Flexible Template and Model Matching Using Image Intensity

Bernard F. Buxton, Vasileios Zografos
University College London
Department of Computer Science
Gower Street, London, United Kingdom
{B.Buxton, V.Zografos}@cs.ucl.ac.uk

Abstract

Intensity-based image and template matching is briefly reviewed with particular emphasis on the problems that arise when flexible templates or models are used. Use of such models and templates may often lead to a very small basin of attraction in the error landscape surrounding the desired solution and also to spurious, trivial solutions. Simple examples are studied in order to illustrate these problems which may arise from photometric transformations of the template, from geometric transforms of it or from internal parameters of the template that allow similar types of variation. It is pointed out that these problems are, from a probabilistic point of view, exacerbated by a failure to model the whole image, i.e. both the foreground object or template and the image background, which a Bayesian approach strictly requires. Some general remarks are made about the form of the error landscape to be expected in object recognition applications and suggestions made as to optimisation techniques that may prove effective in locating a correct match. These suggestions are illustrated by a preliminary example.

1 Introduction

Object recognition and image matching are amongst the most familiar and long standing activities in computer vision and image processing and are, for example discussed in such long-standing textbooks as [22] and [37]. Up to this day, there exist many applications in a variety of different areas ranging from: image stabilisation [15, 30] and the registration of medical images [3, 16] with standard images from a medical encyclopaedia [20], with similar images [53], or across imaging modalities [28]; to the detection of objects and even people in video [39, 27], for security [29] and inspection (image matching) in manufacturing and industrial quality control [51, 33].

Most researchers would feel they know how to tackle such tasks and would select from a variety of feature and pixel matching methods depending on the characteristics of the imagery and of the objects of interest, and application requirements. As expected from the scope of the techniques used and applications considered a huge variety of match and mismatch scoring criteria have been employed. Even within the subset of methods which employ some form of pixel matching, there is itself a large variety of matching metrics to which the researcher can refer [36]. Thus, for example for grey-level imagery, if the intensities in the images or within the objects to be matched are expected to be similar except for small deviations caused by random noise, a sum of squared differences (SSD) measure might be used. However, if the deviations are large and may include gross deviations a sum of absolute differences (SAD) measure would be preferable as it is more robust. If on the other hand there are occlusions, such measures might be weighted [25] or 'gated' only to apply to the visible regions of the object [34]. Colour adds a further range of choices to the variety. Finally we note that, when the pixels in the images or objects to be matched do not correspond in a way that can be represented by means of a simple mapping function, mutual information measures are often used, which can nevertheless detect the underlying patterns of commonality [52].

In the remainder of this paper, in Section 2 we discuss briefly how our recent work on the development of implicit 3D or view-based models of three-dimensional objects has exposed a number of unexpected difficulties, in particular, that the basin of attraction surrounding the optimal match solution is very small. Such difficulties have prompted us to examine closely traditional template matching approaches. In Section 3, we show that the smallness of the basin of attraction is a difficulty also in traditional template matching approaches and in Section 4 that incorporation of extrinsic variations by means of photometric or geometric transformations of a template exacerbates the problem and can lead to spurious, trivial solutions. From a probabilistic point of view many of these deficiencies are related to a failure to

model the whole image, i.e. both the foreground and background. Following a very brief summary of the traditional ways of dealing with photometric variability in Section 5 we turn in Section 6 to the importance of modelling the background and indicate some possible ways of doing so. A simple model is then introduced in Section 7 to illustrate how such modelling alleviates some of the difficulties and inconsistencies discussed and some general remarks are made in Section 8. A brief discussion of some of the optimisation methods that it would therefore seem beneficial to use is given in Section 9 together with some illustrative results. Conclusions are drawn in Section 10.

2 Image-based object modelling

In recent work, we have been interested in the development of iconic, image-based models of three-dimensional objects imaged by means of ordinary cameras. In constructing such models we have to take into account the main factors which might influence an object's appearance, in particular including:

1. variations in viewpoint, which can drastically affect the apparent shape of an object;

2. variations in the illumination of an object, which can change its apparent colour and shading and, even its apparent texture;

3. variations in the state of the object itself, which could be as extreme as articulation and changes in reflectance or the emission of light, or more restricted like the change in shape of the human head from individual to individual, or changes brought about by growth and aging or by facial expression;

4. variations in the camera geometry, number of colour channels, and sensor spectral sensitivity.

Which of these is dominant depends on the application and characteristics of the objects and scenes of interest. However, in many cases camera variations 4) might be considered the least serious since we are, in principle, in a position to know, measure or learn something about these. In indoor environments, illumination 2) is also frequently controllable, and even in outdoor applications it might be measured or its variability may be characterised from training examples. If the object has internal degrees of freedom and especially if it can change its state of its own volition, 1) and 3) may be the most difficult or impossible to control.

Our previous work has thus focused on the representation and characterisation of such effects, in particular those arising from geometrical changes. Since characterising the geometry of a three-dimensional object which is not flat requires two or more views, we have shown how the multi-view geometry may be combined with the popular point distribution or flexible shape models introduced by Cootes and Taylor [8] to produce an integrated shape and pose model (ISPM)) [4, 14].

Having achieved this first step, which may be regarded as a proper parameterisation of the linear combination of views (LCV) object representation of Ullman and Basri [50] and integration of it with the flexible shape model (FSM) of Cootes and Taylor, we expected that proceeding to a pixel-based integrated appearance and pose model (IAPM) would be straightforward.

1. Extension of the FSM to a flexible appearance model (FAM), though it has proceeded via a number of variants [26, 6], is now well established and straightforward [7].

