

Summarising Contextual Activity and Detecting Unusual Inactivity in a Supportive Home Environment

Stephen J. McKenna and Hammadi Nait-Charif

Division of Applied Computing,

University of Dundee, Dundee DD1 4HN, UK.

e-mail: stephen@computing.dundee.ac.uk

Running head: Summarising Activity and Detecting Unusual Inactivity

Summarising Contextual Activity and Detecting Unusual Inactivity in a Supportive Home Environment

Abstract. Interpretation of human activity and the detection of associated events are eased if appropriate models of context are available. A method is presented for automatically learning a context-specific spatial model in terms of semantic regions, specifically inactivity zones and entry zones. Maximum a posteriori estimation of Gaussian mixtures is used in conjunction with minimum description length for selection of the number of mixture components. Learning is performed using EM algorithms to maximise penalised likelihood functions that incorporate prior knowledge of the size and shape of the semantic regions. This encourages a one-to-one correspondence between the Gaussian mixture components and the regions. The resulting contextual model enables human-readable summaries of activity to be produced and unusual inactivity to be detected. Results are presented using overhead camera sequences tracked using a particle filter. The method is developed and described within the context of supportive home environments which have as their aim the extension of independent, quality living for older people.

Keywords: Learning spatial context, inactivity detection, fall detection, supportive home environments, Gaussian mixture models, expectation-maximisation.

Originality and Contribution

The paper contributes a novel method for learning models of spatial context from tracking data. This method involves the use of MAP estimation of Gaussian mixtures in combination with minimum description length to determine both the model order and the parameters. The use of appropriate priors encourages a one-to-one mapping between semantic spatial regions and Gaussian components. Appropriate methods for determining these priors are given for the application under consideration. Importantly, the resulting models can be used to provide event detection and meaningful, human-readable summaries of activity.

The application considered is the use of computer vision for monitoring an older person living alone. Supportive home environments are an important application domain with potential for substantial economic, social and health benefits. The use of overhead vision using standard cameras is novel in this application. The use of the learned context models to provide summarisation and unusual inactivity detection (and thus a cue for fall detection) are demonstrated.

The empirical evaluation provides a comparison with the widely used maximum likelihood method. It also demonstrates activity summarisation and unusual inactivity detection on a relatively large video data set.

1. Introduction

In recent years, researchers have developed vision systems that monitor and interpret human activity in a scene based on models learned through extended observation. Monitoring and surveillance applications need models of context in order to provide semantically meaningful summarisation and recognition of activities and events. Manual specification of such models is clearly limiting so several methods have been proposed for learning context models automatically. This paper describes the use of Bayesian Gaussian mixture estimation combined with minimum description length model order selection to learn models of spatial context automatically from tracking data. Semantic regions (zones) of interest are represented in these context models as Gaussian mixture components. Priors are used to encode beliefs about the scale and shape of these zones. This helps to ensure a one-to-one correspondence between the mixture components and the semantic regions. The resulting models can be used to facilitate automated summarisation of activity in a human-readable format and at an appropriate level of abstraction. This is demonstrated here using an experimental design in which an actor is verbally instructed to perform activities by intuitive reference to semantic regions and the automatically reconstructed summary is then compared to the original instructions. The context models can also be used to detect unusual inactivity.

This paper focuses on an indoor application in which ceiling-mounted visual sensors are used for monitoring in a supportive home environment for older people living independently. Possible aims of such systems include detection of important, rare events such as falls and monitoring of activity patterns. Before describing the learning algorithms and experimental results, the following sections give a description of the application and context-specific spatial modelling in general. Section 4 describes an illustrative scenario and the wide-angle video data sets used in the experiments. The tracking method used to obtain spatio-temporal trajectories is outlined for completeness in Section 5. Section 6 describes the methods used to learn spatial context models from the track data. In particular, two types of spatial zone of importance in this application are considered: inactivity zones and entry zones. After an experimental evaluation of these methods, their use in automatic summarisation and detection of unusual inactivity is presented. The use of the latter as a cue for fall detection is demonstrated. Finally, conclusions are drawn in Section 9.

2. Supportive environments for independent living

2.1. Sensors for telecare in the home

There is evidence that home environments able to monitor the activities of their occupants automatically can be used to help extend independent, quality living and reduce healthcare costs [1, 2, 3, 4]. BT and Anchor Trust performed preliminary research into lifestyle monitoring in the home, recording average activity profiles against which current activity could be compared [1, 4]. Other projects such as the Aware Home [5] are investigating ubiquitous sensing in home environments. Unintentional falls in older people impose a substantial burden on health and social services [6]. A ‘long lie’ after a fall may be as relevant to decreasing

the chances of survival as a broken bone [7]. Clearly, the speed with which emergency help is alerted to a fall is crucial to the faller's health, especially when that person is older. Fear of falling is also a major problem, especially for those who have fallen in the recent past.

Monitoring may utilise sensors embedded in the home environment, worn by the older person, or some combination of both. Sensors used in current supportive home environments often have relatively narrow functionality. For example, passive infra-red (PIR) sensors, pressure pads and fridge door sensors enable room occupancy, presence in a particular area or use of a fridge to be monitored respectively [1, 2, 8]. Passive fall detectors worn on the hip are currently available commercially (see Doughty [7] for a description of this technology). However, they are often not worn when returning home, during housekeeping tasks prone to cause false alarms, or when uncomfortable [9]. Embedded sensors, in contrast, have the advantage of ensuring compliance within the home.

2.2. Vision-based monitoring

There are important issues of acceptability and privacy surrounding the use of computer vision for monitoring in the home. These are being explored along with user requirements and design constraints using a novel drama-based methodology described elsewhere [10]. In particular, it must be decided at what level of abstraction to interpret the behaviour of the older person and how to use the resulting data. It is reasonable to envisage a number of interpretive aims for such vision systems. These range in complexity from monitoring room occupancy, to automatically detecting falls and performing detailed analyses of activity patterns. Reduced mobility can be predictive of a fall and has other health implications. Inactivity detection can, in a context-dependent way, indirectly indicate ill-health or a fall. Activity patterns such as repetition as well as significant changes in daily or weekly patterns might also be detected.

The resulting information could be used as part of an alarm system, potentially detailing the nature of the alarm event and providing evidence for this information. Perhaps equally importantly, it could be used for prediction and thus prevention of falls through risk assessment. Retrospective analysis of logged data in human-readable, summary form could also be useful to provide insights into behaviour and health for use by carers, social workers and researchers.

2.3. Ceiling-mounted visual sensors

The monitoring set-up investigated here uses ceiling-mounted, standard cameras with vertically-oriented optical axes, fitted with wide-angle lenses. Older peoples' homes are often highly cluttered with belongings and furniture brought from former, larger homes. The position and orientation of the cameras have thus been chosen to minimise occlusion of the person by furniture. Wall-mounted cameras would clearly result in greater levels of occlusion.

Causes of difficulty include the varied appearance of the occupant (changes of clothing and body posture), varying sources of illumination (both indoor and outdoor) and cast shadows. Standard, as opposed to infra-

red, cameras have been employed. The main advantage of this is the reduced costs; infra-red sensing remains relatively expensive, especially at high resolution. Some preliminary work has been performed using low-resolution infra-red sensing for fall detection [11, 12]. The authors are not aware of any other published work using computer vision systems with overhead cameras for this application.

