Learning Spatial Context from Tracking using Penalised Likelihoods

Stephen J. McKenna and Hammadi Nait-Charif
Division of Applied Computing, University of Dundee, Dundee DD1 4HN, Scotland
{stephen, hammadi}@computing.dundee.ac.uk

Abstract

MAP estimation of Gaussian mixtures through maximisation of penalised likelihoods was used to learn models of spatial context. This enabled prior beliefs about the scale, orientation and elongation of semantic regions to be encoded, encouraging one-to-one correspondences between mixture components and these regions. In conjunction with minimum description length this enabled automatic learning of inactivity zones and entry zones from track data in a supportive home environment.

1. Introduction

Context-specific spatial models can greatly reduce the complexity of behaviour interpretation. Several authors have proposed learning such models automatically from extended observation but the resulting models are difficult to interpret since they use clusters or hidden states that are not in one-to-one correspondence with semantically meaningful spatial regions (e.g. [6, 7, 11]). This paper demonstrates how MAP estimation can be used to obtain Gaussian mixtures in which such a correspondence is more strongly enforced. The method is applied to a supportive home environment scenario in which a ceiling-mounted camera is used to monitor an occupant. Two uses for the model are fall detection (correlated with unusual inactivity) and high-level activity summarisation in human-readable form. The task is to learn, from motion trajectories, semantically meaningful spatial regions of two kinds: inactivity zones and entry zones. Inactivity zones are regions where the person typically exhibits little global motion for an extended period of time (e.g. a chair, a bed). No distinction was made between entry and exit zones in our context model since these zones are dual purpose: they are referred to as entry zones.

A tracker based on an ellipse model and a particle filter yielded temporally discretised, smoothed 2D trajectories (see top of Figure 2). Points at the beginning and end of a track are entry and exit points respectively. Points at which speed in the image plane drops below a threshold are labelled as inactivity points. This inactivity threshold was set to the speed obtained when a person walked at a slow walking pace at the periphery of the field of view. The problems of learning the entry zones and inactivity zones which constitute the spatial context model were formulated as ones of clustering entry/exit points and inactivity points. These unsupervised learning problems are not straightforward because, although reasonable upper bounds can be imposed, the number of zones of each type is not known *a priori*. In other words, model order must be estimated.

2. Gaussian mixture models

A Gaussian mixture model (GMM) is a probability density function (PDF) of the form $p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k p(\mathbf{x}|k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ where $\sum_{k=1}^{K} \pi_k = 1$ and the mixture components are Gaussian densities. The model's parameters, $\boldsymbol{\theta}$, are the mixing weights, π_k, the means, $\boldsymbol{\mu}_k$, and the covariance matrices, $\boldsymbol{\Sigma}_k$, for each Gaussian component $k \in 1 \ldots K$. Given a set $\mathcal{X} = \{\mathbf{x}^1, \ldots, \mathbf{x}^N\}$ of N i.i.d. realisations of $\mathbf{x}$, the log likelihood is:

$$L(\mathcal{X}|\boldsymbol{\theta}) = \log \prod_{n=1}^{N} p(\mathbf{x}^n|\boldsymbol{\theta}) = \sum_{n=1}^{N} \log \sum_{k=1}^{K} \pi_k p(\mathbf{x}^n|k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{1}$$

The EM algorithm [1] provides an iterative method for searching for a local maximum of this likelihood. Each iteration consists of an E-step and an M-step. In the E-step the posterior probability that component k is responsible for $\mathbf{x}^n$ is estimated:

$$h_k^n = \frac{\pi_k p(\mathbf{x}^n|k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{i=1}^{K} \pi_i p(\mathbf{x}^n|i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)}$$

In the M-step, the parameters are re-estimated as:

$$\pi_k^{new} = \frac{1}{N} \sum_{n=1}^{N} h_k^n \qquad \boldsymbol{\mu}_k^{new} = \frac{\sum_{n=1}^{N} h_k^n \mathbf{x}^n}{\sum_{n=1}^{N} h_k^n} \tag{2}$$

$$\boldsymbol{\Sigma}_k^{new} = \frac{\sum_{n=1}^{N} h_k^n (\mathbf{x}^n - \boldsymbol{\mu}_k^{new})(\mathbf{x}^n - \boldsymbol{\mu}_k^{new})^T}{\sum_{n=1}^{N} h_k^n} \tag{3}$$

