z|\rho)$ and $p^{\mathcal{F}}(z|\rho),\ 0 < \rho < 1$ for the background and foreground components of mixed supports must be learned. Then, a particular object hypothesis X is

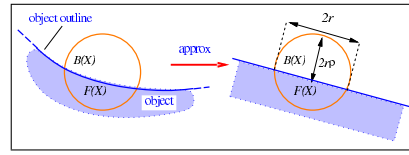


Fig. 3. Foreground and background filter components A circular support set S is illustrated here, split into subsets $F(X)$ from the foreground and $B(X)$ from the background. Assuming that the object's bounding contour is sufficiently smooth, the boundary between foreground and background can be approximated as a straight line. The support therefore divides into segments with offsets $2r\rho$ and $2r(1-\rho)$ for background and foreground respectively.

evaluated as a global likelihood score $p(Z|X)$, based on components $z_1, \ldots, z_K$ which need to have either a known mutual dependence or, simpler still, be statistically independent. Then the observation likelihood can be constructed as a product

$$p(Z|X) = \prod_{k=1}^{K} p(z_k|X). \tag{5}$$

containing terms $p(z_k|X)$ in which the density $p(z_k|X)$ depends, to varying degrees according to the value of X, on each of the learned densities $p^{\mathcal{F}}$ and $p^{\mathcal{B}}$ for the foreground and the background model. This places the requirement on the filter functions W_{X_k}, that they should generate such mutually independent z_k. As mentioned in section 2.1, this is known to be true for z_k over the background. Here we aim to establish independence also over the foreground.

3 Modelling the Foreground Likelihood

The modelling of background components is straightforward [26], simply inferring a density for responses z from a training set of filter outputs z_n, calculated from supports S_n dropped at random over an image [26]. Then $p^{\mathcal{B}}(z|\rho)$ can be learned for some finite set of ρ-values, and interpolated for the ρ-continuum. A similar approach can be used for the foreground case $p^{\mathcal{F}}$ but with some important additional complexities however.

3.1 Spatial Pooling

The distribution $p^{\mathcal{B}}(z|\rho)$ is learned from segments dropped down at random, anywhere on the background. Over the foreground, and in the case that $\rho = 0$, $p^{\mathcal{F}}(z|\rho)$ is similarly learned from a circular support, dropped now at any location wholly inside the training object. However, whenever $\rho > 0$, the support $F(X)$ must touch the object outline; therefore $p^{\mathcal{F}}(z|\rho)$ has to be learned entirely from segments touching the outline. Thus, for $\rho = 0$, statistics are pooled over the whole of the object interior — "spatial pooling", whereas for $\rho > 0$ statistics pooling is restricted to occur over narrow bands, of width $2r(1-\rho)$, running around the inside of the template contour.

Spatial pooling dilutes information contained in the gross spatial arrangement of the grey-level pattern. Sometimes this provides adequate selectivity for the observation likelihood, particularly when the object outline is distinctive, such as the outline of a hand as in figure 2. The outline of a face, though, is less distinctive. In the extreme case of a circular face, and using isotropic filters, rotating the face would not produce any change in the pooled response statistics. In that case, the observation likelihood would carry no information about (2D) orientation. One approach to this problem is to include some anisotropic filters in the filter bank, which would certainly address the rotational indeterminacy. Another approach [26] to enhancing selectivity is to subdivide the interior $\mathcal{F}$ of the object as $\mathcal{F} = \mathcal{F}_0 \cup \ldots \cup \mathcal{F}_{N_F}$, and construct individual distributions $p^{\mathcal{F}_i}(z|\rho = 0)$ for each subregion $\mathcal{F}_i$. However, the choice of the number and shape of subregions is somewhat arbitrary. It would be much more satisfying to find a way of increasing selectivity that is tailored specifically to foreground structure, rather than imposing an arbitrary subdivision, and that is what we seek to do in this paper.

3.2 Warp Pooling

In principle the foreground density $p^{\mathcal{F}}$ depends on the full warp T_X. This means that $p^{\mathcal{F}}(z|\rho)$ must be learned not simply from one image, but from a training set of images containing a succession of typical transformations of the object, and this is reasonable enough. In principle, the learned $p^{\mathcal{F}}$ should be parameterised not merely by $\rho(X)$, as was the case for the background, but by the full, multi-dimensional configuration X itself, and that is not computationally feasible. One approach to this problem is that if these variations cannot be modelled parametrically, they can nonetheless be pooled into the general variability represented by $p^{\mathcal{F}}(z|\rho)$. However, such "warp pooling" dilutes the available information about X, especially given that it is combined with spatial pooling as above.