3. Context-specific spatial models

Patterns of inactivity could be used to make inferences about health and also to help detect falls. It is important to note, however, that the significance of inactivity changes with context. A person lying on a sofa, as she often does, is probably only resting. In contrast, a person lying on the floor, where she has not previously lain, may have fallen and require assistance. The method presented here enables inactivity outside the usual inactivity zones to be detected. When combined with information about body pose and motion this should provide a useful cue for fall detection. In addition, a semantic, human-readable description of behaviour in terms of spatial regions such as normal zones of inactivity (e.g. chairs, beds) provides a useful summary of behaviour. In order to achieve the goals of unusual inactivity detection and behaviour summarisation, a model of spatial context is required.

3.1. Related work

Context-specific spatial models can be used to greatly reduce the complexity of behaviour interpretation [13]. Howarth and Buxton [14] described a spatial model suitable for stylised domains such as road traffic environments. Ayers and Shah [15] monitored activity in an office environment using manually specified prior knowledge of scene layout (e.g. locations of entrances, exits and objects of interest). Nguyen *et al.* [16] recognised high-level behaviours of people in a complex indoor environment from their trajectories using models learned from labelled training data with specified semantic landmarks. The spatial contextual models in these cases can be specified using interactive graphical tools but manual generation of such models is clearly limiting. Manual labelling of training data for each spatial context is even more time consuming. Fernyhough *et al.* [17] showed how contextual spatial models could be automatically generated for strongly stylised domains. Human behaviour is in general, however, far less stylised.

Several authors have used tracking over extended periods of time to learn patterns of human activity in a scene from extracted motion data and subsequently to detect unusual activities in that scene. Johnson and Hogg learned a model of the distribution of trajectories from data supplied by an active shape model tracker [18]. Trajectories were represented in terms of position and instantaneous velocity and a neural network implementing vector quantisation approximated the distribution using prototype vectors. Stauffer and Grimson clustered patterns of activity by overfitting with a large number of Gaussians [19]. While these methods enabled unusual activity to be detected, the clusters and prototypes found did not always have any clear semantic interpretation. This can make automatic production of activity annotations in a human-

readable form problematic. Makris and Ellis [20, 21] described methods for learning a semantic scene model from 2D trajectories in terms of semantic entities such as entry zones, exit zones, routes, path segments and junctions. They used Expectation-Maximisation (EM) to estimate maximum likelihood Gaussian mixture models of entry and exit zones [21], running EM with a large number of Gaussians (ten) and subsequently pruning Gaussians with observation density below a prespecified threshold. The results did not enforce a one-to-one correspondence between Gaussians and subjective semantic entry-exit zones. In related work, the same authors constructed hidden Markov models based on models of the distributions of trajectories across each route [22]. The focus of these authors was on observing the activity of many people in wide-area, outdoor surveillance scenes. This is in contrast to the application considered here in which the activity of a single person is monitored at close range.

Hidden Markov models (HMMs) and related dynamic probabilistic models have been widely used to model and recognise human actions and gestures. Typically, the hidden states are not readily interpretable and cannot be used to obtain a human-readable summary. Brand and Kettner [23] used an entropy minimisation technique to obtain HMMs with hidden states organised to correspond more closely to meaningful activities. After some manual grouping and labelling of states, this was used, for example, to summarise office activity in a human-readable form and to detect unusual activity.

3.2. Spatial context in the home

Within a room in a home, there will typically be a few places in which an occupant spends most of his or her time while in that room. A living room, for example, contains chairs and sofas and the occupant might even have a favourite seat in which he or she invariably sits to watch television, read or sleep. Such places will be referred to here as *inactivity zones* to indicate that occupancy of such a zone tends to involve little global motion of the person. A room will have a fixed set of entrances which also serve as exits. A place in which entry and exit occurs will be referred to as an *entry zone*. These entry zones could, in some cases, be specified manually during system set-up, albeit at extra installation expense. However, a physical doorway used to enter a room might not be entirely within a camera’s field of view. Furthermore, a given person might not ever use a doorway’s entire width. Therefore, it is useful to learn the entry zones automatically.

Typical use of a room involves entering followed by visits to one or more *inactivity zones* and finally exiting the room. Of course other sustained activities may occur but these tend to be more highly variable and transient. It is proposed here that a useful, compact, semantic representation of behaviour in this context can be provided by temporal segmentation of sensor data into time spent (i) entering via *entry zones*, (ii) inactive in *inactivity zones*, (iii) transitioning between *zones*, and (iv) exiting via *entry zones*. Learning a context-specific spatial model then consists of automatically identifying and characterising these zones.

4. Experimental scenario

Figure 1 shows a scene used to illustrate the method. The strong perspective effects due to the wide-angle lens are apparent. An actor was instructed to perform a series of activities in the room designed to emulate aspects of the way an older person might use such a room. The instructions were given only in terms of five atomic instructions relating to four semantic regions and to falling over (F). The four regions were the two entrances to the room (H and R), a chair with a telephone beside it (C), and a sofa (S). Example instructions were “enter through the hall door, sit on the sofa and then exit through the rear door” (HSR) or “enter through the hall door, sit and use the telephone, sit on the sofa and then exit through the hall door” (HCSH). The first three columns in Table 1 summarise the image sequences that were acquired. There were 16 classes of sequences, classified according to the instructions given to the actor. Four classes contained sequences in which the actor was instructed to simulate a fall. (There are obvious barriers to acquiring video data of real falls). The number of examples in each of the first 12 classes was chosen to reflect the frequency with which that class of activity might occur during day-to-day usage of the room by an older person. The relative frequencies were in fact estimated based on subjective knowledge of the person who lives in the house containing this room. In addition, falls were collected to enable evaluation of unusual inactivity. The sequences were acquired using a digital video camera with an IEEE-1394 interface at a frame-rate of $30Hz$. The image resolution was 480×360 pixels. Ninety-seven sequences were acquired, totalling 46755 image frames (26 minutes) over two days of changeable weather. The scene contained multiple light sources (windows and indoor lighting) and no attempt was made to control the extent of lighting changes and cast shadows.

5. Overhead tracking

Although not the main focus of this paper, the tracking method used to obtain the motion trajectories is outlined here for completeness. Vision systems for indoor human tracking and monitoring typically use cameras with relatively narrow fields of view, mounted at a significant angle to the ceiling, e.g. in shopping malls [24], railway stations [25, 26], sports halls [27] and office environments [15, 16]. Tracking has been performed with cameras mounted with near vertical optical axes in environments with very high ceilings (e.g. [28, 29]) or in people counting applications which require only a restricted view (e.g. [30]). Alternatively, omnidirectional cameras have been used (e.g. [31]). Krumm *et al.* described a system consisting of multiple stereo cameras to monitor activity in a living room [32]. The approach proposed in this paper uses wide-angle monocular cameras, mounted with vertically-oriented optical axes. This avoids most other-object occlusions and means that a single camera is often sufficient to monitor an entire room in a normal home.

In a home environment, the lighting and layout are rather poorly constrained. The clothing and body postures that will be encountered are highly variable and it cannot be assumed that the articulated structure of the body will be apparent. Furthermore, it is likely to be very difficult to construct detailed statistical shape models that capture the range of variation in such a way that enables the unusual poses associated

with events such as falls to be tracked and detected. Instead, the person’s position in the image along with a coarse representation of the shape and orientation in the image were tracked using an ellipse so that the state at time t was $\mathbf{e}_t = (x_t, y_t, \psi_t, s_t, e_t)$ where (x_t, y_t) is the ellipse centre and the other parameters are orientation, scale and eccentricity respectively. The authors believe that this representation of a person will be rich enough to support recognition of relevant actions and events such as falling, lying down, sitting and standing. It is also flexible enough to enable a very wide range of body poses and clothing to be tracked. Although the tracking method can be extended to track multiple people, the application considered here only requires a single person to be tracked.

