School of Informatics, University of Edinburgh

On a Connection between Object Localization with a Generative Template of Features and Pose-space Prediction Methods

by

Christopher K. I. Williams, Moray Allan
School of Informatics, University of Edinburgh
5 Forrest Hill, Edinburgh EH1 2QL, UK
c.k.i.williams@ed.ac.uk, moray.allan@ed.ac.uk

Informatics Research Report 0719

School of Informatics **January 2006**
http://www.informatics.ed.ac.uk/

On a Connection between Object Localization with a Generative Template of Features and Pose-space Prediction Methods

Christopher K. I. Williams, Moray Allan
School of Informatics, University of Edinburgh
5 Forrest Hill, Edinburgh EH1 2QL, UK
c.k.i.williams@ed.ac.uk, moray.allan@ed.ac.uk

Informatics Research Report 0719
SCHOOL *of* INFORMATICS
January 2006

Abstract : We address the task of localizing objects from a given object class in an image. The image is represented as a collection of "visual words" at interest points. The generative template of features (GTF) model defines a distribution over visual words and their spatial locations for each part of the object (Sudderth et al., 2005; Fergus et al., 2005). We show how to derive pose-space prediction methods (such as the Hough transform) from the GTF.

Keywords : Object localization, Generative Template of Features, Hough transform

Over the last few years there has been a surge of interest in the problem of object recognition for object classes (as opposed to recognition of specific objects). We distinguish between object classification (detecting whether there is an object of a given class in an image) and object localization (specifying the location of an object in an image).

For the object localization problem there are three main approaches:

- Scanning methods that run a detector at different positions and scales in the image, and report maxima. Such methods have been used e.g. for handwritten digit recognition (LeCun et al., 1990), face detection (Viola and Jones, 2004) and car detection (Agarwal et al., 2004).
- Pose-clustering methods (see Forsyth and Ponce, 2003, §18.3). This method can be traced back at least to Hough (1962). If the detected features are simple then it can be useful to group features to make tight predictions in pose space. However, with richer features a single feature can be informative, see e.g. the work of Leibe et al. (2004).
- Interpretation-tree algorithms (Grimson, 1990) that match image features to features in a model. An example of such an approach for object class recognition is the work on "constellation models" (Weber et al., 2000; Fergus et al., 2003). Such methods potentially search over all correspondences between image and object features. As this is a combinatoric search it is very slow unless aggressive pruning of the search tree can be achieved.

Recently a generative model for object recognition has been proposed by Sudderth et al. (2005); see also Fergus et al. (2005) for related work. We call this model the generative template of features (GTF), and it is explained further in section 1. The GTF was described in the context of an unsupervised learning problem, but it could equally be trained in a supervised fashion. One way to carry out object localization is to scan this template over an image at various locations and scales. The goal of this paper is to give a principled derivation of pose-space prediction methods from the GTF, which we do in section 2.

1 The Generative Template of Features

We assume that an interest point detector has been run on each image, and that a local image descriptor like Lowe's SIFT descriptor (Lowe, 2004) has been computed at each interest point. These descriptors are clustered over a set of training images to produce a dictionary of "visual words". Thus for image m with N_m interest points we have pairs $(\mathbf{x}_{mi}, w_{mi})$, $i = 1, \ldots N_m$, where $\mathbf{x}_{mi}$ denotes the position and w_{mi} denotes the identity of the visual word of feature i in image m. Let these be collected into the matrices X_m, W_m for image m.

Consider an object which has pose variables $\boldsymbol{\theta}$. $\boldsymbol{\theta}$ could denote the (x, y) position of the object in the image, position plus scale and rotation, or be more complex, e.g. including internal degrees of freedom. Under the model defined in Sudderth et al. (2005) we have

$$p(\boldsymbol{\theta}, X_m, W_m) = p(\boldsymbol{\theta}) \prod_{i=1}^{N_m} p(\mathbf{x}_{mi}, w_{mi} | \boldsymbol{\theta}). \tag{1}$$

Notice that each feature $(\mathbf{x}_{mi}, w_{mi})$ is generated i.i.d., conditional on $\boldsymbol{\theta}$. Of course image features may be generated either from background clutter or from an object of interest. Thus we choose the mixture model

$$p(\mathbf{x}_{mi}, w_{mi}|\boldsymbol{\theta}) = (1-\alpha)p_b(\mathbf{x}_{mi}, w_{mi}) + \alpha p_f(\mathbf{x}_{mi}, w_{mi}|\boldsymbol{\theta}), \tag{2}$$

where p_b denotes a background model, p_f a foreground model (of the object) and α is the probability of choosing to generate from the foreground. The background model will typically generate features anywhere in the image and with a broad distribution of visual word types, with $p_b(\mathbf{x}, w) = p_b(\mathbf{x})p_b(w)$.

