A BAYESIAN APPROACH TO 3D OBJECT RECOGNITION USING LINEAR COMBINATION OF 2D VIEWS

Vasileios Zografos and Bernard F. Buxton
Department of Computer Science, University College London, Malet Place, London, WC1E 6BT, UK

Keywords: Object recognition, linear combination of views, Bayes, Markov-Chain Monte-Carlo.

Abstract: We introduce Bayes priors into a recent pixel-based, linear combination of views object recognition technique. Novel views of an object are synthesized and matched to the target scene image using numerical optimisation. Experiments on a real-image, public database with the use of two different optimisation methods indicate that the priors effectively regularize the error surface and lead to good performance in both cases. Further exploration of the parameter space has been carried out using Markov Chain Monte Carlo sampling.

1 INTRODUCTION

In this work, we examine computational aspects of a *pixel-based* linear combination of views approach to the recognition of objects that vary due to changes in the viewpoint from which they can be seen. This method works directly with a search over pixel values and avoids the need for low-level feature extraction and solution of the correspondence problem. In this paper we illustrate how, by using a Bayesian approach, we can restrict our search to regions where valid and meaningful solutions are likely to exist.

The method works by recovering a set of linear coefficients that will combine a small number of 2-D views of an object and synthesise a novel image which is as similar as possible to a target image of the object. Bayes priors are constructed and shown to regularize optimisation of the synthesized image. For one selected object recognition example, Monte Carlo Markov Chain (MCMC) sampling is used to explore the form of the posterior distribution.

2 PROPOSED METHOD

By using the linear combination of views (LCV) theory (Shashua, 1995; Ullman and Basri, 1991) we can deal with the variations in an object's appearance due to viewpoint changes. Thus, given two different views $I'(x',y')$ and $I''(x'',y'')$ of an object (Fig. 1(a), (b)), we can represent any corresponding point (x,y) in a novel target image I_T as:

$$\begin{array}{l} x = a_0 + a_1x' + a_2y' + a_3x'' + a_4y'' \\ y = b_0 + b_1x' + b_2y' + b_3x'' + b_4y'' \end{array} . \qquad (1)$$

These equations are overcomplete (Ullman and Basri, 1991) and other choices may be made (Koufakis and Buxton, 1998). The novel view may then be synthesised by warping and blending the images I' and I'' as follows:

$$I_T(x,y) = w'I'(x',y') + w''I''(x'',y'') + \varepsilon(x,y). \qquad (2)$$

Only 5 or more corresponding landmark points are necessary in the two views (Fig. 1(a), (b)), and the weights w' and w'' are calculated as described in (Koufakis and Buxton, 1998).

We extend (1) and (2) by incorporating prior information on the coefficients (a_i, b_j), based on previous training with synthetic data, and building a Bayesian model. On the assumption that $\varepsilon(x,y)$ in (2) is i.i.d. random noise drawn from a Gaussian distribution and similarly using Gaussian priors for the LCV coefficients (here considered statistically independent) with means and standard deviations estimated from training data, we get the log posterior as, devoid of any uninteresting constants:

Zografos V. and F. Buxton B. (2008).
A BAYESIAN APPROACH TO 3D OBJECT RECOGNITION USING LINEAR COMBINATION OF 2D VIEWS.
In *Proceedings of the Third International Conference on Computer Vision Theory and Applications*, pages 295-298
DOI: 10.5220/0001081902950298

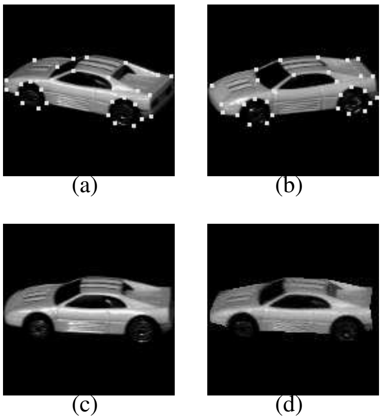


Figure 1: Example of real data from the COIL-20 database. The two basis view images I' (a) and I'' (b) with landmark points selected at prominent features. I_T (c) is the target image. The synthesised image (d) is at the correct pose identified by our algorithm. .