Trajectories were extracted and represented directly in the image plane. The use of a ground-plane constraint was inappropriate because the distance from the person’s ‘centre’ to the floor was large relative to the camera distance and varied significantly with body pose. The camera was uncalibrated and the person was tracked without performing image rectification.

Several authors have tracked objects and people using either ellipses or Gaussian ‘blobs’ which have elliptical isoprobability contours in image space (e.g. [33, 34, 35, 36, 37]). Image measurements made when tracking with a contour model assume a reasonably accurate 2D shape model and that image features such as edges will therefore lie close to the model contour [33, 38, 36]. In the case of overhead person tracking, the body shape is highly deformable and is in fact poorly modelled by an ellipse. Modelling the object using a spatial Gaussian distribution can be effective even when the object is not elliptical but it is not robust to clutter in the image. If the model is fitted to a noisy background subtraction, for example, cast shadows can lead to highly inaccurate results.

The tracker used here employs a particle filtering method with image evidence provided using an adaptive background model with shadow detection [39]. Hypothesised ellipses were scored using a function that provided some robustness to noisy background cues (e.g. due to shadows) and highly non-elliptical poses (such as outstretched arms). Specifically, pixels exterior to the ellipse but close to the ellipse contour were considered to constitute an adjacent annular region (see Figure 2). The score was computed such that an ellipse hypothesis was penalised for having pixels in this adjacent region that were likely to be foreground and for having pixels in the ellipse interior that were likely to be background.

Particle filters have become popular for tracking in computer vision since they were applied by Isard and Blake [40]. The nonparametric representation of the state density enables tracking through ambiguous situations and is important when dealing with strong shadows and highly cluttered scenes. Tracking was performed here using a particle filtering method called Iterated Likelihood Weighting (ILW) [41]. Empirical comparisons of person tracking using standard Condensation [40], auxiliary particle filtering [42] and ILW using multiple runs of each method demonstrated that ILW can result in reduced variance and more reliable trajectory estimation [41].

Figure 3 illustrates the need for a high-level tracker using a method such as particle filtering. Shown are examples of the low-level adaptive background subtraction cue temporarily yielding a poor result due

to strong shadowing. Low-level region extraction methods would fail to give a reasonable localisation of the person given such data.

Figure 4 shows typical estimates obtained during tracking. The tracker lost lock in 3 of the 96 sequences. In each of these 3 cases, lock was lost near the image border shortly after initialisation due to strong shadow and background clutter, leading to premature termination.

The tracker provided a trajectory in the 5D ellipse parameter space for the tracked person in each sequence. These trajectories were temporally smoothed using a moving average filter. The person’s speed in the image plane was estimated using finite differences at each point on each smoothed trajectory. The smoothed ellipse centre trajectories and speeds were subsequently used to provide a compact representation of the person’s global motion. The remaining parameters provided information about pose but these were not used further in the experiments described here. Figure 6 shows smoothed ellipse centre trajectories overlaid on the scene.

6. Learning spatial context

The tracker yields temporally discretised, smoothed 2D trajectories. 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, τ_s , are labelled as inactivity points. This inactivity threshold was set to $\tau_s = 25$ pixels per second. It was set using an estimate of the minimum image speed of a person walking slowly near the image boundary in a sequence. If this inactivity threshold was set too large, erroneous inactivity points would be detected. Such outliers can cause problems for the estimation of inactivity zones. If the threshold was too small, on the other hand, a reduced set of inactivity points would be detected in the inactivity zones. This has relatively little effect on the learning of inactivity zones since inactivity points are plentiful. The threshold can therefore be set conservatively low. 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. 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, the model order must be estimated.

Gaussian mixture models (GMMs) are a popular clustering approach for reasons which include their analytic properties and the fact that many data sets contain clusters that are approximately Gaussian. Of course, GMMs cannot be expected to discover meaningful structure if data clusters are highly non-Gaussian (see e.g. [43]). Adopting a Gaussian mixture model, the goal here is to obtain a mixture with Gaussian components that correspond directly to entry and inactivity zones. Penalised likelihood functions that encode priors on cluster scale and shape will be used to estimate model order and obtain semantically meaningful Gaussians. Before describing these algorithms, the widely used EM algorithm for maximum likelihood (ML) estimation of GMM parameters and methods for model order selection are briefly reviewed.

6.1. Gaussian mixture models and maximum likelihood estimation

A Gaussian mixture model is a probability density function 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 of dimensionality d :

$$p(\mathbf{x}|k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = \frac{1}{(2\pi)^{\frac{d}{2}} |\boldsymbol{\Sigma}_k|^{\frac{1}{2}}} \exp\left[-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\mathbf{x} - \boldsymbol{\mu}_k)\right]$$

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 $k \in 1 \dots K$. Given a set $\mathcal{X} = \{\mathbf{x}^1, \dots, \mathbf{x}^N\}$ of N i.i.d. realisations of \mathbf{x} , the log likelihood is:

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

The EM algorithm [44] 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:

$$\pi_k^{new} = \frac{1}{N} \sum_{n=1}^N h_k^n \quad (2)$$

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

$$\boldsymbol{\Sigma}_k^{new} = \frac{\sum_{n=1}^N h_k^n (\mathbf{x}^n - \boldsymbol{\mu}_k^{new})(\mathbf{x}^n - \boldsymbol{\mu}_k^{new})^T}{\sum_{n=1}^N h_k^n} \quad (4)$$

This maximum likelihood estimation algorithm, although sensitive to initial conditions, can provide an effective method for parameter estimation. As is well known, however, maximum likelihood cannot be used to determine the number of Gaussians in the mixture. (For example, allocating one Gaussian at each data point, the likelihood can be made to tend to infinity by shrinking the variance parameters.)

6.2. Model Order Selection

The problem of determining the number of Gaussians is one of model order selection. Roberts *et al.* [45] compared six model order selection techniques for GMMs and found that those methods with some information theoretic basis outperformed other more heuristic methods. In particular, a method based on the minimum description length (MDL) principle [46] was comparatively 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 computing layered representations of multiple motions [47], clustering human gestures [48], and clustering spatial or spatio-temporal regions for indexing and retrieval [49, 50]. Given parameter estimates, $\hat{\boldsymbol{\theta}}$, the model order is selected so as to minimise the description length, \mathcal{C} , in Equation (5) where

M is the number of free parameters in the model. In the case of a GMM with full covariance matrices, $M = \frac{1}{2}Kd^2 + \frac{3}{2}Kd + K - 1$.

$$\mathcal{C} = -L(\mathcal{X}|\hat{\boldsymbol{\theta}}) + \frac{1}{2}M \ln N \quad (5)$$

This is in fact a simplified, *two-stage* description length criterion [51]. The first term represents the number of nats¹ 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 $1/\sqrt{N}$. This is motivated by the fact that the magnitude of the parameter estimation error is $1/\sqrt{N}$, so the parameters need only be coded with this much precision. The criterion in (5) is valid for any $1/\sqrt{N}$ -consistent estimator (including the maximum likelihood estimator). It should be noted that the Bayesian Information Criterion of Schwarz [52], derived as an approximation to the posterior distribution on model order, also consists of choosing the model with the smallest value of \mathcal{C} in (5).