This maximum likelihood (ML) estimation algorithm, although sensitive to initial conditions, can provide effective parameter estimation. As is well known, however, ML cannot be used to determine the number of Gaussians, i.e. the model order. Roberts *et al.* [10] compared six model order selection techniques for GMMs and found that those methods with some information theoretic basis outperformed more heuristic methods. In particular, a method based on the minimum description length (MDL) principle [9] was strong. This principle can be concisely stated as *select the model that gives the shortest description of the data set*. MDL has been used to select GMM model order for clustering human gestures [13] and space-time regions for video indexing [4]. Given parameter estimates, $\hat{\boldsymbol{\theta}}$, the model order is selected so as to minimise the description length, $\mathcal{C}$, in Eqn. (4) where M is the number of free parameters in the model.

$$\mathcal{C} = -L(\mathcal{X}|\hat{\boldsymbol{\theta}}) + \frac{1}{2}M \ln N \tag{4}$$

This is in fact a simplified, *two-stage* description length criterion [5]. The first term represents the number of nats needed to encode the data set, $\mathcal{X}$, given the estimated model, $\hat{\boldsymbol{\theta}}$. The second term represents the number of nats needed to encode the model parameters, $\hat{\boldsymbol{\theta}}$, to precision $1/\sqrt{N}$, which is the magnitude of the parameter estimation error. Note that $\ln 2$ nats $\equiv 1$ bit.

3. Maximum penalised likelihood

The EM algorithm in conjunction with the MDL criterion can be used to estimate the order and parameters of a GMM PDF. This has been found to work well on several synthetic and real-world data sets in the literature although some authors report a tendency for MDL to underestimate model order. In this paper, GMMs are used to identify semantic regions for spatial context modelling. Here, the aim is not an accurate overall density estimation. Rather, the model order should correspond to the number of semantic regions, and the Gaussian parameters should provide a probabilistic description of the spatial characteristics of these regions. Data from a region might be distributed only approximately normally.

In order to obtain Gaussian components that correspond to meaningful semantic regions, a penalised likelihood approach is adopted. A penalty term is added to the log-likelihood function such that maximising this penalised likelihood is equivalent to the Bayesian approach of maximising the posterior (i.e. MAP estimation) where the penalty term is the log of the prior.

Gauvain and Lee [3] proposed MAP estimation of GMMs for speech data using a product of a Dirichlet density and normal-Wishart densities as a prior joint density, $p(\boldsymbol{\theta})$. This choice of prior was justified by the fact that the Dirichlet density is a conjugate density for the multinomial distribution (for the mixing weight parameters) and the normal-Wishart density is a conjugate density for the Gaussian distribution. It assumes independence between the parameters of each Gaussian component and the mixing weights. This choice of prior enables EM to be applied to MAP estimation, i.e. to maximise the penalised likelihood:

$$L(\mathcal{X}|\boldsymbol{\theta}) + \log \mathcal{D}(\boldsymbol{\pi}|\gamma) + \sum_{k=1}^{K}[\log \mathcal{N}(\boldsymbol{\mu}_k|\boldsymbol{\nu}_k, \eta_k^{-1}\boldsymbol{\Sigma}_k) + \log \mathcal{W}(\boldsymbol{\Sigma}_k^{-1}|\alpha_k, \boldsymbol{\beta}_k)] \tag{5}$$

where $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$, $\mathcal{D}$ is a Dirichlet density, $\mathcal{N}$ is a normal density and $\mathcal{W}$ is a Wishart density. The hyperparameters (α_k, $\boldsymbol{\beta}_k$, γ_k, η_k and $\boldsymbol{\nu}_k$) can be interpreted as sufficient statistics of an additional, notional data set. In fact, notional data sets of different sizes, ω_π, ω_μ and ω_Σ can be associated with each of the different model parameters. If a non-informative, uniform prior on the hyperparameters is assumed, then $\alpha = \omega_\Sigma + d$ and $\boldsymbol{\beta} = \omega_\Sigma \mathbf{S}$, where $\mathbf{S}$ is the estimate of $\boldsymbol{\Sigma}$ obtained from the notional data set. The E-step is as before but the M-step is modified [3, 8]. In particular, in the absence of prior knowledge about the means and mixing parameters (i.e. $\omega_\pi = \omega_\mu = 0$), the updates for these parameters remain unchanged while the covariance update becomes:

$$\boldsymbol{\Sigma}_k^{new} = \frac{\sum_{n=1}^{N} h_k^n (\mathbf{x}^n - \boldsymbol{\mu}_k^{new})(\mathbf{x}^n - \boldsymbol{\mu}_k^{new})^T + \boldsymbol{\beta}_k}{\sum_{n=1}^{N} h_k^n + \alpha_k - d} \tag{6}$$