3.3 Foreground Distribution

The predictable behaviour of filter responses over natural scenes, which applies well to background modelling, could not necessarily be expected to apply for foreground models. Filter response z over background texture assumes a characteristic kurtotic form, well modelled as by an exponential (Laplace) distribution [6]. The foreground,

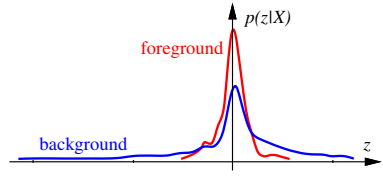


Fig. 4. Foreground and background distributions for support radius $r = 20$ pixels. The background distribution has higher kurtosis, having extended tails.

being associated with just a single object, is less variable and does not have extended tails (figure 4). Hence the exponential distribution that applies well to the background [6] is inapplicable and a normal distribution is more appropriate.

As for independence, filter outputs over the background are known to be uncorrelated at a displacement of r or 3σ but this need not necessarily hold over the foreground. Nonetheless, autocorrelation experiments done over the foreground have produced evidence of good independence for $\nabla^2 G$ filters, as in figure 6 (b).

4 Conditioned Foreground Likelihood: Warping and Illumination Modelling

It was demonstrated in section 1 that greater selectivity is needed in the foreground model. Generally this can be approached by reducing the degree of pooling in the learning of $p^{\mathcal{F}}$. A previous attempt at this inhibited spatial pooling by subdivision, but this is not altogether satisfactory, as explained in the previous section. The alternative investigated here simultaneously diminishes both warp pooling and spatial pooling. It involves warping a template image $\bar{I}$, onto the test image I and taking the warped $T_X(\bar{I})$ to be the mean of the distribution for I. This warping scheme is described in the next section, together with a further elaboration to take account of illumination variations.

4.1 Approximating Warps

Two-dimensional warps T_X could be realised with some precision, as thin plate splines [8]. A more economical, though approximate, approach is proposed here. First the warped outline contour is represented as a parametric spline curve [2], over a configuration-space $\mathcal{X}$, define to be a sub-space of the spline space. Then the warp of the *interior* of the object is approximated as an affine transform by *projecting* the configuration X onto a space of planar-affine transformations [7, ch 6]. The fact that this affine transformation warps the interior only approximately is absorbed by pooling approximation error, during learning, into the foreground distribution $p^{\mathcal{F}}$. The resulting warp of the interior then loses some *specificity* but is still "fair" in that the variability is fairly represented by probabilistic pooling. (A similar approach was taken with pooled camera calibration errors in mosaicing [24].)

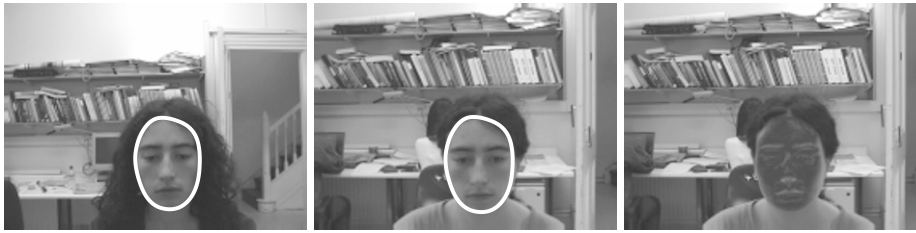
To summarise, the warp model is bipartite: an accurate mapping of outline contour coupled with an approximate (affine) mapping of the interior. The precision of the mapped contour ensures that foreground/background discrimination is accurate, and this is essential for precise contour localisation. The approximate nature of the interior mapping is however acceptable because it is used only for intensity compensation in which, especially with large filter scale σ, there is some tolerance.

4.2 Single Template Case

Given a hypothesised warp T_X, the output $z(\mathbf{x})$ of a filter $W_\mathbf{x}$ centred at $\mathbf{x}$ is modelled as

$$z(\mathbf{x}) = \left\langle W_\mathbf{x}, T_X \cdot \bar{I} + \mathbf{n} \right\rangle = \left\langle W_\mathbf{x}, T_X \cdot \bar{I} \right\rangle + \langle W_\mathbf{x}, \mathbf{n} \rangle \\ = \tilde{z}(\mathbf{x}, X) + Y_\mathbf{x} \tag{6}$$

where $\tilde{z}(\mathbf{x}, X)$ is the predicted filter output and where $Y_\mathbf{x}$ is a random variable, whose distribution is to be learned, assumed to be symmetric with zero mean. It is the residue (6) of the predicted intensity from the image data and is likely to have a narrow distribution if prediction is reasonably effective as in figure 5. Thus the distribution p_Y is far more restrictive than $p^\mathcal{F}$. Using the $Y_\mathbf{x}$'s instead of the $z(\mathbf{x})$'s in the calculation of the global likelihood $p(Z|X)$ results in more powerful and specific detection.