6.3. Maximum penalised likelihood

The EM algorithm in conjunction with the MDL criterion can be used to estimate the order and parameters of a GMM probability density function. 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 the 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. This biases estimation so that the Gaussian components in the mixture become more likely to correspond to the semantic regions.

Gauvain and Lee [53] proposed MAP estimation of GMMs in the context of hidden Markov models of speech. They used a product of a Dirichlet density and normal-Wishart densities as a prior joint density, $p(\boldsymbol{\theta})$:

$$p(\boldsymbol{\theta}) \propto \mathcal{D}(\pi_1, \dots, \pi_K | \gamma) \prod_{k=1}^K \mathcal{N}(\boldsymbol{\mu}_k | \boldsymbol{\nu}_k, \eta_k^{-1} \boldsymbol{\Sigma}_k) \mathcal{W}(\boldsymbol{\Sigma}_k^{-1} | \alpha_k, \boldsymbol{\beta}_k) \quad (6)$$

where \mathcal{D} is a Dirichlet density, \mathcal{N} is a normal density and \mathcal{W} is a Wishart density². This choice of prior

¹ 1 bit $\equiv \ln 2$ nats

² The Wishart distribution has the form:

$$p(\boldsymbol{\Sigma}^{-1} | \alpha, \boldsymbol{\beta}) = c |\boldsymbol{\Sigma}^{-1}|^{(\alpha-d-1)/2} \exp[-\frac{1}{2} \text{tr}(\boldsymbol{\beta} \boldsymbol{\Sigma}^{-1})]$$

where c is a normalisation factor and α is the degrees of freedom

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 (for the means and covariance matrices)³. It assumes independence between the parameters of each Gaussian component and the mixing weights. This choice of prior enables the EM algorithm to be applied to MAP estimation, i.e. to maximise the following penalised likelihood:

$$L_{pen}(\mathcal{X}|\boldsymbol{\theta}) = L(\mathcal{X}|\boldsymbol{\theta}) + \log \mathcal{D}(\pi_1, \dots, \pi_K | \gamma) + \sum_{k=1}^K [\log \mathcal{N}(\boldsymbol{\mu}_k | \boldsymbol{\nu}_k, \eta_k^{-1} \boldsymbol{\Sigma}_k) + \log \mathcal{W}(\boldsymbol{\Sigma}_k^{-1} | \alpha_k, \boldsymbol{\beta}_k)] \quad (7)$$

The EM algorithm to maximise $L_{pen}(\mathcal{X}|\boldsymbol{\theta})$ has the same E-step as before but the M-steps are modified as follows [53, 54].

$$\pi_k^{new} = \frac{\sum_{n=1}^N h_k^n + \gamma_k - 1}{N + \sum_{k=1}^K \gamma_k - K} \quad (8)$$

$$\boldsymbol{\mu}_k^{new} = \frac{\sum_{n=1}^N h_k^n \mathbf{x}^n + \eta_k \boldsymbol{\nu}_k}{\sum_{n=1}^N h_k^n + \eta_k} \quad (9)$$

$$\boldsymbol{\Sigma}_k^{new} = \frac{\sum_{n=1}^N h_k^n (\mathbf{x}^n - \boldsymbol{\mu}_k^{new})(\mathbf{x}^n - \boldsymbol{\mu}_k^{new})^T + \eta_k (\boldsymbol{\mu}_k^{new} - \boldsymbol{\nu}_k)(\boldsymbol{\mu}_k^{new} - \boldsymbol{\nu}_k)^T + \boldsymbol{\beta}_k}{\sum_{n=1}^N h_k^n + \alpha_k - d} \quad (10)$$

The hyper-parameters (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 of size ω . Consider a notional data set $\mathcal{X}^* = \{\mathbf{x}^{N+1}, \dots, \mathbf{x}^{N+K\omega}\}$ of size $K\omega$ generated by a mixture of K Gaussians. Given a prior belief that each Gaussian is equally likely to have generated each data point, assume that each Gaussian generated ω of these data points. Let \mathcal{X}_k^* denote the data subset generated by the k^{th} Gaussian. In order to use a prior on the covariance matrix parameters of the GMM, imagine using \mathcal{X}^* to estimate these parameters. If a non-informative, uniform hyperprior is assumed (as in [54]), then having observed the data \mathcal{X}^* , the parameters are distributed as follows.

$$\boldsymbol{\Sigma}_k^{-1} \sim \mathcal{W}(\omega + d, \sum_{\mathbf{x} \in \mathcal{X}_k^*} (\mathbf{x} - \boldsymbol{\nu}_k)(\mathbf{x} - \boldsymbol{\nu}_k)^T) \quad (11)$$

Therefore, the hyperparameters (the parameters of the Wishart distribution) can be expressed as $\alpha = \omega + d$ and $\boldsymbol{\beta} = \omega \mathbf{S}_k$, where \mathbf{S}_k is the estimate of $\boldsymbol{\Sigma}_k$ obtained from \mathcal{X}_k^* . Other hyperparameters are set to neutral values, $\gamma_k = 1$ and $\eta_k = 0$, so that the M-steps in Equations (2) and (3) are recovered for the means and mixing weights. Other work on the use of priors in this and related models can be found elsewhere [55, 56]. In what follows, this penalised likelihood method is used in conjunction with the MDL criterion to obtain models of spatial context. Ormoneit and Tresp [54] used the above EM algorithm as a regularisation method to help avoid overfitting when performing density estimation using GMMs. In this paper the aim is somewhat different. Penalised likelihood estimation is used along with MDL to obtain mixture models whose Gaussian components correspond to semantically meaningful regions.

³ Analysis is simplified if the prior is chosen such that the posterior has the same functional form as the prior. The prior and posterior are then said to be *conjugate*.

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. 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. Even when inactive, a person is never entirely motionless. For example, he/she might make small adjustments of body posture for comfort, turn the pages of a book or operate a TV remote control. Such local movements will lead to motion of the ellipse centroid. We expect that over many visits to an inactivity zone, variation in the placement of the body in the zone and the local body movements made will not result in the ellipse centre position having a highly non-isotropic distribution. This leads to the subjective prior that inactivity zones are 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 $\mathbf{S}_k = \mathbf{S} = \sigma^2 \mathbf{I}$ 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 (Equations (2) and (3)). The covariance update in the M-step is:

$$\Sigma_k^{new} = \frac{\sum_{n=1}^N h_k^n (\mathbf{x}^n - \boldsymbol{\mu}_k^{new})(\mathbf{x}^n - \boldsymbol{\mu}_k^{new})^T + \omega \mathbf{S}}{\sum_{n=1}^N h_k^n + \omega} \quad (12)$$

The values of ω and σ need to be determined in advance. The σ parameter encodes a prior belief about spatial scale while ω 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.

6.5. Learning entry zones

The methods used to learn entry zones also adopt penalised likelihoods but they are formulated somewhat differently. Entry zones are elongated rather than isotropic and they are expected to occur near the image borders in the application considered here. Two solutions are now described. 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.

6.5.1. Entry zones on a 1D closed contour

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 in the range $[0, 2\pi)$. The von Mises distribution is the circular analog of the Gaussian distribution on a line so a mixture of von Mises distributions could be estimated [57]. 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 (on an open contour). A Gaussian mixture clustering method similar to the one used to identify inactivity zones was then used to identify entry zones. The only difference was that the mixture was 1D rather than 2D. The scale parameter, σ , was set to reflect a prior belief about the width of room entrances (doors).