2. The LCV has itself been extended so that images rather than geometric arrays of landmark points or line drawings can be represented, with results that are frequently very good, as illustrated in Figure 1, and can be utilised to interpolate movie sequences from a small number of snapshots [19, 18].

Preliminary work [41], however, quickly showed that, even for the LCV alone, this was not the case and that the basin of attraction within which such a procedure would converge was very small and that the approach was therefore sensitive to the initialisation of the algorithm.

3 Comparison with other approaches

Similar effects are encountered in the use of flexible appearance models, but may often be overcome by use of standard multi-resolution techniques and image pyramid representations [10, 9]. Even though in an FAM the object model may possess many (over a 100) internal (or as we termed them in [4, 13, 14], intrinsic) geometric and (appearance) pixel degrees of freedom, the high dimensionality of this space is not a problem. The magnitude of such intrinsic variations is usually quite restricted (to the appropriate shape and appearance of the object of interest which, if a face, is usually imaged under quite well-controlled conditions). The difficulties arise from the small number of extrinsic geometric degrees of freedom in such models, such as the location of the object in the image and it is for this reason that a multi-resolution iterative search is often employed. If we regard the FAM as a flexible template representing the object in the image, we see that it is the interaction of the structure within the template and the structure within the image that causes the difficulties. Indeed, a similar effect may be seen if the geometry of the object is

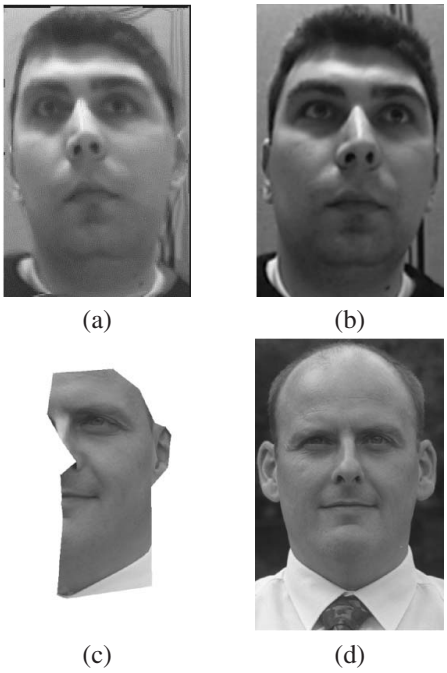


Figure 1. Example face images constructed by using variants of the Ullman and Basri LCV technique, (a) from the early work of Koufakis and Buxton [24] and (c) an improved version from the work of Kennedy et al. [23] applied to a colour image. The original or target images which here we are modelling by use of a linear combination of the basis views are shown in (b) and (d). For details see the references given.

constrained to that of a square and we just use a rigid template to represent the intensity variation within this patch. The resulting SSD difference windowed matching score or squared error is illustrated in Figure 2. It has a small basin of attraction, at the bottom of which the SSD is reduced almost to zero, surrounded by a rugged landscape with many local summits, immits and cols.

4 The difficulty of using a flexible template with additional extrinsic variations

The discussion in Section 3 begins to illustrate some of the difficulties that we might encounter in our bid to develop an IAPM. In fact, the difficulties encountered were much more severe. The imaged object has many more extrinsic degrees of freedom. Furthermore, the range of this extrinsic variation allows the image object model to adopt a wide variety of shapes and appearances.

Thus, for example, the centred affine trifocal tensor (the 'CATT') which correctly describes the geometrical variation of the shape of the image object has 9 degrees of freedom [48] which, since the object is centred, are in addition to the two translational degrees of freedom discussed in Section 3 above. Similarly, linear models of the photometric variation in an object's appearance in an image [46, 42] introduce at least 3 degrees of freedom even if individual pixel (texture) variations are ignored [12].

The main problem however, is not the number of additional degrees of freedom, but the range of variation the models permit, i.e. their lack of specificity. To illustrate such problems, we will consider the simpler problem of matching a flexible template $I_m^{'}(x^{'}, y^{'})$, say, to an image $I(x, y)$,under affine photometric (grey-level transformations) and affine geometric transformations [17], of the kind:

$$I_m(x^{'}, y^{'}) = aI_m^{'}(x^{'}, y^{'}) + b \tag{1}$$

$$\begin{aligned} x &= a_0 + a_1 x^{'} + a_2 y^{'} \\ y &= b_0 + b_1 x^{'} + b_2 y^{'} . \end{aligned} \tag{2}$$

In (1), $I_m^{'}(x^{'}, y^{'})$ and $I_m(x^{'}, y^{'})$ stand respectively for the template intensities at pixel $(x^{'}, y^{'})$ before and after the photometric transformation whilst in (2) the pixel co-ordinates $x^{'}, y^{'}$ before the geometric transformation are mapped into image co-ordinates (x, y). The net effect of the two transformations is to map $I_m^{'}(x^{'}, y^{'})$ into $I_m(x, y)$.