$$-\log[P(\vec{a}_i,\vec{b}_j|I_T,I',I'')] \propto \frac{\sum_{x,y}[I_T(x,y)-I_S(x,y)]^2}{\sigma_\varepsilon^2}+ \sum_{i=0}^{4}\frac{(\vec{a}_i-\mu_{a_i})^2}{\sigma_{a_i}^2}+\sum_{j=0}^{4}\frac{(\vec{b}_j-\mu_{b_j})^2}{\sigma_{b_j}^2}. \tag{3}$$

This defines the probability of observing the target image I_T given the vectors of coefficients $(\vec{a}_i,\vec{b}_j)$ and the basis views I' and I''. We usually require a single synthesised image to be presented as the most probable result. A typical choice is the one which maximises the posterior probability (MAP) or minimises the negative log-posterior (3) with respect to the parameters a_i and b_j. The latter can be minimised using standard optimisation techniques.

Priors were constructed by examination of the variation of the LCV coefficients using a synthetic 3D model. It was found that a_0 follows a quadratic curve, coefficients a_1 and a_3 are linear whilst the remaining coefficients are almost constant. Appropriate Gaussian priors were defined whose effect can be seen in Fig. 2(a). Here we show negative log probability of the likelihood, prior and posterior for the coefficient a_2. The plot was generated by isolating and varying this coefficient while having conditioned the remaining coefficients to the optimal prior values identified previously during training. We note the effect of the prior on the likelihood, especially near the tails of the p.d.f. where we have large error residuals. The prior widens the likelihood's basin of attraction resulting in much easier minimisation, even if we initialise our optimisation algorithm far away from the optimal solution.

On the other hand, near the global optimum we wish the prior to have as little impact as possible in order for the detailed information as to the value of a_2 to come from the likelihood alone. This allows for small deviations from the values for the coefficients encoded in the prior means, since every synthesis and recognition problem differs slightly due to object type, location, orientation and perspective camera effects.

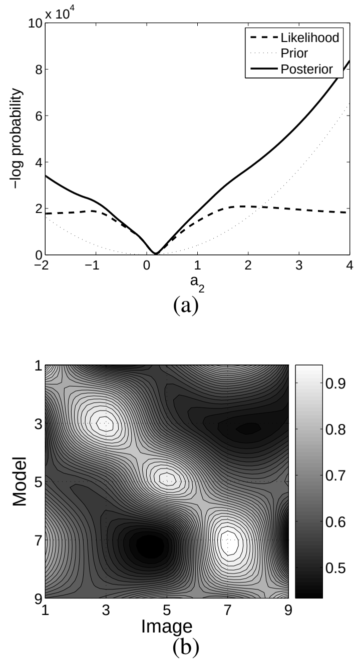
Using the LCV for object recognition is straightforward. The first component of our system is the two stored basis views I' and I'' which define the library of known modelled objects. These are rectangular bitmap images that contain grey-scale (or colour) pixel information of the object without any additional background data. It is important not to choose a very wide angle between the basis views to avoid I' and I'' belonging to different aspects of the object with landmark points being occluded.

Having selected the two basis views, we pick a number of corresponding landmark points, in particular lying on edges and other prominent features. We then use constrained Delaunay triangulation (Shewchuk, 2002) and the correspondence to produce similar triangulations on both the images. The above processes may be carried out during an off-line, model-building stage and are not examined here.