The term $p_f(\mathbf{x}_{mi}, w_{mi}|\boldsymbol{\theta})$ can be further decomposed using the notion of parts. If z_{mi} is an indicator variable for each of the P possible parts in the object, then we have

$$p_f(\mathbf{x}_{mi}, w_{mi}|\boldsymbol{\theta}) = \sum_{z_{mi}=1}^{P} p_f(\mathbf{x}_{mi}|z_{mi}, \boldsymbol{\theta})p(w_{mi}|z_{mi})p(z_{mi}). \tag{3}$$

The meaning of eq. 3 is that the location of $\mathbf{x}_{mi}$ depends on the part it is derived from and the object's instantiation parameters, and w_{mi} is generated from a multinomial distribution[1] that depends on only the identity of the part that the visual word is generated from. $p_f(\mathbf{x}_{mi}|z_{mi}, \boldsymbol{\theta})$ could, for example, be a Gaussian distribution. We note that Revow et al. (1996) defined a similar generative model but for black (ink) pixels, without the visual words component.

One can have a deep discussion over the meaning of the term part, but here we will take it to mean that if we have a collection of images of a particular object class (e.g. cars) which have been normalized (e.g. translated, rescaled etc.) so as to be in as good alignment as possible, then there will tend to be regions in this set of images which have propensities to generate particular visual words. For example, the wheels on cars tend to be detected by interest point operators and found towards the bottom left and right hand sides of side views of cars.

Localization of an object with respect to the $\boldsymbol{\theta}$ parameters such as position and scale can be carried out with the GTF by scanning the template over the image at a dense grid of $\boldsymbol{\theta}$ settings, detecting maxima in this space and verifying if these maxima are sufficiently strong to support the hypothesis. Note that the GTF also allows the probabilistic assignment of features to the foreground or background given $\boldsymbol{\theta}$ by computing the posterior probability

$$\frac{\alpha p_f(\mathbf{x}_{mi}, w_{mi}|\boldsymbol{\theta})}{(1-\alpha)p_b(\mathbf{x}_{mi}, w_{mi}) + \alpha p_f(\mathbf{x}_{mi}, w_{mi}|\boldsymbol{\theta})}. \tag{4}$$

If there are a number of object models $O_1, O_2, \ldots$ then to carry out model comparison we seek to compute the marginal likelihood

$$p(X_m, W_m|O_j) = \int p(\boldsymbol{\theta}, X_m, W_m|O_j)d\boldsymbol{\theta}. \tag{5}$$

In general this integral will not be analytically tractable but if $\boldsymbol{\theta}$ is low dimensional it can be approximated, e.g. by using numerical quadrature based on a grid of points in $\boldsymbol{\theta}$-space, or by using Laplace's approximation at a mode of $p(\boldsymbol{\theta}|X_m, W_m, O_j)$.

We note that the i.i.d. assumption in the GTF model is quite weak as it does not enforce generation from particular parts, in contrast to the constellation model.

Dealing with multiple objects in a scene. One way to handle multiple objects in a scene is to follow the treatment of Sudderth et al. (2005). They extend $\boldsymbol{\theta}$ to hold the instantiation

[1] In fact it is not absolutely necessary to cluster the descriptors into a discrete set of visual words; one could define $p(w|z)$ over real-valued descriptors, e.g. with a mixture of Gaussians.

parameters for each object, and define mixing proportions for each object and the background. This approach ignores occlusion, but it would be quite straightforward to use a layered model and to reason about occlusion so as to generate only from visible components. Alternatively, we might expect that individual models could be run to find good regions of $\boldsymbol{\theta}$-space for the given model, and that the robust background model would explain features from other objects. This parallels the work of Williams and Titsias (2004) where such an approach was used to propose good locations for sprite models individually, and a layer ordering was determined in a second pass.