4. Learning inactivity zones

The approach adopted when learning inactivity zones is that, *a priori*, there is no reason to prefer any image location over any other, nor to bias the mixing weights. There is, however, a strong prior belief about inactivity zones' scale and shape. In particular, the distribution characterising a zone is expected to be approximately isotropic. The penalised likelihood method is therefore used to penalise non-isotropic Gaussians that differ from the expected scale. These beliefs are encoded by setting $\omega_\pi = \omega_\mu = 0$ and $\mathbf{S} = \sigma^2\mathbf{I}$ where σ is a scale parameter and $\mathbf{I}$ is the identity matrix. The EM algorithm needed then uses the original M-steps for the mixing parameters and means (Eqn. (2)) and a covariance update based on Eqn. (6). The values of ω_Σ and σ need to be determined in advance. The σ parameter encodes a prior belief about spatial scale (the variation in image translation of a person when at rest in an inactivity zone) while ω_Σ encodes the strength of this prior belief. These values do not have to be chosen very accurately because the results obtained are similar over a large range of values.

5. Learning entry zones

Entry zones are elongated and expected to occur near the image borders in the application considered here. Two solutions are described for learning them. The first models entry zones as 1D distributions on a closed contour near the image borders. The second models entry zones as elongated 2D distributions.

Rather than treat entry zones as 2D regions, they can be treated as 1D regions on some closed contour, $\mathcal{B}$, specified to be near the image borders where entry zones will be located. The problem is then that of clustering entry/exit points after projecting them onto a closed contour. (Either each 2D point is mapped to the nearest point on the contour or the points at which trajectories cross $\mathcal{B}$ for the first and last time are recorded). One approach would be to treat these points as circular data and estimate a mixture of von Mises distributions [12]. However, the data are not truly circular and so a simpler approach was preferred here that takes advantage of the fact that every room will have a relatively large distance between at least two neighbouring entry zones. A point on $\mathcal{B}$ was found in a region with a low density of entry-exit points. This point was used to 'break' $\mathcal{B}$ so as to treat the data as linear. A 1D Gaussian mixture clustering method similar to the one used to identify inactivity zones was then used to identify entry zones. The scale parameter, σ, was set to reflect a prior belief about the width of room entrances (doors). The point at which to break the contour was found using the following simple algorithm. Points on $\mathcal{B}$ were ordered to give a set $\{x_1, \ldots, x_N\}$ of points on the 1D contour relative to an arbitrary origin on $\mathcal{B}$. The break point on $\mathcal{B}$ was then found as $(x_{j+\delta} - x_j)/2$ where $j = \arg\max_j |x_{j+\delta} - x_j|$ and arithmetic was performed modulo N. The offset δ was set to a small fraction of the data set size to give some robustness to outliers. In experiments described here, $\delta = \lceil 0.01N \rceil$. However, the breakpoint found was rather insensitive to the value of δ.

Alternatively, the spatial extent of an entry zone can be modelled as an elongated 2D elliptical region with an appropriate orientation angle, ϕ. In the special case of an entry zone which is elongated along the image's x-axis (i.e. $\phi = 0°$), a diagonal covariance matrix $\mathbf{C} = diag[\sigma_x^2, \sigma_y^2]$ characterises the zone, where $\sigma_x > \sigma_y$. The determinant $|\mathbf{C}|$ encodes the spatial scale and the ratio σ_x/σ_y encodes the elongation. However, the orientation, ϕ, of an entry zone is expected to change with image location in the application considered here. Assuming that the image coordinates are relative to an origin in the centre of the image, a Gaussian centred at $\boldsymbol{\mu} = (\boldsymbol{\mu}_x, \boldsymbol{\mu}_y)$ is expected to be oriented with an angle which can be approximated as $\phi = \tan^{-1}(\frac{w\boldsymbol{\mu}_y}{h\boldsymbol{\mu}_x})$ where w and h are the width and height of the image. The corresponding covariance matrix can then be obtained as $\mathbf{R}_\phi \mathbf{C} \mathbf{R}_\phi^T$ which is a transformation of $\mathbf{C}$ such that the corresponding ellipse is rotated by ϕ where $\mathbf{R}$ is a rotation matrix. This suggests a modification to the M-step for updating the covariance matrices by setting $\mathbf{S}$ in Eqn. (6) to $\mathbf{R}_\phi \mathbf{C} \mathbf{R}_\phi^T$. In this way, the current estimate of a Gaussian component's mean is used to determine ϕ. The prior for a Gaussian's covariance matrix thus depends on its mean.