This loses the degrees of freedom associated with the view geometry of 3-dimensional objects and is a far from sufficient model of the potential photometric variations of such objects but suffices, even with further specialisation as indicated below, to elucidate a number of important points. First however, we note that, if our matching criterion is a SSD error measure, i.e. we

$$\min \left\{ \sum_{x,y} (I(x, y) - aI_m^{'}(x, y) - b)^2 \right\}, \tag{3}$$

minimisation over the parameters (a, b) of the photometric transformation may be carried out analytically and the result written in the following form:

$$\min \left\{ \left\langle \Delta I^2 \right\rangle (1 - r^2) \right\}, \tag{4}$$

where $\langle \ldots \rangle$ stands for a sum or average over the pixels (x, y), $\Delta I = I - \langle I \rangle$ and r is the correlation coefficient defined as:

$$r = \frac{\langle \Delta I \Delta I_m \rangle}{\sqrt{\langle \Delta I^2 \rangle \langle \Delta I_m^2 \rangle}} . \tag{5}$$

Except for the term in $\left\langle \Delta I^2 \right\rangle$, (4) is one of many familiar image matching criteria whose performance in template matching have been evaluated several times [49, 3]. Other

familiar forms in which the deviations from the mean intensity are used, or the intensities normalised for the image brightness or level of illumination may similarly be derived by using the photometric transformations which respectively include only the bias b or gain a.

The result (4), in particular the presence of the term $\langle \Delta I^2 \rangle$ deserves closer scrutiny. First, we note that, in conventional template matching, the SSD is usually computed by summing over the pixels lying within the image area, A_m say, covered by the transformed template $I_m(x, y)$. If, the geometric transformation (2) is restricted to translation of the template and the image or images in which we are trying to find the object of interest are stationary, the variance $\langle \Delta I^2 \rangle$ is independent of the position of the template and (4) reduces simply to maximisation of the magnitude of the correlation coefficient, r. However, if the image is not stationary, $\langle \Delta I^2 \rangle$ cannot be removed from (4) without changing the matching criterion. Moreover, in such cases a number of difficulties become apparent.

1. Bland regions of the image where there is little or no variation produce good matches with little error,
2. If we retain the affine geometric transformation (2), the area A_m covered by the transformed template may under scaling or shearing shrink to zero resulting in a zero variance $\langle \Delta I^2 \rangle$ and spurious matches.

One way to remove such spurious matches is to normalise by the area A_m, but this means that the matching score becomes very noisy whenever A_m is small. Another is to introduce suitable priors which will add regularising terms to criterion (3) and bias against spurious solutions in which the template is shrunk to cover only a very small area [57]. Adopting the probabilistic viewpoint is very satisfying, but exposes a more fundamental failing of the approach outlined above. By using only the area under the transformed template in the match criterion (3), from a probabilistic viewpoint the observations we are using to test our hypothesis as to where the object is in the image (which may include the null hypothesis that the object of interest isn't present) are dependent on the parameters of our model, ie on the hypothesis. As pointed out by Sullivan et. al [46, 47], this is not allowed in a Bayesian approach. Simply put, our observation is the whole of the image and we should have a model of the background as well as of the foreground object or objects of interest. Thus, we should utilise not only positive evidence of where we are hypothesizing the object or objects may be, but also negative evidence from elsewhere in the image where the observed image intensity does not accord with our expectations for the background.

We should therefore include *all* pixels in the image in the sum in our SSD score (3). The variance $\langle \Delta I^2 \rangle$ is then evaluated over the *whole of the image area* A, say. Spatial stationarity within an image $I(x, y)$ is no longer an issue, though stationarity over the set of images of interest, for example the images in a database, or temporal stationarity of a sequence in a tracking application remain as significant questions. One nice outcome of this view is that we do not have to worry about the possibility of the variance $\langle \Delta I^2 \rangle$ vanishing unless there are trivial, totaly bland images in the data, which can easily be detected and removed. Another is that the procedure leads naturally to a mechanism for anomaly or novelty detection.

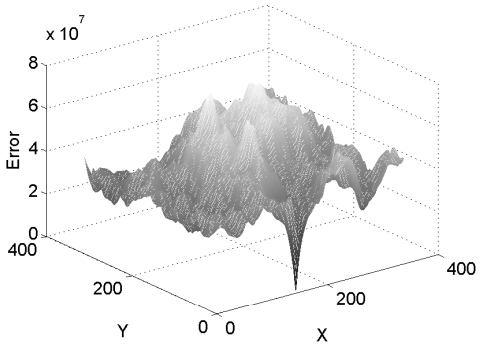


Figure 2. Example SSD error landscape for the 2-dimensional translation space, computed from a traditional windowed template matching paradigm.

5 Traditional solutions for photometric variability

A variety of approaches are used to overcome the problem that the photometry of the image and template may not match. In practice, it is rare to see explicit application of transformations such as that defined in (1), probably because doing so is equivalent to (or almost equivalent to) use of various normalised matching scores. An alternative, obvious from the linearity of the photometric transformation, is to high-pass filter the image and template prior to matching. This can bring other advantages as noted below.

If the filtering is appropriately designed and combined with a decision function, localised features may be extracted and the features matched. Examples include edge detection, which may be generic as for example in the design of a Canny edge detector [5] or tuned to particular types of edges relevant to particular applications, such as the 'valley edges' often seen on the extremal contour of a curved object such as a face [35]. The disadvantage of using localised features is that we may then be confronted with a difficult correspondence problem. Unless there is some simplifying aspect of the application, as for example in stereo

imaging or tracking, solving the correspondences may be of combinatorial complexity. Occlusion of localised features and the inevitability of features missed as false negatives and, more importantly, of distracting clutter from false positives, can severely exacerbate the problems encountered in such approaches. For example, such effects often mean that, even when there is some geometric simplification as when the epipolar constraint is employed in binocular stereo, the problem may in practice still be difficult.

6 Background modelling

One downside of the remarks in Section 4 is that we have to construct a background model as well as a foreground model for the object(s) of interest. Since the combined model will necessarily be more complicated than the foreground template model alone, the danger is that the combined models will be less applicable and therefore more fragile than a model which only includes the foreground.