A point at which to break the closed contour in order to obtain an open contour must be selected. If the closed contour were broken at a point within an entry zone, this would lead to more than one Gaussian component being assigned to that zone. This is because the data arising from the zone would thus be split with some near one end of the resulting open contour and the rest of the data near the other end. A point at which to break the contour safely was found using the following simple algorithm. Points on \mathcal{B} were ordered to give a set $\{x_1, \dots, x_N\}$ 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 the experiments described here, $\delta = \lceil 0.01N \rceil$. However, the breakpoint found was rather insensitive to the value of δ . (In fact, a value of $\delta = 0$ gave satisfactory results here because the tracking data obtained were relatively clean.)

6.5.2. Entry zones as elongated 2D regions

Alternatively, the spatial extent of an entry zone (within which tracks begin and end) can be modelled as an elongated 2D elliptical region with an appropriate orientation angle, ϕ . Prior beliefs about the scale, elongation and orientation of these ellipses can be characterised by specifying priors for the covariance matrices. In the special case of an entry zone which is elongated along the image's x -axis (i.e. $\phi = 0^\circ$), a diagonal covariance matrix $\mathbf{C} = \text{diag}[\sigma_x^2, \sigma_y^2]$ 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 (the ratio of the axes of an ellipse). 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

$$\mathbf{R}_\phi = \begin{bmatrix} \sin(\phi) & \cos(\phi) \\ -\cos(\phi) & \sin(\phi) \end{bmatrix} \quad (13)$$

This suggests the following modification to the M-step for updating the covariance matrices.

$$\boldsymbol{\Sigma}_k^{new} = \frac{\sum_{n=1}^N h_k^n (\mathbf{x}^n - \boldsymbol{\mu}_k^{new})(\mathbf{x}^n - \boldsymbol{\mu}_k^{new})^T + \omega \mathbf{R}_\phi \mathbf{C} \mathbf{R}_\phi^T}{\sum_{n=1}^N h_k^n + \omega} \quad (14)$$

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.

7. Empirical evaluation of context learning

The above GMM methods were evaluated and compared on the trajectory data obtained from the scenario described in Section 4. The sequences in the data set were allocated at random to training and test sets such that half of the examples in each class were reserved for testing. The training and test sets were subsequently swapped and the results on the two test sets combined. Entry/exit and inactivity points were obtained from the training data and used to learn spatial context models. In each case, the EM algorithms were initialised by running K-means clustering and then setting the mixing weights to the proportion of data points in each cluster and the covariance matrices to the sample covariances for each cluster. It should be noted that, in what follows, the prior parameters were deliberately not set carefully: the scale parameters chosen were in fact rather too large for the scene used here. They were $\sigma = 40$, $\sigma_x = 40$ and $\sigma_y = 20$ unless otherwise stated.

7.1. Inactivity zones

When learning inactivity zones, the value of σ represented subjective, prior knowledge about the variation in image translation of a person when at rest in an inactivity zone. Figure 5 shows the description lengths obtained using Equation (5) from 10 different runs for each value of K between 1 and 9. Maximum likelihood estimation resulted in a minimum at $K = 6$ indicating that a mixture of this many components best estimated the density. However, MAP estimation using the penalised likelihood resulted in a minimum at $K = 2$, the true number of semantic regions (inactivity zones). Figure 6 shows example results obtained using ML with $K = 6$ and MAP with $K = 2$. The MAP-MDL estimates resulted in correct models of the inactivity zones.

An experiment was conducted in order to investigate the sensitivity of the results to the values of the notional prior sample size, ω , and scale, σ . Figure 7 plots the proportion of the mixing weights accounted for by the strongest two Gaussian components when clustering the inactivity points using $K = 6$. The leftmost plot shows that this proportion was greater than 0.99 for $0.05 < \omega/N < 10$, indicating that the result was rather insensitive over this large range of values. The abscissa shows the relative strength of the prior (i.e. ω/N) on a logarithmic scale. As $\omega \rightarrow 0$, the unpenalised maximum likelihood result is recovered and the probability mass is distributed across more than two Gaussians. As $\omega \rightarrow \infty$, the image data have less influence on the covariance matrix estimates and the (inaccurate) prior leads again to more than two Gaussians having significant mixing weights. In all other experiments reported in this paper, $\omega = 0.2N$. The rightmost plot shows that the proportion of the mixing weights accounted for by the two strongest Gaussians was greater than 0.99 over a large range of values of σ , indicating that this subjective scale prior did not need to be very accurate in order for the correct model order (number of Gaussian mixture components) to be found.

7.2. Entry zones

Figure 8 plots the description lengths obtained using Equation (5) from 10 runs for each value of K between 1 and 9 when entry zones were learned as 2D regions. Maximum likelihood estimation resulted in a minimum at $K = 5$. However, estimation using the penalised likelihood resulted in a minimum at $K = 2$ which is the true number of semantic regions (entry zones). Figure 9 shows example results obtained using ML with $K = 5$ and using the penalised likelihood with $K = 2$.

Figure 10 plots the description lengths obtained using Equation (5) from 10 runs for each value of K between 1 and 9 when learning entry zones on a 1D contour. In this case, maximum likelihood estimation resulted in a minimum at $K = 2$ which is the true number of semantic regions (entry zones). The penalised likelihood also resulted in a minimum at $K = 2$. Figure 12 shows example results obtained using ML with $K = 2$ and using the penalised likelihood with $K = 2$. In this case, the 1D mixture distributions are shown overlaid on the scene. Although unpenalised ML resulted in the correct minimum, it should be noted that there was increased certainty when using the penalised likelihood due to the reduced variance. In fact, the variance over runs was effectively zero. In all runs with $K \geq 2$, only two Gaussian components had significantly non-zero mixing weights. Furthermore, these two Gaussians' means and covariance matrices did not vary over runs, resulting in essentially the same mixture model being obtained in each case. This is what gave rise to the linear increase on the righthand plot in Figure 10 which is due only to the description length penalty (i.e. the second term in Equation 5). Figure 11 shows the penalised likelihood result obtained when the scale parameter was halved ($\sigma = 20$) and illustrates that variance is reduced compared to ML even when the scale prior is quite inaccurate.

8. Temporal segmentation and recognition of unusual inactivity

The learned model of spatial context can be used to temporally segment trajectories and to detect unusual inactivity. The Gaussian components in the mixtures correspond to inactivity zones and entry zones. Each Gaussian thus provides a model for the spatial extent of a zone. Entry zones can be used to focus tracker initialisation and to semantically label points of entry and exit. When a person's speed drops to an extent that indicates inactivity, the inactivity zone PDFs provide a way of checking whether the inactivity is occurring in an unusual location or whether the person is resting in a known inactivity zone. A simple algorithm was used to decide whether the person was inactive in a known inactivity zone. The speed, s_t , at each time-step was estimated using a finite difference over a 40-frame temporal window. Whenever $s_t < \tau_s$, the Mahalanobis distance, $D_k = ((\mathbf{x}_t - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}_t - \boldsymbol{\mu}_k))^{\frac{1}{2}}$, of the ellipse centre position from each inactivity zone was evaluated. The person was recognised as being inactive in the k^{th} inactivity zone if $k = \operatorname{argmin}_{j \in \{1, \dots, K\}} D_j$ and $D_k < \tau_D$. Here results are reported for a detection threshold of $\tau_D = 2$. (Similar results were obtained using $\tau_D = 3$.)