(a) Template (b) Image Data (c) Differenced Image

Fig. 5. Template subtraction.(a) The white contour marks the outline of the intensity template $\bar{I}$. When subtracted from an image I (b), the residue (c) is relatively small, as indicated by the dark area over the face.

Note that the predicted output $\tilde{z}(\mathbf{x}, X)$ can be approximated as

$$\tilde{z}(\mathbf{x}, X) \approx (T_X \cdot W_\mathbf{x} * \bar{I})(\mathbf{x})$$

which is computationally advantageous as the filtered template $W_\mathbf{x} * \bar{I}$ can be computed in advance. The approximation is valid provided T_X is not too far from being a Euclidean isometry. (An affine transformation, which is of course non-Euclidean, will change a circular filter support S, and this generates some error.)

4.3 Light Source Modelling

A family of templates $\bar{I}_1, \ldots, \bar{I}_K$ is generated corresponding to K lighting conditions, and typically $K = 4$ to span a linear space of shadow-free, Lambertian surfaces under variable lighting [5]. So the image data is can be modelled as $I = T_X(\boldsymbol{\alpha} \cdot \bar{\mathbf{I}}) + \mathbf{n}$. Now

the predicted filter outputs are defined to be

$$\tilde{z}(\mathbf{x}, X, \boldsymbol{\alpha}) = \left\langle W_{\mathbf{x}}, T_X(\boldsymbol{\alpha} \cdot \bar{\mathbf{I}}) \right\rangle = \sum_k \alpha_k \left\langle W_{\mathbf{x}}, T_X \cdot \bar{I}_k \right\rangle$$
$$= \sum_k \alpha_k \tilde{z}_k(\mathbf{x}, X) \tag{7}$$

Illumination modelling in this way makes for better prediction allowing the distribution of the residual p_Y to become even narrower (see figure 8).

4.4 Joint and Marginal Distributions for Illumination-Compensated Foreground

In order to preserve the validity of (5), the independence of the $Y_{\mathbf{x}}$ for sufficiently separated $\mathbf{x}$ should be checked. For instance, the correlation

$$C[\mathbf{x}, \mathbf{x}'] = \mathcal{E}[Y_{\mathbf{x}}, Y_{\mathbf{x}'}].$$

should $\to 0$ sufficiently fast as $|\mathbf{x} - \mathbf{x}'|$ increases. As figure 6 shows, the correlation

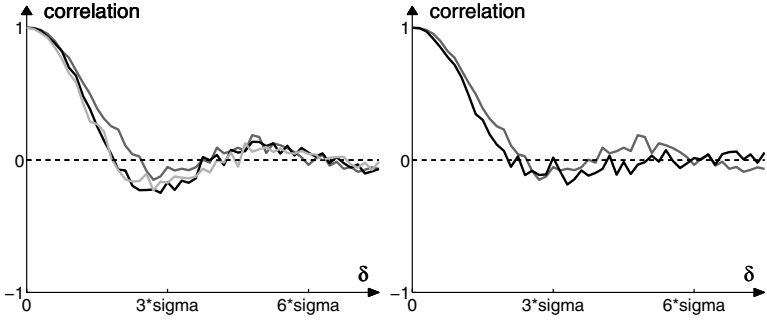


Fig. 6. Foreground correlation. The correlation between filter outputs at various displacements is shown (black) for $Y_{\mathbf{x}}$, the resuidual between the image data and the template and this is very similar to the correlation of the $z(\mathbf{x})$ (grey), and the $Y'_{\mathbf{x}}$ obtained by taking illumination factors into account (light grey). Right: the foreground correlation (grey) is similar to background correlation (black).

has fallen close to zero at a displacement of r, giving independence of adjacent outputs for the support-tessellation of figure 2. Correlation functions for foreground and background are broadly similar and so fit the same grid of filters. Finally, de-correlation is a necessary condition for statistical independence but is not sufficient. Independence properties can be effectively visualised via the conditional histogram [25]. Figure 7 displays histograms which estimate $p(Y_{\mathbf{x}}, Y_{\mathbf{x}'} \mid |\mathbf{x} - \mathbf{x}'| = \delta)$ where $\delta = \sigma, 2\sigma, 3\sigma$ and $\mathbf{x}$ and $\mathbf{x}'$ are diagonally displaced ($r = 3\sigma$). The greylevel in each histogram represents the frequency in each bin. White indicates high frequency and black none. From these it is clear that at the grid separation $r = 3\sigma$, $Y_{\mathbf{x}}, Y_{\mathbf{x}'}$ are largely independent. It might