We thus either have to know what the background is, build a very simple model, or have a statistical model of what it is expected to be like. In fact, it is surprisingly often the case that we know the background or may learn it. Examples include: medical applications, many monitoring and some inspection systems. Indeed, in monitoring and inspection, it is often an essential requirement that the background is known or has to be modelled [56]. In some cases, as in the CMU PIE database [43], the background has been recorded with no objects present (in this case human faces) for the convenience of researchers.

Moreover, much research has been carried out on the statistics of natural and man-made imagery [44, 21, 40] and theoretical and practical results are available that can, and should be utilised. Here, the important point is that the statistics of the intensity of such wide classes of imagery are not very predictable, but thanks to the approximate fractal-like nature of such images, the distributions of filtered versions of the images are predicable. Application of a bank of suitable filters then leads to predictable background statistics which can also, if desired, be learned anew for each application, for example by 'layered' sampling [47]. Arranging the filter banks, for example by use of appropriate wavelet filters [1], to produce pyramid representations also leads to computational efficiencies.

7 A simple model

In order to illustrate several of the above points, it is instructive to take our cue from the idea that, in order to be robust, background models should be simplistic. We thus construct a very simple example which can act as a surrogate for more realistic models, for example of a probabilistic kind. Our basic assumption is that there is an object of area A_O of constant intensity I_O in the foreground of an image $I(x, y)$ of area A which otherwise is of constant intensity I_B. The model correspondingly has a foreground object of intensity I_m of area A_m centred at (x_m, y_m) and a background intensity I_b. The model and object may have an overlap area A_{Om} as sketched in Figure 3 (a). For simplicity, given that the model contains foreground and background intensities I_m and I_b that we may vary, we shall ignore the photometric transformation (1) and, since we have not specified the size or shape of the model of area A_m, we will similarly ignore the geometric transformations (2). This also in this case avoids the possibility of having an ill-posed problem, though for more realistic examples, such transformations will in general be essential.

For our simple model, calculation of the match score such as the SSD is a matter of counting the number of pixels in, or the areas of, four contributions where: the model object overlaps the image object, the model object overlaps the image background and vice-versa, and where the two backgrounds overlap, leading to:

$$\min \left\{ \left[\begin{array}{l} (A_m - A_{Om})(I_B - I_m)^2 + \\ A_{Om}(I_O - I_m)^2 + \\ (A_O - A_{Om})(I_O - I_b)^2 + \\ (A - A_m - A_O + A_{Om})(I_B - I_b)^2 \end{array} \right] \right\}. \quad (6)$$

In (6) the area of the overlap A_{Om} is a function of the coordinates (x_m, y_m). Even for simple objects such as rectangles and circles A_{Om} is complicated and non-analytic. Optimisation over (x_m, y_m) (and in general any other model parameters determining the orientation, size, and shape of the model object, i.e. affecting A_m and A_{Om}) thus has to be carried out numerically. However, we may choose in the above whether to treat the photometric values in the model, I_m and I_b as constants or as variables, and in the latter case carry out optimisation with respect to them analytically. Thus, for a traditional rigid, windowed template, I_m would be constant and, since we only need the first two contributions in (6) from under the template, I_b is irrelevant. It follows that, in this case, (6) becomes simply:

$$\min \left\{ \left[(A_m - A_{Om})(I_B - I_m)^2 + A_{Om}(I_O - I_m)^2 \right] \right\}, \quad (7)$$

in which, if we choose the foreground and background intensities correctly to match the image, $(I_B - I_m)^2$ may be replaced by $(I_B - I_O)^2$ and $(I_O - I_m)^2$ by zero. However, if the object model intensity I_m is not fixed and we optimise (7) with respect to it we find that (7) is replaced by:

$$\min \left\{ \left[(A_m - A_{Om})(I_O - I_B)^2 A_{Om}/A_m \right] \right\}. \quad (8)$$

Whilst (7) has, as expected, a single basin of attraction of area $4A_O$ containing at its unique minimum the correct location of the object (see Figure 3 (b)), (8) does not behave

in such a nice way. There is a much smaller basin of attraction, and it is surrounded by a rim beyond which there is no overlap and the matching score becomes zero as I_m adapts to the image background level (Figure 3 (c))). This simple behaviour is symptomatic of what can happen if adaptive or flexible models are not used carefully.

Somewhat surprisingly, simply taking into account all the evidence from the whole of the image as described in Section 4 largely alleviates the problem. In this case, we need to optimise (6) with respect to both I_m and I_B which, if $A_m = A_O$, leads to:

$$\min \left\{ \begin{array}{l} (F_O - F_B)^2 (A_m - A_{Om}) \\ \left[A_{Om}/A_m + \frac{(A - A_m) - (A_m - A_{Om})}{A - A_m} \right] \end{array} \right\}. \qquad (9)$$

This has a single basin of attraction, slightly smaller than that in the examples above, with a small rim, and when there is no overlap, a plateau slightly less high than that obtained when a rigid, windowed template was used (Figure 3 (d)).

8 Some general comments

In the above, the basin of attraction has an area of approximately $4A_O$ and the landscape outside the basin is flat (see Figure 3 (b)). Structure within the object and in the background will, as illustrated in Figure 2, lead to considerable variation of the SSD outside the basin of attraction. Also, the area of the basin of attraction is larger in our simple model (probably considerably much larger) than we should expect in general because:

1. Perfect correlation of the pixel intensities with each other will not persist right across the object. The object may be patterned or have systematic variation across it that will reduce the strength of the correlation and may change its sign, and the range of the correlations is unlikely to extend fully across the object.

2. Structure in the foreground and background will tend to decrease the size of the basin of attraction and make the rim irregular. Noise will have a similar, but unless the images are very noisy, less pronounced effect.