The regions in the automatically learned spatial model corresponded well (and one-to-one) with the

semantic zones referred to in the natural language instructions given to the actor. The annotation results are summarised in Table 2. An overall error rate was computed as:

$$E = 100 \times \frac{E_{sub} + E_{ins} + E_{del}}{N_{test}}$$

where $E_{sub} = 1$, $E_{ins} = 11$ and $E_{del} = 2$ were the number of atomic instructions erroneously substituted, inserted and deleted respectively and $N_{test} = 270$ was the total number in the test set. The error rate was therefore $E = 5.2\%$. Insertion errors occurred, for example, due to the person leaning forward on the sofa and the algorithm therefore labelling him as temporarily leaving and then reentering the S inactivity zone. Deletion and substitution errors occurred due to the tracker losing lock.

In summary, the automatic summarisation of the sequences obtained compared well with the instructions originally given to the actor. In other words, correct, intuitive, linguistic descriptions of activity were generated automatically. Figure 13 shows some example trajectories with the current ellipse overlaid. The trajectories here are colour-coded to indicate the temporal segmentation obtained. In each of these cases, the trajectory was correctly segmented into transitions between zones, inactivity within a known inactivity zone and inactivity while not in a known inactivity zone.

9. Discussion and Conclusions

MAP estimation of Gaussian mixture models through EM-based 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 thus encouraging a one-to-one correspondence between the mixture components and these regions. In conjunction with minimum description length model order selection this enabled automatic learning of inactivity zones and entry zones from tracking data. This method was demonstrated in a supportive home environment where it can be used to provide human-readable summarisation of activity and detection of unusual inactivity. High-level activity summarisation and context-dependent inactivity detection are also important in other applications. The former provides an efficient coding for storage and retrieval. The latter is useful in monitoring and surveillance.

Others have reported that MDL used with maximum likelihood estimation tends to underestimate model order. Modifications to the simplified MDL to more accurately estimate description length are likely to increase the model order estimated (e.g. [58]). However, the task addressed in this paper was not that of obtaining an accurate mixture density estimation but that of obtaining a semantically meaningful clustering. Due to non-Gaussian clusters and noise, maximum likelihood density estimation with MDL actually *overestimated* in this respect. When MDL was used in conjunction with penalised likelihood estimation, on the other hand, correct semantic clusterings were obtained.

The data sets used in the experiments reported here did not contain a significant number of gross statistical outliers. However, the method could be extended to cope with data sets that did have significant numbers of outlying data points. This could be done, for example, by assigning one Gaussian large, fixed variance parameters. The mixture model can then use this Gaussian to explain those outlying points which

cannot be easily accounted for using other Gaussians in the model whose priors are set to reflect prior beliefs about the semantic regions of interest.

In the acted sequences used in the experiment reported here, the duration of inactivity in the activity zones was typically only a few seconds. In reality, much longer periods of inactivity would be normal and this would make the learning easier since the inactivity data would be both more plentiful and more concentrated in the inactivity zones. Future work could usefully explore learning and inference using temporal context from the track data. An obvious extension is to use the methods described here to learn hidden Markov models with Gaussian or Gaussian mixture observation densities. Another extension would be to perform learning and inference using the further ellipse parameters in addition to the centre trajectories. In particular, ellipse eccentricity is likely to be useful for the detection of falls.

A method has been described for learning aspects of the activity patterns of a single person when alone. It was assumed that there was never more than one person in the environment. In situations when other people were present in the home environment, it might be reasonable to suspend learning and recognition because (a) the behaviour patterns of the person being modelled are likely to alter and (b) the need for an automatic fall alarm system would be greatly reduced. Other moving objects such as pets are likely to be distinguishable from the person being modelled. However, such situations were not addressed explicitly in this paper. Several extensions have been proposed to particle filter tracking to cope with multiple objects [59, 60, 61, 62] and investigation of their applicability to this application is left for future work.

Passive fall detection has been identified as a priority for supportive home environments for older people. The methods presented here have gone some way to providing useful cues for fall detection. In future work it would be interesting to combine these cues (unusual inactivity) with dynamic models of falling. It is conceivable that the 5D spatio-temporal trajectories extracted here in terms of ellipse centre, scale, elongation and orientation contain sufficient information to enable many falls to be detected. Exactly how these parameters can be used for this purpose and to what extent the camera set-up needs to be calibrated in order to achieve this goal are interesting open questions. Visual environmental factors such as lighting levels and changes of room layout can be significant for many older people with poor vision and are implicated in falls. The automatic detection of such changes is also under investigation.

Acknowledgments

Dr. Nait Charif was supported by UK EPSRC EQUAL grant GR/R27419/01. The authors are grateful to the reviewers for helpful comments on an earlier version of this paper.

References

- [1] N. M. Barnes, N. H. Edwards, D. A. D. Rose, and P. Garner. Lifestyle monitoring: technology for supported independence. *IEE Computing and Control Engineering Journal*, pages 169–174, August 1998.
- [2] S. Bonner. Assisted interactive dwelling house: Edinvar housing association smart technology demonstrator and

- evaluation site. In I. P. Porrero and E. Ballabio, editors, *Improving the Quality of Life for the European Citizen (TIDE: Technology for Inclusive Design and Equality)*, pages 396–400, 1998.
- [3] F. Marquis-Faulkes, S. J. McKenna, P. Gregor, and A. F. Newell. Gathering the requirements for a fall monitor using drama and video with older people. *Technology and Disability*, 2003. In Press.
- [4] J. Porteus and S. Brownsell. Using telecare: Exploring technologies for independent living for older people. Technical report, Report on the Anchor Trust/BT Telecare Research Project, Anchor Trust, 2000.
- [5] C. D. Kidd, R. Orr, G. D. Abowd, C. G. Atkeson, I. A. Essa, B. MacIntyre, E. D. Mynatt, T. Starner, and W. Newstetter. The aware home: A living laboratory for ubiquitous computing research. In *Cooperative Buildings*, pages 191–198, 1999.
- [6] P. Scuffham, S. Chaplin, and R. Legood. Incidence and costs of unintentional falls in older people in the United Kingdom. *Journal of Epidemiology and Community Health*, 57:740–744, 2003.
- [7] K. Doughty. Fall prevention and management strategies based on intelligent detection, monitoring and assessment. In *New Technologies in Medicine for the Elderly*, Charing Cross Hospital, London, November 2000.
- [8] M. Chan, H. Bocquet, E. Campo, and J. Pous. Remote monitoring system to measure indoors mobility and transfer of the elderly. In I. P. Porrero and E. Ballabio, editors, *Improving the Quality of Life for the European Citizen (TIDE: Technology for Inclusive Design and Equality)*, pages 379–383, 1998.
- [9] SeniorWatch. Fall detector. Technical report, A Case Study of the European IST Seniorwatch Project, IST-1999-29086, www.seniorwatch.de, 2001.
- [10] S. J. McKenna, F. Marquis-Faulkes, A. F. Newell, and P. Gregor. Using drama for requirements gathering: a case study on advanced sensors in supportive environments for the elderly. *International Journal of Human-Computer Studies*, 2004. Submitted.
- [11] P. A. Bromiley, P. Courtney, and N. A. Thacker. Design of a visual system for detecting natural events by the use of an independent visual estimate: A human fall detector. In H. I. Christensen and P. J. Philips, editors, *Empirical Evaluation Methods in Computer Vision*, volume 50 of *Machine Perception and Artificial Intelligence*. World Scientific Publishing, 2002.
- [12] A. Sixsmith and N. Johnson. SIMBAD: Smart inactivity monitor using array-based detector. *Gerontechnology*, 2(1):110–110, 2002.
- [13] H. Buxton and S. Gong. Visual surveillance in a dynamic and uncertain world. *Artificial Intelligence*, 78(1–2):431–459, 1995.
- [14] R. J. Howarth and H. Buxton. An analogical representation of space and time. *Image and Vision Computing*, 10(7):467–478, 1992.
- [15] D. Ayers and M. Shah. Monitoring human behavior from video taken in an office environment. *Image and Vision Computing*, 19(12):833–846, October 2001.
- [16] N. Nguyen, H. Bui, S. Venkatesh, and G. West. Recognising and monitoring high-level behaviours in complex spatial environments. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 620–625, Madison, Wisconsin, USA, June 2003.
- [17] J. H. Fernyhough, A. G. Cohn, and D. Hogg. Generation of semantic regions from image sequences. In *European Conference on Computer Vision*, volume 2, pages 475–484, Cambridge, England, April 1996.
- [18] N. Johnson and D. Hogg. Learning the distribution of object trajectories for event recognition. *Image and Vision Computing*, 14(8):609–615, August 1996.
- [19] 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.
- [20] D. Makris and T. Ellis. Path detection in video surveillance. *Image and Vision Computing*, 20(12):895–903, October 2002.