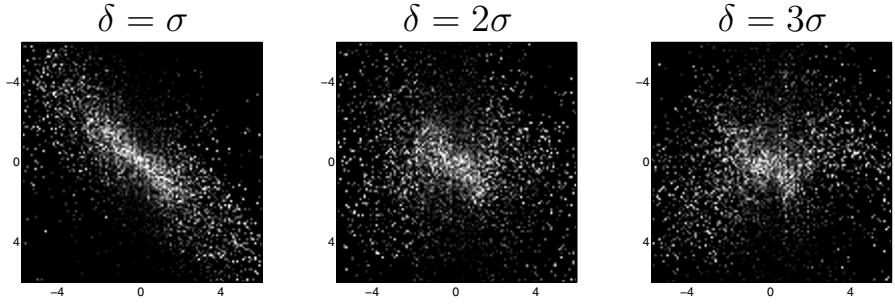


Fig. 7. Joint conditional histograms of pairs of filter responses. As δ increases the structure of the histograms decreases. When $\delta = \sigma$ the white diagonal ridge indicates the correlation between the filter responses. While at $\delta = 3\sigma$ this ridge has straightened and diffused. The two rows of figures are extremely similar and show that the template subtraction and illumination compensation have at most a marginal effect as regards whitening the data.

have been expected that template subtraction, especially with illumination compensation, would have significantly decreased correlation of the foreground but that was not the case. Where there is a significant effect is in the marginal distribution for $p^{\mathcal{F}}$ which becomes significantly narrower, as figure 8 shows.

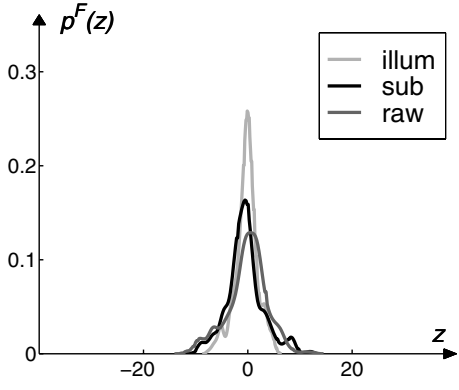


Fig. 8. Illumination compensation narrows $p_Y(Y_{\mathbf{x}})$. Each of the graph displays p_Y or $p^{\mathcal{F}}(z)$ learnt from data at different stages of preprocessing, Grey:raw filter responses ($p^{\mathcal{F}}(z)$), Black:template subtracted residual responses and Light Grey:template subtracted plus illumination compensated residual responses.

This is a measure of the increased selectivity of modelling the foreground with template subtraction, especially when this is combined with illumination compensation.

5 Learning and Inference

The goal is to infer the value of X from $p(X|Z)$ via Bayes' Rule and the constructed likelihood function $p(Z|X)$. If the test data was labelled with the value of the $\boldsymbol{\alpha}$ of the illumination/object inference would be straight forward. Modelling the illumination results in the fact that we have instead $p(Z|X, \boldsymbol{\alpha})$. In principle the correct way to proceed would be to integrate $\boldsymbol{\alpha}$ out of $p(Z|X, \boldsymbol{\alpha})$ to construct

$$L(X) = p(Z|X) = \int_{\boldsymbol{\alpha}} p(Z|X, \boldsymbol{\alpha}) p_0(\boldsymbol{\alpha}|X) d\boldsymbol{\alpha} \tag{8}$$

However, due to the probable dimensionality of $\boldsymbol{\alpha}$ and the computational expense of exhaustively calculating $p(Z|X, \alpha)$ it is not feasible to compute this integration numerically. In fact maximisation of $p(Z|X, \boldsymbol{\alpha})$ over $\boldsymbol{\alpha}$ in place of integration is an well known alternative that is simply an instance of the model selection problem. A factor $G(Z, X)$ known as the "generacity" factor (and has elsewhere been known as the "Occam" factor [20]) is a measure of robustness [12] of the inferred $\hat{\boldsymbol{\alpha}}$ — the stability of Z with respect to fluctuations in $\boldsymbol{\alpha}$:

$$G(Z, X) = \frac{\int_{\boldsymbol{\alpha}} p(Z|X, \boldsymbol{\alpha}) p_0(\boldsymbol{\alpha}|X) \mathrm{d}\, \boldsymbol{\alpha}}{p(Z|X, \hat{\boldsymbol{\alpha}}(X, Z))} \tag{9}$$

The generacity G is then the additional weight that would need to be applied to the maximised likelihood

$$\hat{L}(X) \equiv L(X, \hat{\boldsymbol{\alpha}})$$

to infer the posterior distribution for X:

$$p(X|Z) \propto \hat{L}(X) G(Z, X) p_0(X). \tag{10}$$

If $G(Z, X)$ does not vary greatly then it is reasonable to use $\hat{L}(X)$ instead of $p(Z|X)$.