Smoothing the image and model will tend to increase the range of the correlations and also, probably, their strength. However, neither effect is necessarily guaranteed in the sense that we can expect such increases to occur monotonically as the smoothing is increased. In general, increased smoothing will eventually tend to wash-out distinctive features on the object and structure in the background, leading to a decrease in the depth of the basin of attraction and, with enough smoothing, the merging and disappearance of some spurious basins of attraction (see Figure 4). This is the basic reason why multi-resolution pyramid methods are useful in

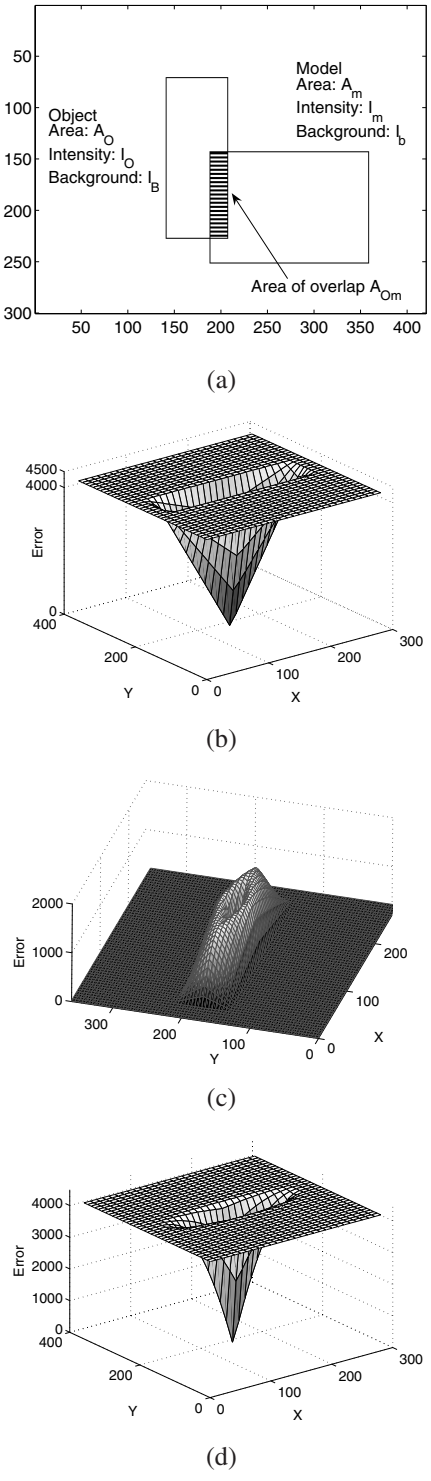


Figure 3. The simple matching example. A sketch of the object and template overlap geometry (a), the SSD error surface for (b) a traditional, fixed template matching algorithm, (c) according to equation (8), and (d) equation (9).

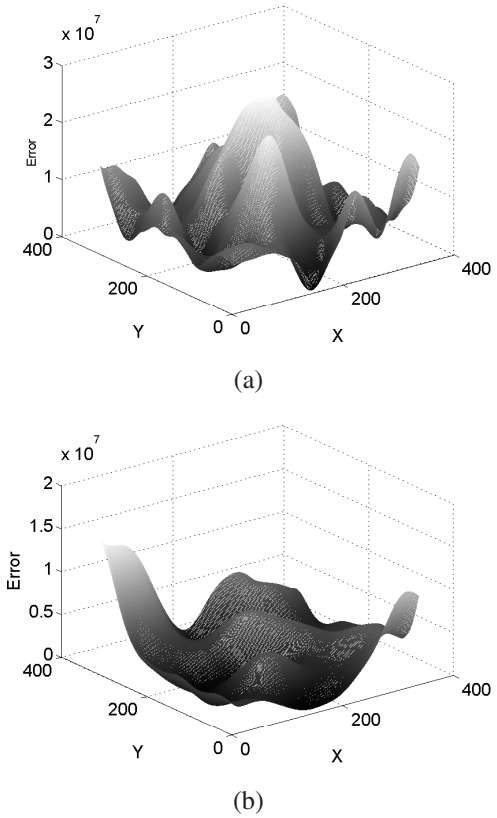


Figure 4. The effects of successive smoothing (Gaussian filter) of the image and the model. Resulting surfaces (a) with a 10x10 pixel smoothing mask and (b) with a 20x20 pixel mask

template matching. However, such methods are not foolproof, cannot be guaranteed to find the correct match to the object and involve inevitable trade-offs between usefulness and reliability.

1. Ideally, the top-most level of the pyramid within which an exhaustive search is carried out is chosen so that there is sufficient smoothing that it contains one (or just a few) basins of attraction. These basins of attraction must, however, be sufficiently well-defined for their minima to be detected and located.

2. Ideally, we can track the locations of these minima down the pyramid to the bottom-most, highest-resolution level of the image itself. However, as the smoothing is reduced at levels of higher resolution, either new basins of attraction are born and/or existing basins split. Tracking the minima can thus be problematic and we have no guarantee of arriving at the desired matching solutions in the image level.