- [21] D. Makris and T. Ellis. Automatic learning of an activity-based semantic scene model. In *IEEE Conference on Advanced Video and Signal Based Surveillance*, Miami, FL, USA, July 2003.
- [22] D. Makris and T. Ellis. Spatial and probabilistic modelling of pedestrian behaviour. In *British Machine Vision Conference*, volume 2, pages 557–566, Cardiff, UK, September 2002.
- [23] M. Brand and V. Kettner. Discovery and segmentation of activities in video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):844–851, August 2000.
- [24] J. Ferryman, editor. *Third IEEE International Workshop on Performance Evaluation of Tracking and Surveillance*, Copenhagen, June 2002.
- [25] D. Comaniciu, V. Ramesh, and P. Meer. Kernel-based object tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(5):564–575, May 2003.
- [26] L. M. Fuentes and S. A. Velastin. People tracking in indoor surveillance applications. In *2nd IEEE Workshop on Performance Evaluation of Tracking and Surveillance*, Hawaii, 2001.
- [27] C. J. Needham and R. D. Boyle. Tracking multiple sports players through occlusion, congestion and scale. In *British Machine Vision Conference*, volume 1, pages 93–102, Manchester, 2001.
- [28] A. Bobick, S. Intille, J. Davis, F. Baird, C. Pinhanez, L. Campbell, Y. Ivanov, A. Schutte, and A. Wilson. The kidsroom: A perceptually-based interactive and immersive story environment. *PRESENCE: Teleoperators and Virtual Environments*, 8(4):367–391, August 1999.
- [29] I. Yoda, D. Hosotani, and K. Sakaue. Ubiquitous stereo vision for controlling safety on platforms in railroad stations. In *Asian Conference on Computer Vision*, volume 2, pages 770–776, Jeju, Korea, 2004.
- [30] J.-W. Kim, K.-S. Choi, B.-D. Choi, and S.-J. Ko. Real-time system for counting the number of passing people using a single camera. In *Lecture Notes in Computer Science*, volume 2781, pages 466–473, 2003.
- [31] X. Chen and J. Yang. Towards monitoring human activities using an omnidirectional camera. In *International Conference on Multimodal Interfaces*, pages 423–428, Pittsburgh, 2002.
- [32] J. Krumm, S. Harris, B. Meyers, B. Brummitt, M. Hale, and S. Shafer. Multi-camera multi-person tracking for EasyLiving. In *Third IEEE Workshop on Visual Surveillance*, pages 3–10, Dublin, 2000.
- [33] S. Birchfield. Elliptical head tracking using intensity gradients and color histograms. In *IEEE Conference on Computer Vision and Pattern Recognition*, Santa Barbara, CA, June 1998.
- [34] M. Isard and A. Blake. Icondensation: Unifying low-level and high-level tracking in a stochastic framework. In *European Conference on Computer Vision*, volume 1, pages 893–908, 1998.
- [35] F. Liu, X. Lin, S. Z. Li, and Y. Shi. Multi-modal face tracking using bayesian network. In *Proceedings of IEEE International Workshop on Analysis and Modeling of Faces and Gestures*, Nice, France, 2003.
- [36] H. Nait-Charif and S. J. McKenna. Head tracking and action recognition in a smart meeting room. In *4th IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS-ICVS)*, Graz, Austria, 2003.
- [37] C. R. Wren, A. Azarbayejani, T. Darrell, and A. Pentland. Pfunder: Real-time tracking of the human body. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):780–785, 1997.
- [38] A. Blake and M. Isard. *Active Contours*. Springer-Verlag, 2000.
- [39] S.J. McKenna, S. Jabri, Z. Duric, A. Rosenfeld, and H. Wechsler. Tracking groups of people. *Computer Vision and Image Understanding*, 80(1):42–56, 2000.
- [40] M. Isard and A. Blake. Contour tracking by stochastic propagation of conditional density. In *European Conference on Computer Vision*, volume 1, pages 343–356, 1996.
- [41] H. Nait-Charif and S. J. McKenna. Tracking poorly modelled motion using particle filters with iterated likelihood weighting. In *Asian Conference on Computer Vision (ACCV)*, pages 156–161, Jeju Island, Korea, January 2004.
- [42] M. Pitt and N. Shephard. Filtering via simulation: Auxiliary particle filters. *J. American Statistical Association*, 94(446):590–599, 1999.

- [43] S. J. Roberts. Parametric and non-parametric unsupervised cluster analysis. *Pattern Recognition*, 30(2):261–272, 1997.
- [44] 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.
- [45] 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.
- [46] J. Rissanen. Modelling by shortest data description. *Automatica*, 14:465–471, 1978.
- [47] H. S. Sawhney and S. Ayer. Compact representations of videos through dominant and multiple motion estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(8):814–830, August 1996.
- [48] M. Walter, A. Psarrou, and S. Gong. Data driven gesture model acquisition using minimum description length. In *British Machine Vision Conference*, Manchester, UK, 2001.
- [49] C. Carson, S. Belongie, H. Greenspan, and J. Malik. Blobworld: Color- and texture-based image segmentation using EM and its application to image querying and classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(8):1026–1038, August 2002.
- [50] 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.
- [51] 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.
- [52] G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.
- [53] J.-L. Gauvain and C.-H. Lee. Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. *IEEE Transactions on Speech and Audio Processing*, 2(2):291–298, April 1994.
- [54] 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.
- [55] P. Cheeseman and J. Stutz. Bayesian classification (Autoclass): Theory and results. In U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors, *Advances in Knowledge Discovery and Data Mining*, pages 153–180. AAAI Press and MIT Press, 1996.
- [56] D. Geiger and D. Heckerman. Parameter priors for directed acyclic graphical models and the characterization of several probability distributions. *The Annals of Statistics*, 30:1412–1440, 2002.
- [57] 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.
- [58] 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.
- [59] C. Hue, J.-P. Le Cadre, and P. Pérez. Sequential Monte Carlo methods for multiple target tracking and data fusion. *IEEE Transactions on Signal Processing*, 50(2):309–325, February 2002.
- [60] M. Isard and J. MacCormick. BraMBLe: A Bayesian multiple-blob tracker. In *IEEE International Conference on Computer Vision*, volume 2, pages 34–41, Vancouver, 2001.
- [61] Z. Khan, T. Balch, and F. Dellaert. An MCMC-based particle filter for tracking multiple interacting targets. In *European Conference on Computer Vision*, volume 4, pages 279–290, Prague, 2004.
- [62] J. MacCormick and A. Blake. A probabilistic exclusion principle for tracking multiple objects. In *IEEE International Conference on Computer Vision*, pages 572–587, Kerkyra, Greece, 1999.