5.1 MLE for Illumination Parameters

As stated it has been assumed that the residual variable Y_x in 6 is drawn from the stationary distribution p_Y. The likelihood function for particular values of X and $\boldsymbol{\alpha}$ is the product of three separate components, the likelihood of the hypothesised background, foreground and mixed measurements as:

$$\begin{aligned} L(X, \boldsymbol{\alpha}) = p(Z|X, \boldsymbol{\alpha}) &= \prod_{i \in \mathcal{I}_F} p_Y(Y_{\mathbf{x}_i, \boldsymbol{\alpha}}) \prod_{i \in \mathcal{I}_B} p^{\mathcal{B}}(z_i) \prod_{i \in \mathcal{I}_M} p(z_i|\rho(X)) \\ &= L_F(X, \boldsymbol{\alpha}) L_B(X) L_M(X) \end{aligned} \tag{11}$$

where $\mathcal{I}_{\{F,B,M\}}$ are the sets containing the foreground,background and mixed measurements. In the implementation of template subtraction and illumination compensation only the foreground measurements are affected. Therefore only L_F is dependent upon $\boldsymbol{\alpha}$.

Intuitively it would seem reasonable to solve for $\boldsymbol{\alpha}$ by maximising L_F in 11 with respect to $\boldsymbol{\alpha}$:

$$\hat{\boldsymbol{\alpha}}(Z, X) = \arg\max_{\boldsymbol{\alpha}} L_F(X, \boldsymbol{\alpha}) \tag{12}$$

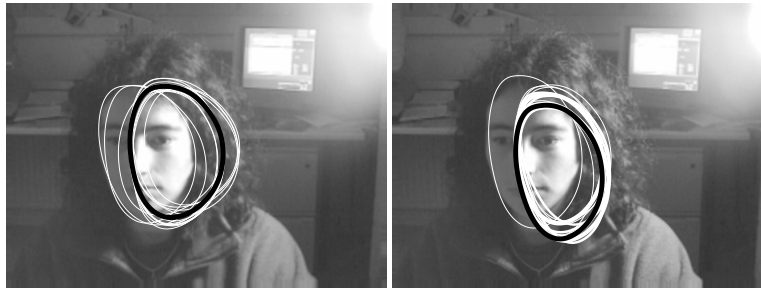
and then proceed with $\boldsymbol{\alpha}$ fixed as $\hat{\boldsymbol{\alpha}}$ and $p(Z|X, \boldsymbol{\alpha})$ is used. A functional form of p_Y is needed though in order to be able to differentiate equation 12. From figure 8 it is plausible to assume that p_Y is a zero mean Gaussian with variance γ^2. It then follows that

$$L_F(X, \boldsymbol{\alpha}) \sim \text{MVN}(\mathbf{0}, \gamma^2 I_{K\times K}) \tag{13}$$

where $I_{K\times K}$ is the identity matrix and MVN stands for the multi-variate normal distribution. Obviously maximisation of equation 13 is equivalent to the least squares minimisation

$$\hat{\boldsymbol{\alpha}}(Z, X) = \arg\min_{\boldsymbol{\alpha}} \sum_n (z(\mathbf{x}_i) - \tilde{z}(\mathbf{x}_i, X, \boldsymbol{\alpha}))^2 \tag{14}$$

Thus in the factored sampling algorithm for inferring X the following is implemented. For each hypothesis X_h a corresponding MLE $\hat{\boldsymbol{\alpha}}_h$ is calculated and the likelihood $L(Z|X)$ is approximated by $L(Z|X, \hat{\boldsymbol{\alpha}}_h)$.



No illumination Modelling

Illumination Modelling

Fig. 9. Illumination modelling improves detection results. Layered sampling at two levels (r=40 and 20 pixels) with the *conditioned* foreground likelihood model in which illumination is not modelled and it is. In the latter case $\boldsymbol{\alpha}$ is inferred by its MLE value. The gross change in the illumination conditions foils the naive *conditioned* foreground likelihood.