The form of the matching error surface is also significant. If the background were bland and the image contained several instances of the object of interest, the matching error surface would consist of a plateau, containing several pits, rather like a cake-tin (see Figure 5 (a)). If the mouth of each

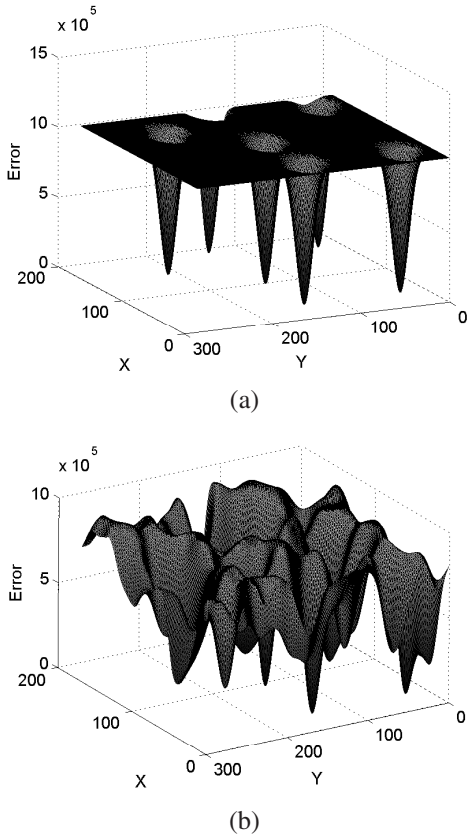


Figure 5. Error surfaces produced when there are many instances of the object in question in the image. With a bland background (a) and with clutter in the background (b). Both of these surfaces pose difficult problems for most optimisation algorithms.

of the pits were small and the plateau itself distorted into a rugged landscape by clutter and structure in the background (see Figure 5 (b)), we will have an optimisation problem in which finding the pits can be difficult. Certainly, under such circumstances, gradient descent cannot be expected to work well nor will multiple starts help, though with the caveats noted above, smoothing and use of a multi-resolution pyramid may help. Since such a landscape is not of the nested, ultrametric kind, stochastic gradient methods and simulated annealing will not necessarily work well either. Similarly, evolutionary and genetic methods may also run into difficulties, as it may be hard to ensure that the sampling will find the important areas within the pits.

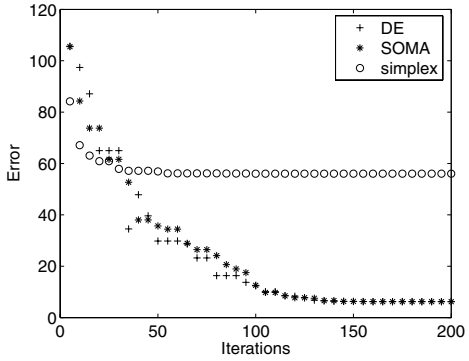
9 Optimisation techniques

There are some techniques however, such as Differential Evolution (DE) [45] and the Self-Organizing Migrating Algorithm (SOMA) [54] that it seems, may be able to cope with the above problems. In image processing, however, the matching error can only be computed numerically and, at best, interpolated from a discrete sampling a fact which may reduce the effectiveness of such algorithms.

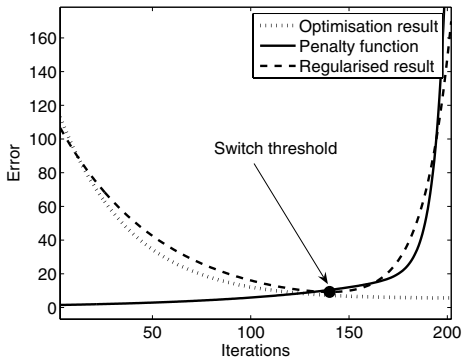
Nevertheless, a preliminary exploration of the two algorithms with a fairly complicated example and comparison of the results with a common local search algorithm within a traditional windowed matching paradigm was promising. For the test, we used the image of Figure 6 (a) and the template in Figure 6 (b). The template is taken from the image (same lighting conditions) and is subject to a random affine geometric transformation so that, in the matching we are seeking to find the optimal 6-parameter affine transform that will match the template with the image.

Figure 6. A typical template matching scenario between a scene (a) and a template that has been randomly affine transformed (b). The significant background clutter makes this example difficult to optimise. Unlike local methods, both DE and SOMA will converge to the optimal solution and transform the template to match the scene (detail)(c).

Comparison of the results obtained from the DE and SOMA algorithms with those obtained from a local, simplex search algorithm [32] can be seen in Figure 7 (a) and the resulting template is overlayed on the original image in



(a)



(b)

Figure 7. (a) The convergence rate of the evolutionary algorithms (DE and SOMA) compared with that of a local algorithm (simplex). It is clear that the simplex algorithm cannot overcome local minima. A hybrid method using a penalty function to determine the switching threshold is shown in (b).

Figure 6 (c). We can see that both DE and SOMA produce similar results and converge to the optimal result, unlike the simplex method which gets stuck in local minima. However, in order to obtain these promising results we had to initialise the populations in the DE and SOMA algorithms fairly close to the final solution. This seems to be necessary because the translations are the most difficult parameters to optimise amongst the 6 dimensions of the affine transform as discussed in Section 8, (see Figure 2) though it may, as noted above, be exacerbated by the fact that the translations are discretised at the pixel resolution. Such good initialisation was only required for the translation space and initial positions may be provided by algorithms which compute global features of the object of interest [31] such as colour, brightness, colour co-occurrence, and texture. Alternatively, one could envisage using algorithms which detect regions that cannot be well explained by a background

model.

In addition we would like to point out that since the convergence speed of both DE and SOMA is initially fast but then becomes quite slow, a better solution in practice would be to adopt a hybrid approach, whereby we begin with DE or SOMA, reduce the error quickly and significantly and then switch over to a local search method when we are near or inside the basin of attraction and the local method is guaranteed to converge fast. The switch threshold may be selected using a penalty function, which penalises high iteration numbers. By adding the penalty function to the optimisation result, we obtain a regularised optimisation result which has a single global minimum at the threshold point. When the global optimiser reaches this point, we can switch over to a local optimisation method on the original objective function. This can be seen in Figure 7 (b).