Limitations of the GTF. The GTF does not enforce generation of features from each part, as under the generative model there is a non-zero probability that some part is not chosen on any of the draws from the foreground. This problem was observed by Revow et al. (1996) where it was called the "beads in white space" problem, as ink generators (beads) could occur in regions where there were no black (ink) pixels. One way to deal with this is to use the GTF to find promising regions of $\boldsymbol{\theta}$ space, and then evaluate measures of model fit for such instantiated models which can then be fed to discriminatively-trained classifiers. This strategy was used by Revow et al. (1996) on the digit recognition problem, and more recently has been used, for example, by Fritz et al. (2005) for the recognition of cars, motorbikes, cows and horses.

Learning the GTF. In general learning the GTF requires estimation of both the $p(\mathbf{x}|z)$ and $p(w|z)$ distributions for each part. In Sudderth et al. (2005) and Fergus et al. (2005) training is carried out by using unsupervised learning. However, if we have training data for each object class annotated with bounding boxes (as in the PASCAL VOC dataset[2]) then one can use supervised learning. Each bounding box for a given object class is rescaled so as to be centered and have the same area as the template. (If separate x and y scaling factors are used then rectangular bounding boxes can be brought into perfect alignment.) Given these aligned data it is straightforward to learn the parameters of the template by EM.

2 Making Predictions in Pose-space

As we have said, one way to localize objects in images using the GTF is to scan the template over the image in a dense grid of $\boldsymbol{\theta}$ settings (e.g. position, scale). An alternative approach is to combine predictions from the various features in pose space to approximate $p(\boldsymbol{\theta}|X_m, W_m, O_j)$ for object class O_j. This idea goes back to the Hough transform (Hough, 1962), although note that below we do not bin up $\boldsymbol{\theta}$-space. For example, the classical Hough transform for detecting lines divides parameter space up into bins and each observed point votes for all lines consistent with it; peaks in the accumulator array are candidates for lines.

By taking logs of eq. 1 we obtain

$$\log p(\boldsymbol{\theta}, X_m, W_m) = \log p(\boldsymbol{\theta}) + \sum_{i=1}^{N_m} \log p(\mathbf{x}_{mi}, w_{mi}|\boldsymbol{\theta}). \qquad (6)$$

As the data (X_m, W_m) are fixed we have $p(\boldsymbol{\theta}, X_m, W_m) \propto p(\boldsymbol{\theta}|X_m, W_m)$ with $p(\mathbf{x}_{mi}, w_{mi}|\boldsymbol{\theta})$ viewed as a function of $\boldsymbol{\theta}$. Thus the generative model can be used to hypothesize detections in $\boldsymbol{\theta}$-space by finding the maxima of $p(\boldsymbol{\theta}|X_m, W_m)$, e.g. by hill-climbing. Such an explanation of the probabilistic Hough transform can be found, for example, in Stephens (1991), although

[2] http://www.pascal-network.org/challenges/VOC/.

without the use of specific visual word features (which provide more information and thus tighter distributions).

To spell this out further, consider a distinctive visual word which occurs in only one position on an object. This feature will be predictive of the location of the centre of the object, but as it can also be generated from the background part there is also an associated broad outlier distribution as derived from eq. 3.

Equation 6 shows how to run the generative model backwards to provide predictions in parameter space. However, given training data with features $\{(\mathbf{x}_{mi}, w_{mi})\}$ it is natural to build predictors for $p(\boldsymbol{\theta}|\mathbf{x}_{mi}, w_{mi})$, e.g. by creating a Parzen windows estimator for $p(\boldsymbol{\theta}|\mathbf{x}_{mi}, w_{mi})$. How should we then combine these predictions from each feature in order to obtain $p(\boldsymbol{\theta}|X_m, W_m)$? Fortunately Bayes' rule comes to our aid, as

$$p(\mathbf{x}_{mi}, w_{mi}|\boldsymbol{\theta}) = \frac{p(\boldsymbol{\theta}|\mathbf{x}_{mi}, w_{mi}) p(\mathbf{x}_{mi}|w_{mi}) p(w_{mi})}{p(\boldsymbol{\theta})}. \tag{7}$$