6. Experiments

Evaluation was performed on trajectory data obtained in a supportive home environment scenario. EM algorithms were initialised by running K-means and setting mixing weights to the proportion of data points in each cluster and covariance matrices to the sample covariances for each cluster. Figure 1 shows the description lengths obtained using Eqn. (4) from 10 different runs of EM for each value of K between 1 and 9. Plotted are means obtained over ten runs for each model order. Error bars denote $\pm$ one standard deviation. ML estimation of inactivity zones resulted in a minimum at $K = 6$ indicating that a mixture of this many components best estimated the density. However, MAP estimation resulted in a minimum at $K = 2$, the true number of semantic regions. ML and MAP estimation of 1D entry zones both resulted in a minimum at $K = 2$ which is the true number of semantic regions (doors). Note, however, that there is increased certainty about the MAP model order due to the reduced variance. ML estimation of 2D entry zones resulted in a minimum at $K = 5$. However, estimation using the penalised likelihood resulted in a correct minimum at $K = 2$. Figure 2 shows example results obtained using the model orders suggested by MDL for ML (left) and MAP (right). Image resolution was 480×360 pixels.

It should be noted that prior parameters ($\omega_\Sigma = 0.2N$, $\sigma = 40$, $\sigma_x = 40$ and $\sigma_y = 20$ pixels) were deliberately not set carefully: the spatial scale parameters chosen were in fact rather too large for the scene used here. Sensitivity to the value of ω_Σ was investigated by examining the proportion of the mixing weights accounted for by the strongest two Gaussian components when clustering inactivity points using $K = 6$. This proportion was greater than 0.99 for $0.05 < \omega_\Sigma < 10$, indicating that the result was rather insensitive over this large range of values.

7. Conclusions

In summary, the use of the penalised likelihoods resulted in MDL estimates that recovered the true semantic regions. On the other hand, unpenalised ML estimation with MDL resulted in the number of Gaussians being overestimated. Furthermore, modifications to the simplified MDL to more accurately estimate description length (e.g. [2]) are likely to further increase the model order estimated with ML.

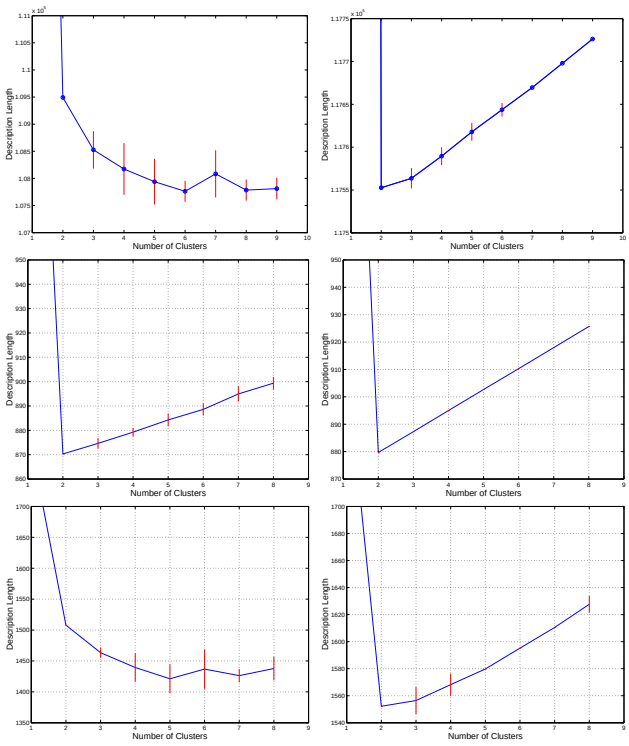

Figure 1. Description lengths for inactivity zones (top), 1D entry zones (middle) and 2D entry zones (bottom). Left: ML. Right: MAP.