The set of LCV coefficients is then determined by minimising the negative log posterior (3) and the object of interest in the target image I_T recognised by selecting the best of the models, as represented by the basis views, that explain I_T sufficiently well. Essentially, we are proposing a flexible template matching system in which the template is allowed to deform in the LCV space, restricted by the Bayesian priors to regions where there is a high probability of meaningful solutions, until it matches the target image. To do this we need to search a high-dimensional parameter space using an efficient optimisation algorithm. We have tested Differential Evolution (Storn and Price, 1997) and a reformulation of the simplex algorithm (Zografos and Buxton, 2007; Nelder and Mead, 1965).

3 EXPERIMENTS

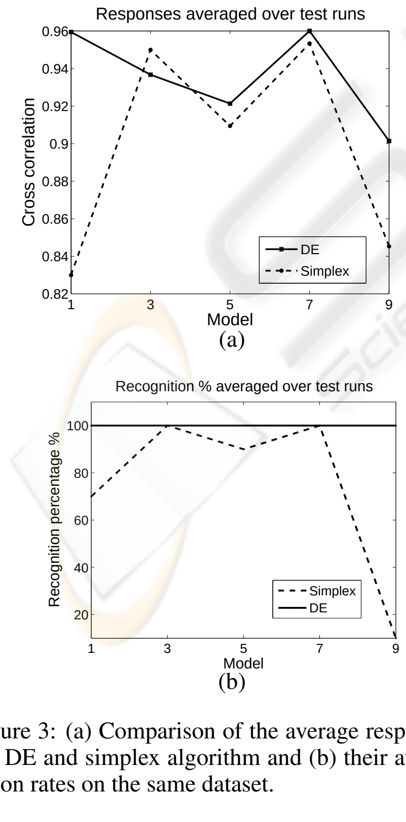
We have performed a number of experiments on real images using the publicly available COIL-20 database (Nene et al., 1996). This database contains examples of 20 objects imaged under varying pose (horizontal rotation around the view-sphere at 5^o intervals) against a constant background with the camera position and lighting conditions kept constant. We constructed LCV models from 5 objects, using as basis views the images at $\pm 20^o$ from a frontal view, while



(a)

(b)

Figure 2: (a) The negative log posterior resulting from the combination of the prior and likelihood. (b) Model × image heatmap array with high cross-correlation in the main diagonal.



(a)

(b)

Figure 3: (a) Comparison of the average response between the DE and simplex algorithm and (b) their average recognition rates on the same dataset.

ensuring that the manually chosen landmarks were visible in both I' and I'' (Fig. 1). Comparisons were carried out against target images from the same set of modelled objects taken in the frontal pose at 0^o.

In total, we carried out 500 experiments (250 with each optimisation method × 10 tests for each model-target image combination) and constructed two 5 × 5 arrays of model×image results. Each array contains information about the matching scores represented by the cross-correlation coefficient. The highest scores were along the main diagonal where each model of an object is correctly matched to a target image of the same object.

For the simplex method, we set the maximum number of function evaluations (NFEs) to 1000 and a fixed initialisation of: a_o,a_1,a_3,b_0,b_1=1, a_2,a_4,b_3=0.5, b_2=0.9, b_4=1.4, deliberately chosen far away from the expected prior solution in order not to influence the optimisation algorithm with a good initialisation. In the case of DE, we chose a much higher NFEs=20000 (100 populations × 200 generations) and specified the boundaries of the LCV space as: -5≤ a_0, b_0 ≤5, -1≤ a_1,a_2,a_3,a_4 ≤1, -1≤ b_1,b_2,b_3,b_4 ≤1.

The results of the above experiments, averaged over 10 test runs, are summarised in the heatmap plot Fig. 2(b). As expected, we can see a well defined diagonal of high cross-correlation where the correct model is matched to the target image. This observation, combined with the absence of any significant outlying good matches when model≠image, leads us to the conclusion that, on average, both methods perform well in terms of recognition results. The question is how close these methods can get to the global optimum, and in how great a NFEs.

We have also included the plots in Fig. 3(a) and (b) which compare the average cross-correlation responses and the recognition rates for both methods respectively. A recognition is deemed a failure if the recovered cross-correlation value is below the 95_{th} percentile of the ground truth solution in each case. This threshold is empirical and some test runs with much lower scores produce visually acceptable matching results.