For the experiments performed in figure 9, the MLE method is used to infer $\boldsymbol{\alpha}$. However, it remains to be confirmed experimentally that $G(Z|X)$ remains more or less constant. In this experiment a shadow basis was formed by taking three images with the point light source to the left, right and behind of the subject. Then a sequence of 267 frames in which the light source moved around the subject was used as test data. In every fifth frame the face person was searched for using layered sampling with the *conditioned* foreground likelihood, independent of the results from the previous search. Two levels of layered sampling were applied ($r = 40, 20$) and 900 samples at each level. The prior for the object's affine configuration space was uniform over $x, y-$translation and Gaussian over the other parameters allowing the contour to scale to $\pm 20\%$ horizontally, vertically or diagonally and rotate 20 degrees from its original position. (Each of the 6 parameters were treated independently). Using the proposed method the face was successfully located at each frame. However, when illumination was not modelled detection was not always successful. Two frames in which this happened are shown in figure 9. To see the results of the whole sequence please see **http://www.robots.ox.ac.uk/ sullivan/Movies/FaceIlluminated.mpg**.

5.2 Sampling Illumination Parameters

In the previous subsection a method for inferring $\boldsymbol{\alpha}$ was described. This method though is not Bayesian. The alternative is to extend the state vector to $X' = (X, \boldsymbol{\alpha})$ and to sample this in order to obtain a particle estimate of $p(X'|Z)$. This however, is likely to be computationally burdensome because of the increased dimensionality and also due to the broad prior from which $\boldsymbol{\alpha}$ must be drawn. Usually no particular prior for $\boldsymbol{\alpha}$ will be known and in accordance a uniform one will be generally used.

The alternative is to use an importance sampling function [18] $g_X(\boldsymbol{\alpha})$ that restricts $\boldsymbol{\alpha}$ to its likely range. It is possible to incorporate this importance function into the factored sampling process as follows. Draw a sample X_h from $p_0(X)$. Given this fixed value of X draw a sample $\boldsymbol{\alpha}_h$ from $g_{X_h}(\boldsymbol{\alpha})$. The corresponding weight associated with the particle $X'_h = (X_h, \boldsymbol{\alpha}_h)$ is $L(X'_h)/g_{X_h}(\boldsymbol{\alpha}_h)$ (the denominator is the correction factor applied to compensate for the bias shown towards certain $\boldsymbol{\alpha}$ values).

The most important question has yet to be answered. From where can an appropriate importance function $g(\boldsymbol{\alpha})$ be found ? In fact we don't have to look any further than the partial likelihood function L_F. From equation 13 this can be approximated by a multivariate normal distribution with diagonal covariance matrix. This then implies that $\boldsymbol{\alpha}$ given a fixed value of X is also a multi-variate normal distribution whose covariance matrix and mean can be easily calculated. Allowing this distribution to be $g(\boldsymbol{\alpha})$ results in an importance function that can be sampled from exactly and greatly narrows the range of possible $\boldsymbol{\alpha}$ values.

6 Results and Conclusions

Results of localisation by factored sampling, using the new *conditioned* foreground likelihood are shown next, and compared with the "partitioned foreground" approach

of [26]. Each figure displays an image plus the particle representation of the posterior distribution for the configuration of the target object. For clarity, just the 15 most highly weighted particles are displayed. The weight of each particle being represented, on a log scale, by the width of the contour. The black contour represents the mean configuration of the particle set. Three different sets of experiments were carried out. Firstly it was checked if the new likelihood was prone to highlighting the same false positives as the partitioned likelihood and this is investigated with the decoy test. Then does the new method work for face detection and finally can it detect other textured objects.

The decoy test In figure 1, it was shown using a face decoy that the partitioned foreground model was prone to ghost object hypotheses. Results of this experiment with the new, conditioned foreground likelihood are shown in figure 10. Note the effect on the mean configuration (the black contour): for the partitioned foreground, the mean lies between the two peaks in the posterior. With the conditioned likelihood the posterior is unimodal however, as evidenced by the coincidence of the mean configuration with the main particle cluster. Experiments were carried out at one scale level $r = 40$ and using 1200 particles, uniformly distributed, the translational component of the prior being drawn deterministically (ie on a regular grid), for efficiency. For computational