10 Conclusions

Our main conclusion is that flexible template matching needs to be carried out with care and that, in particular, the whole image, both foreground and background should be modelled. Doing so is necessary in order to be able to make a valid probabilistic interpretation of the matching process, avoids at least some spurious, trivial solutions and leads to a more satisfying approach capable of encompassing both novelty/anomaly detection and expectation driven object recognition and location. In addition, background modelling seems to improve the form of the error surface and to make the basin of attraction of the desired solution less difficult to find. It is argued that, nevertheless, the characteristics of the matching problem are such that the error surface will in general be rugged and of a form that renders many of the common algorithms ineffective or unreliable, in particular for finding the object location. It is suggested that some evolutionary algorithms that have recently proved useful in a variety of engineering and signal processing applications [38, 55] may be appropriate and some preliminary, illustrative results given. Much remains to be done to explore and characterise the form of the error surface more fully, but we hope that doing so will enable us to tailor such algorithms more closely to the problem and thereby develop more effective solutions.

References

[1] S. Basalamah, A. Bharath, and D. McRobbie. Contrast marginalised gradient template matching. *Proceedings of the ECCV 2004: 8th European Conference on Computer Vision*, 3023:417–429, 2004.

[2] L. Bretzner and T. Lindeberg. Use your hands as a 3-d mouse, or, relative orientation from extended sequences of sparse point and line correspondences using the affine trifocal tensor. *Proc. 5th ECCV, LNCS*, 1406:141–157, June 1998.

[3] L. G. Brown. A survey of image registration techniques. *ACM Computing Surveys*, 24(4):325–376, 1992.

[4] B. F. Buxton and M. B. Dias. The principles of view invariant, image-based linear flexible shape modelling. In D. Mukherjee and S. Pal, editors, *Proceedings of the Fifth International Conference on Advances in Pattern Recognition ICAPR-2003*, pages 19–24, Indian Statistical Institute, Kolkata, India, 10-13 December 2003. Allied Publishers.

[5] J. F. Canny. A computational approach to edge detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 8(6):679–698, 1986.

[6] T. F. Cootes and C. J. Taylor. Modelling object appearance using the grey-level surface. in proc. british machine vision conference, 1994, pp.479-488. In *Proc. British Machine Vision Conference*, pages 479–488, 1994.

[7] T. F. Cootes and C. J. Taylor. Statistical models of appearance for computer vision. Technical report, University of Manchester, 2004.

[8] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham. Active shape models - their training and application. *Computer Vision and Image Understanding*, 61:38–59, 1995.

[9] T. F. Cootes, C. J. Taylor, and A. Lanitis. Active shape models: Evaluation of a multi-resolution method for improvingimage search. *Proc. British Machine Vision Conference*, pages 327–336, 1994.

[10] T. F. Cootes, C. J. Taylor, and A. Lanitis. Multi-resolution search with active shape models. *Proc. International Conference on Pattern Recognition*, I:610–612, Odtober 1994.

[11] M. B. Dias and B. F. Buxton. Integrated shape and pose modelling. *Proceedings of the British Machine Vision Conference (BMVC 2002)*, pages 827–836, September 2002.

[12] M. B. Dias and B. F. Buxton. Estimating bas-relief angles, illumination direction and surface texture. *In Proceedings of the 5th International Conference on Advances in Pattern-Recognition*, pages 383–386, December 2003.

[13] M. B. Dias and B. F. Buxton. Separating shape and pose variations. *Image and Vision Computing*, 22(10):851–861, September 2004.

[14] M. B. Dias and B. F. Buxton. Implicit, view invariant, linear flexible shape modelling. *Pattern Recognition Letters*, 26(4):433–447, 2005.

[15] Z. Duric and A. Rosenfeld. Image sequence stabilization in real time. *Real-Time Imaging*, 2(5):271–284, 1996.

[16] A. A. Gohatsby. *2-D and 3-D image registration for medical, remote sensing and industrial applications*. Wiley, 2005.

[17] G. D. Hager and P. N. Belhumeur. Efficient region tracking with parametric models of geometry and illumination. *IEEE PAMI*, 20(10):1025–1039, 1998.

[18] M. E. Hansard and B. F. Buxton. Image-based rendering via the standard graphics pipeline. *IEEE International Conference on Multimedia and Expo*, 3:1437–1440, 2000.

[19] M. E. Hansard and B. F. Buxton. Parametric view-synthesis. *In Proc. 6th ECCV*, 1:191–202, 2000.

[20] D. L. G. Hill, P. G. Batchelor, M. Holden, and D. J. Hawkes. Medical image registration [invited topical review]. *Physics in Medicine and Biology*, 46(3):R1–R45, 2001.

[21] J. Huang and D. Mumford. Statistics of natural images and models. *Computer Vision and Pattern Recognition*, 1:1541–1547, 1999.

[22] A. K. Jain. *Fundamentals of digital image processing*. Prentice Hall, Englewood Cliffs, NJ, 1989.

[23] D. M. Kennedy, B. F. Buxton, and J. H. Gibly. Application of the total least squares procedure to linear view interpolation. *Proc. BMVC*, 1:305–314, 1999.

[24] I. Koufakis and B. F. Buxton. Very low bit-rate face video compression using linear combination of 2dface views and principal components analysis. *Image and Vision Computing*, 17:1031–1051, 1998.