Both methods have a consistently good performance with the DE converging to solutions of higher cross-correlation in most cases while producing results over 95% of the ground truth in every case. The simplex failed to converge to the correct solution in a few cases, particularly in some of the tests for models 1 and 9, while producing acceptable recognition results in the majority of test runs. This of course may be explained in part by the smaller NFEs that were allowed for this algorithm although preliminary experiments had indicated the NFE value chosen should generally have sufficed.

From these experiments we have also observed that there is little diversity in the 10 coefficients in the recovered solutions along the main diagonal indicating a stability in the coefficients across different objects that is consistent with the prior training data. Also, we have detected a difference in the optimisation behaviour of the two algorithms, DE and simplex, and how much earlier the latter can reach the global minimum. DE is much slower, but it has the advantage that it can avoid locally optimum solutions, which the simplex sometimes cannot.

Finally, in order to obtain a more specific and complete idea of the characteristics of the posterior surface, we have used a Markov-Chain Monte Carlo (MCMC) (Gelman et al., 1995) approach in order to generate a sample of the distribution and further analyse it. We chose a single experiment (matching to a frontal view of object 1 at 0^0) and generated a set of 10000 samples of the posterior (3) from areas of high probability using a single Markov Chain. We then ran a k-means clustering algorithm (Bishop, 1995) which recovered 3 main clusters in close proximity and all near the global optimum. This indicates that, for this example, the distribution is approximately unimodal though perhaps with some subsidiary, nearby peaks caused by noise effects. The main point is that there is no significant local optimum elsewhere nearby in the distribution.

A final examination of the kurtosis and skewness of the sample has shown that the distributions of the samples of all coefficients, except b_1, are quite strongly skewed, reflecting strong influence of the likelihood near the optimum posterior, a property that is highly desirable. This is due to the shape of the likelihood function since the priors are symmetric. The values for the kurtosis are small for some coefficients whose posteriors are therefore almost Gaussian near the optimum, whilst other coefficients strongly affected by the priors are leptokurtic.

4 CONCLUSIONS

Our approach to view-based object recognition involves synthesising intensity images using a linear combination of views and comparing the sythesised images to the target, scene image. We incorporate prior probabilistic information on the LCV parameters by means of a Bayesian model. Matching and recognition experiments carried out on data from the COIL-20 public database have shown that our method works well for pose variations where the target view lies between the basis views. The experiments further show the beneficial effects of the prior distributions in "regularising" the optimisation. In particular, priors could be chosen that produced a good basin of attraction surrounding the desired optimum without unduly biasing the solution.

Nevertheless, additional work is required. In order to avoid the overcompleteness of the LCV equations, we would like to reformulate the LCV equations (1) by using the affine tri-focal tensor and introducing the appropriate constraints in the LCV mapping process. In addition, in this paper we have only addressed extrinsic viewpoint variations, but it should also be possible to include intrinsic, shape variations using the approach described by (Dias and Buxton, 2005).

REFERENCES

Bishop, C. M. (1995). *Neural Networks for Pattern Recognition.* Oxford University Press.

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

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1995). *Bayesian Data Analysis*. Chapman and Hall, London, 2nd edition.

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

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

Nene, S. A., Nayar, S. K., and Murase, H. (1996). Columbia Object Image Library (COIL-20). Technical Report CUCS-006-96, Department of computer science, Columbia University, New York, N.Y. 10027.

Shashua, A. (1995). Algebraic functions for recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(8):779–789.

Shewchuk, J. R. (2002). Delaunay refinement algorithms for triangular mesh generation. *Computational Geometry: Theory and Applications*, 22:21–74.

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

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

Zografos, V. and Buxton, B. F. (2007). Pose-invariant 3d object recognition using linear combination of 2d views and evolutionary optimisation. *ICCTA*, pages 645–649.