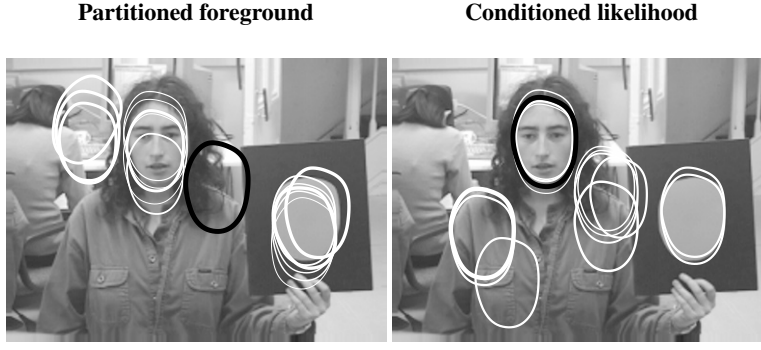


Fig. 10. Conditioned foreground likelihood eliminates ghosting. Foreground partitioning produces a bimodal posterior distribution (plainly visible from the position of the mean contour) while conditioned foreground gives a unimodal distribution.

efficiency, multi-scale processing can be applied via "layered sampling" [26] and this is demonstrated with person-specific models, for two different people, in figure 11. The prior for the affine configuration space is uniform over $x, y-$translation and Gaussian over the other parameters allowing the contour to scale to $\pm 20\%$ horizontally, vertically or diagonally and rotate 20 degrees from its original position. (Each of the 6 configuration space parameters are treated independently) This prior is used for the rest of the experiments unless otherwise stated.

Fig. 11. Layered sampling demonstrated for individuals, using individual-specific models this figure. The prior for the position in the face is uniform in the $x, y-$translation over the image. The search takes place over two scales ($r = 20, 10$) implemented via layered sampling, using 1500 samples in each layer.

Generalisation Experimentally, a model trained on one individual turns out to be capable of distinguishing the faces of a range of individuals from general scene background. The experiment used the learnt model from figure 11 (a) and applied it to the images displayed in figure 12. Once again two levels of layered sampling were applied ($r = 20, 10$), now increasing the number of samples increased to 3000. This performance is achieved

Fig. 12. Generalisation of face detection. Training on a single face generates a model that is still specific enough to discriminate each of a variety of faces against general scene background.

without resorting to the more complex, multi-object training procedure of 7.1, though it remains to test what improvements in multi-object training would bring.

Detecting various textured objects Finally, the conditioned foreground likelihood model has been tested on a variety of other objects, as in figure 13. Note that even in the case of a the textured vase resting against a textured sofa, the vase object is successfully localised. Given that the boundary edge of the vase is not distinct, edge based methods would not be expected to work well here. (Layered sampled was applied at scales $r = 20, 10$ pixels with 1200 particles in each layer.) The prior in the clown

example allows for a greater rotation, while in the shoe example the prior has been narrowed.

Fig. 13. Textured inanimate objects can also be localised by the algorithm. Special note should be taken of the detection of the vase against the textured sofa.

7 Discussion and Future Work

7.1 Modelling Object Variability

In addition to lighting variations, a further generalisation is to allow object variations. For example, in the case of faces, varying physiognomy and/or expression. This could be dealt with in conventional fashion [19] by training from a set $I_1^*, I_2^*, \ldots$ covering both object and illumination variations, and using Principal Components Analysis (PCA) to generate templates $\bar{I}_1, \ldots, \bar{I}_K$ that approximately spans the training set. Then the methodology of the previous section can be followed as before.

Alternatively, it may be the case that the training set is explicitly labelled with illumination conditions $k = 1, \ldots, K$ and basis-object index $j = 1, \ldots, M$, in which case the training set is organised as $\{I_{jk}\}$ and these could be used directly as templates $\{\bar{I}_{jk}\}$. Then a general image is

$$I = \sum_{j,k} \alpha_{jk} \bar{I}_{jk} = \sum_{j,k} \beta_j \gamma_k \bar{I}_{jk}$$

where γ_k weights light-sources and β_j weights basis objects. Thus the KM weights α_{jk} applied to the templates decompose as $\beta_j \gamma_k$, and so have just $K + M$ degrees of freedom. This is a familiar type of bilinear organisation, the "style and content" decomposition [13], that occurs also with the decomposition of facial expression and pose [3]. Imposing the bilinear constraint that $\boldsymbol{\alpha} = \boldsymbol{\beta}\boldsymbol{\gamma}^\top$, which stabilises the estimation of α, can be performed as usual by SVD.

In this bilinear situation, the earlier model (7) is extended to take account of light source variations as follows.

$$\tilde{z}(\mathbf{x}, X, A) = \sum_{j,k} \alpha_{jk} \tilde{z}_{j,k}(\mathbf{x}, X) \tag{15}$$

where

$$\tilde{z}_{j,k}(\mathbf{x}, X) = \left\langle W_{\mathbf{x}}, T_X \cdot \bar{I}_{j,k} \right\rangle .$$

and A is a matrix whose entries are α_{jk}.