Here $p(\boldsymbol{\theta}|\mathbf{x}_{mi}, w_{mi})$ is obtained from the predictive model, $p(w_{mi})$ is just the marginal probability of visual word w_{mi} over the training set, and $p(\mathbf{x}_{mi}|w_{mi})$ is the probability of seeing a visual word of type w_{mi} in position $\mathbf{x}_i$. This could be estimated, e.g. using a density estimator for the location of features of a given type in the collection of training data. Alternatively, if $p(\boldsymbol{\theta})$ has a non-informative location component then we might expect that $p(\mathbf{x}_{mi}|w_{mi})$ should be uniform across locations in the image. This use of Bayes' theorem to replace likelihood terms with predictive distributions has been called the *scaled likelihood* method, see, e.g. Morgan and Bourlard (1995); Feng et al. (2002).

Putting equations 6 and 7 together we obtain

$$\log p(\boldsymbol{\theta}|X_m, W_m) = \sum_{i=1}^{N_m} \log p(\boldsymbol{\theta}|\mathbf{x}_{mi}, w_{mi}) - (N_m - 1) \log p(\boldsymbol{\theta}) + c, \tag{8}$$

where c is a constant independent of $\boldsymbol{\theta}$. Thus we have shown rigorously how to obtain $p(\boldsymbol{\theta}|X_m, W_m)$ from individual predictions $p(\boldsymbol{\theta}|\mathbf{x}_{mi}, w_{mi})$ up to a normalization constant. Note, however, that to compute the marginal likelihood (eq. 5) from eq. 8 requires additional terms involving $p(\mathbf{x}_{mi}|w_{mi})$ and $p(w_{mi})$ to be included.

Recently, Leibe et al. (2004) have used such ideas to predict an object's location based on the observed position of visual words. However, we note that the equation they use (their eq. 6), is, in our notation,

$$\text{score}_m(\boldsymbol{\theta}) = \sum_{i=1}^{N_m} p_f(\boldsymbol{\theta}|\mathbf{x}_{mi}, w_{mi}). \tag{9}$$

Eq. 9 does not at first sight agree with eq. 1; for a start it sums probabilities rather than multiplying probabilities or summing log probabilities. However, using eq. 2 we have

$$\prod_{i=1}^{N_m} p(\mathbf{x}_{mi}, w_{mi}|\boldsymbol{\theta}) = \prod_{i=1}^{N_m} p_b(\mathbf{x}_{mi}, w_{mi}) \Big[(1-\alpha)^{N_m} + \alpha(1-\alpha)^{N_m - 1} \sum_{i=1}^{N_m} \frac{p_f(\mathbf{x}_{mi}, w_{mi}|\boldsymbol{\theta})}{p_b(\mathbf{x}_{mi}, w_{mi})} + O(\alpha^2) \Big]. \tag{10}$$

If α is small and $p(\boldsymbol{\theta})$ is non-informative w.r.t. location then using eq. 7 for $p_f(\mathbf{x}_{mi}, w_{mi}|\boldsymbol{\theta})$ we obtain to first order

$$p(\boldsymbol{\theta}|X_m, W_m) = c_0 + c_1 \sum_{i=1}^{N_m} \frac{p_f(\mathbf{x}_{mi}, w_{mi})}{p_b(\mathbf{x}_{mi}, w_{mi})} p_f(\boldsymbol{\theta}|\mathbf{x}_{mi}, w_{mi}), \tag{11}$$

where $c_0,\ c_1$ depend on the image features but not on $\boldsymbol{\theta}$ and $p_f(\mathbf{x}_{mi}, w_{mi}) = \int p_f(\mathbf{x}_{mi}, w_{mi}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}$. Minka (2003) has also discussed how a robustified product of probabilities gives rise to a sum of probabilities to first order.