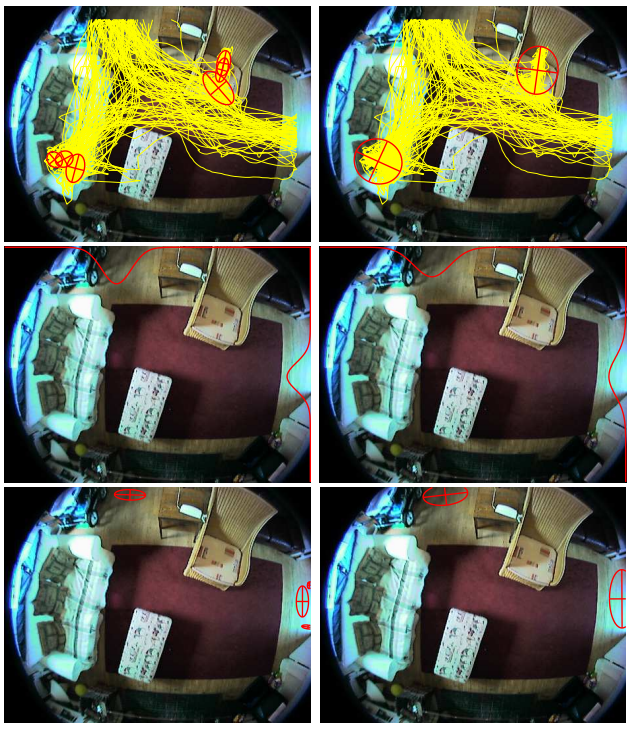

Figure 2. Example results for inactivity zones (top), 1D entry zones (middle) and 2D entry zones (bottom). Left: ML. Right: MAP.

Future work could usefully explore learning temporal context with this approach. The method could also be extended to cope with outliers by, for example, assigning one Gaussian in the mixture a low mixing weight and large variance priors.

References

[1] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. Royal Statistical Society B*, 39:1–38, 1977.

[2] M. A. T. Figueiredo, J. M. N. Leitao, and A. K. Jain. On fitting mixture models. In E. Hancock and M. Pellilo, editors, *Energy Minimization Methods in Computer Vision and Pattern Recognition*, pages 54–69. Springer-Verlag, 1999.

[3] J.-L. Gauvain and C.-H. Lee. Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. *IEEE Trans. Speech and Audio Proc.*, 1994.

[4] H. Greenspan, J. Goldberger, and A. Mayer. A probabilistic framework for spatio-temporal video representation and indexing. In *European Conference on Computer Vision*, Copenhagen, 2002.

[5] M. H. Hansen and B. Yu. Model selection and the principle of minimum description length. *Journal of the American Statistical Association*, 96(454):746–774, 2001.

[6] N. Johnson and D. Hogg. Learning the distribution of object trajectories for event recognition. *Image and Vision Computing*, 14(8):609–615, August 1996.

[7] D. Makris and T. Ellis. Automatic learning of an activity-based semantic scene model. In *IEEE Conf. Advanced Video and Signal Based Surveillance*, Miami, July 2003.

[8] D. Ormoneit and V. Tresp. Averaging, maximum penalized likelihood and Bayesian estimation for improving Gaussian mixture probability density estimates. *IEEE Transactions on Neural Networks*, 9(4):639–650, 1998.

[9] J. Rissanen. Modelling by shortest data description. *Automatica*, 14:465–471, 1978.

[10] S. Roberts, D. Husmeier, I. Rezek, and W. Penny. Bayesian approaches to Gaussian mixture modelling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1133–1142, 1998.

[11] C. Stauffer and W. E. L. Grimson. Learning patterns of activity using real-time tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):747–757, 2000.

[12] C. S. Wallace and D. L. Dowe. MML clustering of multi-state, Poisson, von Mises circular and Gaussian distributions. *Statistics and Computing*, 10(1):73–83, January 2000.

[13] M. Walter, A. Psarrou, and S. Gong. Data driven gesture model acquisition using minimum description length. In *British Machine Vision Conference*, Manchester, UK, 2001.