[25] Z. D. Lan, R. Mohr, and P. Remagnino. Robust matching by partial correlation. In D. Pycock, editor, *British Machine Vision Conference*, volume 2, 1995.

[26] A. Lanitis, C. J. Taylor, and T. F. Cootes. An automatic face identification system using flexible appearance models. *Image and Vision Computing*, 13(5):393–402, June 1995.

[27] M. K. Leung and Y. H. Yang. Human body motion segmentation in a complex scene. *Pattern recognition*, 20(1):55–64, 1987.

[28] F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, and P. Suetens. Multimodality image registration by maximization of mutual information. *IEEE Transactions on Medical Imaging*, 16(2):187–198, 1997.

[29] S. McKenna and S. Gong. Non-intrusive person authentication for access control by visual tracking and face recognition. In *Proc. IAPR international conference on audio-video based biometric person authentication*, pages 177–184, 1997.

[30] C. H. Morimoto and R. Chellappa. Automatic digital image stabilization. In *Proc. IEEE conference on pattern recognition*, Vienna, Austria, August 1996.

[31] H. Murase and V. Vinod. Fast visual search using focused color matching - Active Search. *Systems and Computers in Japan*, 31(9):81–88, 2000.

[32] J. A. Nelder and R. Mead. A simplex method for function minimization. *Computer Journal*, 7:308–313, 1965.

[33] T. S. Newman and A. K. Jain. A survey of automated visual inspection. *Computer Vision and Image Understanding*, 61(2):231–262, 1995.

[34] H. T. Nguyen and A. W. M. Smeulders. Fast occluded object tracking by a robust appearance filter. *IEEE Pattern Analysis and Machine Intelligence*, 26(8):1099–1104, August 2004.

[35] D. Pearson. Development in model-based video coding. *Proc. of the IEEE*, 83(6):892–906, June 1995.

[36] G. P. Penney, J. Weese, J. Little, P. Desmedt, D. L. G. Hill, and D. J. Hawkes. A comparison of similarity measures for use in 2d-3d medical image registration. *IEEE Trans. Med. Imag*, 17:586–595, 1998.

[37] W. K. Pratt. *Digital Image Processing*. John Wiley & Sons, New York, 2nd edition, 1991.

[38] A. Rae and S. Parameswaran. Synthesising application-specific heterogenous multiprocessors using differential evolution. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, E84-A(12):3125–3131, 2001.

[39] K. Rohr. Incremental recognition of pedestrians from image sequences. In *Proceedings of the 1993 conference on computer vision and pattern recognition*, pages 8–13, 1993.

[40] D. L. Ruderman and W. Bialek. Statistics of natural images: Scaling in the woods. *Physical Review Letters*, 73(6):814–817, 1994.

[41] F. Segundo. Using the integrated shape and pose model for recognition. Master's thesis, University College London, September 2003.

[42] A. Shashua. *Geometry and Photometry in 3D Visual Recognition*. PhD thesis, Massachusetts Institute of Technology, November 1992.

[43] T. Sim, S. Baker, and M. Bsat. The CMU pose, illumination and expression (PIE) database. In *Proc. of the 5th IEEE international conference on automatic face and gesture recognition*, 2002.

[44] A. Srivastava, A. Lee, E. Simoncelli, and S.-C. Zhu. On advances in statistical modeling of natural images. *Journal of Mathematical Imaging and Vision*, 18:17–33, 2003.

[45] R. Storn and K. V. Price. Differential evolution - a simple and efficient heuristic for global optimization over continuous spaces. *Journal of Global Optimization*, 11(4):341–359, December 1997.

[46] J. Sullivan. *A Bayesian Framework for Object Localisation in Visual Images*. PhD thesis, University of Oxford, September 2000.

[47] J. Sullivan, A. Blake, M. Isard, and J. MacCormick. Object localization by bayesian correlation. *Proc Int. Conf. Computer Vision*, pages 1068–1075, 1999.

[48] T. Thórhallsson and D. W. Murray. The tensors of three affine views. In *Proc. of the Computer Vision and Pattern Recognition conference*, June 1999.

[49] D.-M. Tsai, C.-T. Lin, and J.-F. Chen. The evaluation of normalized cross correlations for defect detection. *Pattern Recognition Letters*, 24(15):2525–2535, November 2003.

[50] S. Ullman and R. Basri. Recognition by linear combinations of models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(10):992–1006, October 1991.

[51] D. Vernon. *Machine Vision: Automated Visual Inspection and Robot Vision*. Prentice Hall, 1991.

[52] P. Viola and W. M. W. III. Alignment by maximization of mutual information. *Proc. 5th Int. Conf. Computer Vision*, pages 16–23, 1995.

[53] J. B. West, J. M. Fitzpatrick, and M. Y. W. et al. Comparison and evaluation of retrospective intermodality image registration techniques. *Journal of Computer Assisted Tomography*, 21(4):554–566, 1997.

[54] I. Zelinka. SOMA-Self Organizing Migrating Algorithm. In G. Onwubolu and B. V. Babu, editors, *New optimization techniques in engineering*. Springer, Berlin, 2004.

[55] I. Zelinka and V. Kresalek. *Fine Mechanics and Optics*, chapter Evolutionary Algorithms and their Applicability in Physics, (Czech ed.). Number ISSN 0447-6441. 2003.

[56] Q. Zhou and J. K. Aggarwal. Tracking and classifying moving objects from video. In *Proc. of the 2nd IEEE internat. workshop on PETS*, Hawaii, December 2001.

[57] V. Zografos and B. F. Buxton. Affine invariant, model-based object recognition using robust metrics and Bayesian statistics. *ICIAR 2005 LNCS*, 2005.