Furthermore, if $p(\boldsymbol{\theta})$ has a non-informative location component then the spatial part of $p_f(\mathbf{x}_{mi}, w_{mi})$ will be non-informative and we can refine eq. 11 to obtain

$$p(\boldsymbol{\theta}|X_m, W_m) = c_0 + c_2 \sum_{i=1}^{N_m} \frac{p_f(w_{mi})}{p_b(w_{mi})} p_f(\boldsymbol{\theta}|\mathbf{x}_{mi}, w_{mi}), \tag{12}$$

where $p_f(w) = \sum_{z=1}^{P} p(w|z)p(z)$, i.e. the weighted average of the multinomial vectors in the foreground parts. Eq. 12 is close to eq. 9, although note the weighting of each predictive distribution $p_f(\boldsymbol{\theta}|\mathbf{x}_{mi}, w_{mi})$ by the factor $p_f(w_{mi})/p_b(w_{mi})$. If visual word w_{mi} is more probable under the background model then its prediction will be discounted. We note that Dorko and Schmid (2005) have discussed selecting discriminative foreground features for use in eq. 9, but that their criterion is based on intuitive arguments rather than on a formal derivation.

In future work we plan to investigate how well the approximation of eq. 12 agrees with eq. 9.

Acknowledgements

MA gratefully acknowledges support through a research studentship from Microsoft Research Ltd. This work was supported in part by the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778. This publication only reflects the authors' views.

References

Agarwal, S., Awan, A., and Roth, D. (2004). Learning to detect objects in images via a sparse, part-based representation. *IEEE PAMI*, 20(11):1475–1490.

Dorko, G. and Schmid, C. (2005). Object Class Recognition using Discriminative Local Features. Technical Report RR-5497, INRIA Rhône Alpes.

Feng, X., Williams, C. K. I., and Felderhof, S. N. (2002). Combining Belief Networks and Neural Networks for Scene Segmentation. *IEEE Trans Pattern Analysis and Machine Intelligence*, 24(4):467–483.

Fergus, R., Fei-Fei, L., Perona, P., and Zisserman, A. (2005). Learning Object Categories from Google's Image Search. In *ICCV 2005*.

Fergus, R., Perona, P., and Zisserman, A. (2003). Object class recognition by unsupervised scale-invariant learning. In *CVPR 2003*.

Forsyth, D. A. and Ponce, J. (2003). *Computer Vision: A Modern Approach*. Prentice Hall, Upper Saddle River, New Jersey.

Fritz, M., Leibe, B., Caputo, B., and Schiele, B. (2005). Integrating Representative and Discriminant Models for Object Category Detection. In *ICCV 2005*.

Grimson, W. E. L. (1990). *Object recognition by computer.* MIT Press, Cambridge, MA.

Hough, P. V. C. (1962). Methods and Means for Recognizing Complex Patterns. U.S. Patent 3069654.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., and Jackel, L. D. (1990). Handwritten digit recognition with a back-propagation network. In *Advances in Neural Information Processing Systems 2*. Morgan Kaufmann.

Leibe, B., Leonardis, A., and Schiele, B. (2004). Combined object categorization and segmentation with an implicit shape model. In *ECCV2004 Workshop on Statistical Learning in Computer Vision*, pages 17–32, Prague, Czech Republic.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110.

Minka, T. (2003). The “summation hack” as an outlier model. Unpublished manuscript available from http://research.microsoft.com/∼minka/papers/.

Morgan, N. and Bourlard, H. A. (1995). Neural Networks for Statistical Recognition of Continuous Speech. *Proceedings of the IEEE*, 83(5):742–770.

Revow, M., Williams, C. K. I., and Hinton, G. E. (1996). Using Generative Models for Handwritten Digit Recognition. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 18(6):592–606.

Stephens, R. S. (1991). Probabilistic approach to the Hough transform. *Image and Vision Computing*, 9(1):66–71.

Sudderth, E., Torralba, A., Freeman, W. T., and Willsky, A. S. (2005). Learning Hierarchical Models of Scenes, Objects and Parts. In *ICCV 2005*.

Viola, P. A. and Jones, M. J. (2004). Robust Real-time Face Detection. *International Journal of Computer Vision*, 57(2):137–154.

Weber, M., Welling, M., and Perona, P. (2000). Unsupervised Learning of Models for Recognition. In *Proceedings of the Fifth European Conference on Computer Vision, ECCV 2000*, pages 18–32.

Williams, C. K. I. and Titsias, M. K. (2004). Greedy Learning of Multiple Objects in Images using Robust Statistics and Factorial Learning. *Neural Computation*, 16(5):1039–1062.