Acknowledgements. We are grateful for the support of the EU (JR) and the EPSRC (JS).

References

1. R. Baddeley. Searching for filters with interesting output distributions: an uninteresting direction to explore?. *Network*, 7(2):409–421, 1996.
2. R.H. Bartels, J.C. Beatty, and B.A. Barsky. *An Introduction to Splines for use in Computer Graphics and Geometric Modeling*. Morgan Kaufmann, 1987.
3. B. Bascle and A. Blake. Separability of pose and expression in facial tracking and animation. In *Proc. 6th Int. Conf. on Computer Vision*, pages 323–328, 1998.
4. B. Bascle and R. Deriche. Region tracking through image sequences. In *Proc. 5th Int. Conf. on Computer Vision*, pages 302–307, Boston, Jun 1995.
5. P.N. Belhumeur and D.J. Kriegman. What is the set of images of an object under all possible illumination conditions. *Int. J. Computer Vision*, 28(3):245–260, 1998.
6. A.J. Bell and T.J. Sejnowski. Edges are the independent components of natural scenes. In *Advances in Neural Information Processing Systems*, volume 9, pages 831–837. MIT Press, 1997.
7. A. Blake and M. Isard. *Active contours*. Springer, 1998.
8. F.L. Bookstein. Principal warps:thin-plate splines and the decomposition of deformations. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 11(6):567–585, 1989.
9. P.J. Burt. Fast algorithms for estimating local image properties. *Computer Vision, Graphics and Image Processing*, 21:368–382, 1983.
10. T.F. Cootes, C.J. Taylor, D.H. Cooper, and J. Graham. Active shape models — their training and application. *Computer Vision and Image Understanding*, 61(1):38–59, 1995.
11. D.J. Field. Relations between the statistics of natural images and the response properties of cortical cells. *J. Optical Soc. of America A.*, 4:2379–2394, 1987.
12. W.T. Freeman. The generic viewpoint assumption in a framework for visual perception. *Nature*, 368:542–545, 1996.
13. W.T. Freeman and J.B. Tenenbaum. Learning bilinear models for two-factor problems in vision. In *Proc. Conf. Computer Vision and Pattern Recognition*, pages 554–560, June 1997.
14. D. Geman and B. Jedynak. An active testing model for tracking roads in satellite images. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 18(1):1–14, 1996.
15. U. Grenander. *Lectures in Pattern Theory I, II and III*. Springer, 1976–1981.
16. U. Grenander, Y. Chow, and D.M. Keenan. *HANDS. A Pattern Theoretical Study of Biological Shapes*. Springer-Verlag. New York, 1991.
17. G.D. Hager and K. Toyama. Xvision: combining image warping and geometric constraints for fast tracking. In *Proc. 4th European Conf. Computer Vision*, pages 507–517, 1996.
18. J.M. Hammersley and D.C. Handscomb. *Monte Carlo methods*. Methuen, 1964.
19. A. Lanitis, C.J. Taylor, and T.F. Cootes. A unified approach to coding and interpreting face images. In *Proc. 5th Int. Conf. on Computer Vision*, pages 368–373, 1995.
20. D.J.C. MacKay. *Bayesian Methods for Adaptive Models*. PhD thesis, California Institute of Technology, 1992.

21. D. Mumford. Pattern theory: a unifying perspective. In D.C. Knill and W. Richard, editors, *Perception as Bayesian inference*, pages 25–62. Cambridge University Press, 1996.
22. B.A. Olshausen and D.J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381:607–609, 1996.
23. P. Perona. Steerable-scalable kernels for edge detection and junction analysis. *J. Image and Vision Computing*, 10(10):663–672, 1992.
24. S.M. Rowe and A. Blake. Statistical mosaics for tracking. *J. Image and Vision Computing*, 14:549–564, 1996.
25. E.P. Simoncelli. Modeling surround suppression in V1 neurons with a statistically-derived normalization model. In *Advances in Neural Information Processing Systems*, volume 11, page in press. MIT Press, 1997.
26. J. Sullivan, A. Blake, M. Isard, and J. MacCormick. Object localization by bayesian correlation. In *Proc. 7th Int. Conf. on Computer Vision*, volume 2, pages 1068–1075, 1999.
27. P. Viola and W.M. Wells. Alignment by maximisation of mutual information. In *Proc. 5th Int. Conf. on Computer Vision*, pages 16–23, 1993.
28. S.C. Zhu and D. Mumford. GRADE: Gibbs reaction and diffusion equation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 19(11):1236–1250, 1997.