Sequence annotation	Number of sequences	Average duration (frames)	Tracking errors
RH	11	103	0
RSR	7	504	0
RSH	4	470	0
RCR	7	514	1
RCH	4	561	0
HR	11	119	1
HSR	4	536	0
HSH	16	506	0
HSCH	5	998	0
HCR	4	619	0
HCH	11	672	1
HCSH	4	1150	0
RF	5	469	0
RSF	1	580	0
RCF	1	608	0
HSF	1	571	0
Totals	96	-	3

Table 1. The image sequences acquired. Labels R , H , S , C and F denote atomic instructions for the actor to go to the rear door, go to the hall door, sit on the sofa, sit on the chair and fall over, respectively. Tracking errors occurred in 3 of the 96 sequences.

Sequence annotation	Atomic instructions	E_{ins}	E_{sub}	E_{del}	Erroneous annotations
RH	22	1	0	0	RFH
RSR	21	1	0	0	RSFR
RSH	12	0	0	0	
RCR	21	0	0	1	RR
RCH	12	0	0	0	
HR	22	0	1	0	HH
HSR	12	0	0	0	
HSH	48	4	0	0	HFSH, HSSH ($\times 3$)
HSCH	20	0	0	0	
HCR	12	0	0	0	
HCH	33	2	0	1	HH, HCCH ($\times 2$)
HCSH	16	1	0	0	HCCSH
RF	10	2	0	0	RFFF
RSF	3	0	0	0	
RCF	3	0	0	0	
HSF	3	0	0	0	
Totals	270	11	1	2	-

Table 2. Annotation errors when $\tau_D = 2$. Insertion, substitution and deletion errors are denoted E_{ins} , E_{sub} and E_{del} respectively.

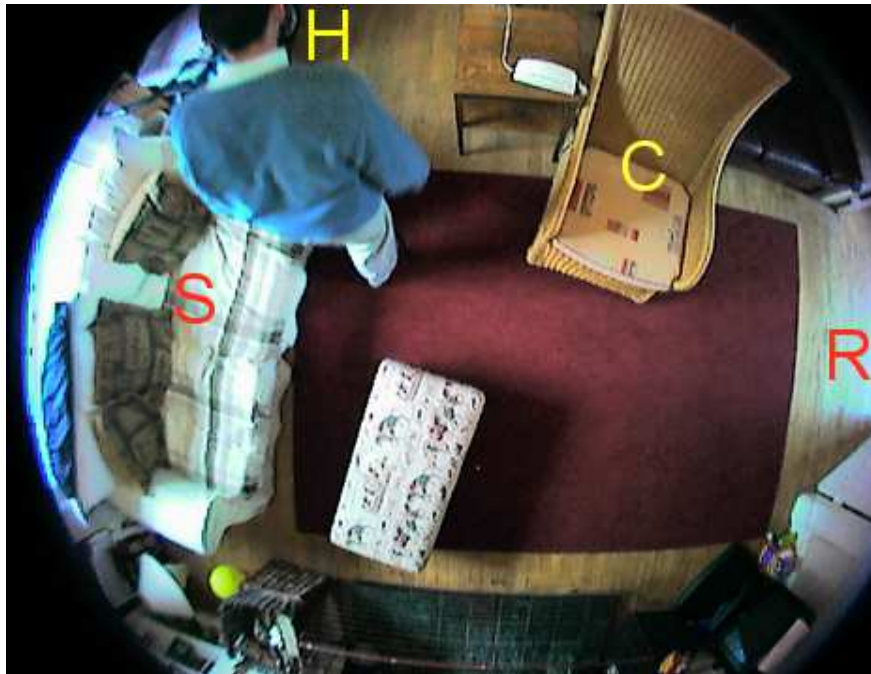


Fig. 1. A scenario used to illustrate the method. A ceiling-mounted, wide-angle camera acquires colour image sequences of a living room. Salient regions are labelled: a sofa (S), a chair (C), the hall door (H) and the rear door (R).

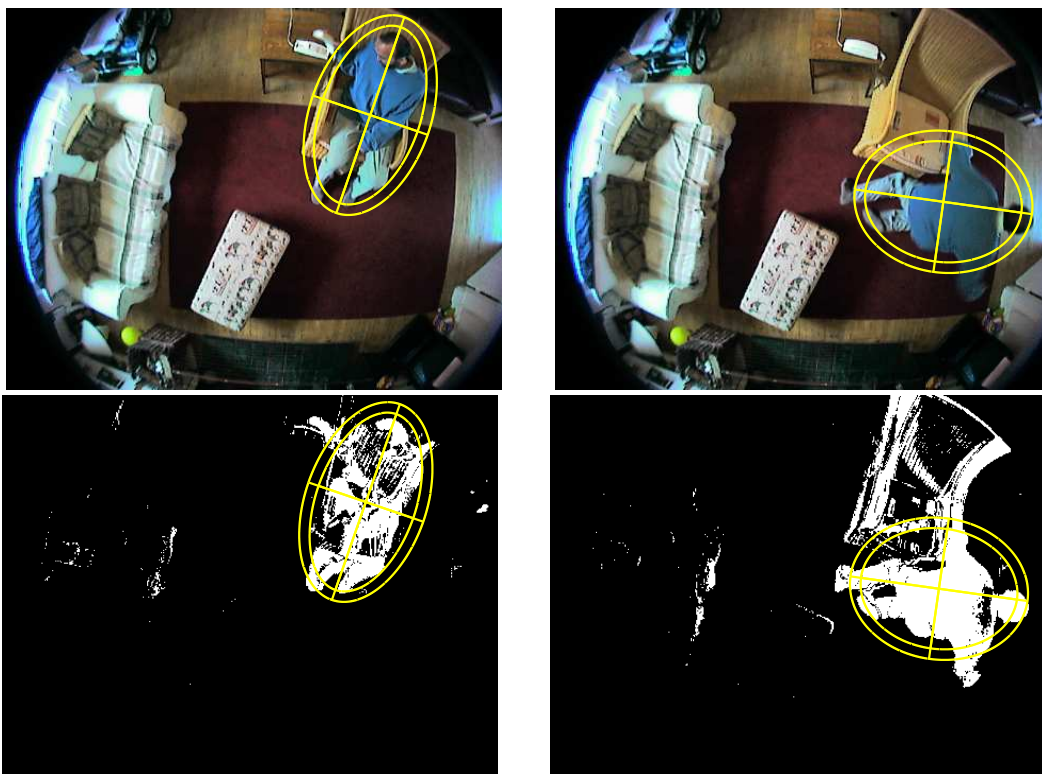


Fig. 2. Two examples of the strongest particle overlaid on the input images (top) and pixel foreground images (bottom). The interior ellipse represents the state estimate and the outer ellipse contains the annular region used to compute the likelihood.

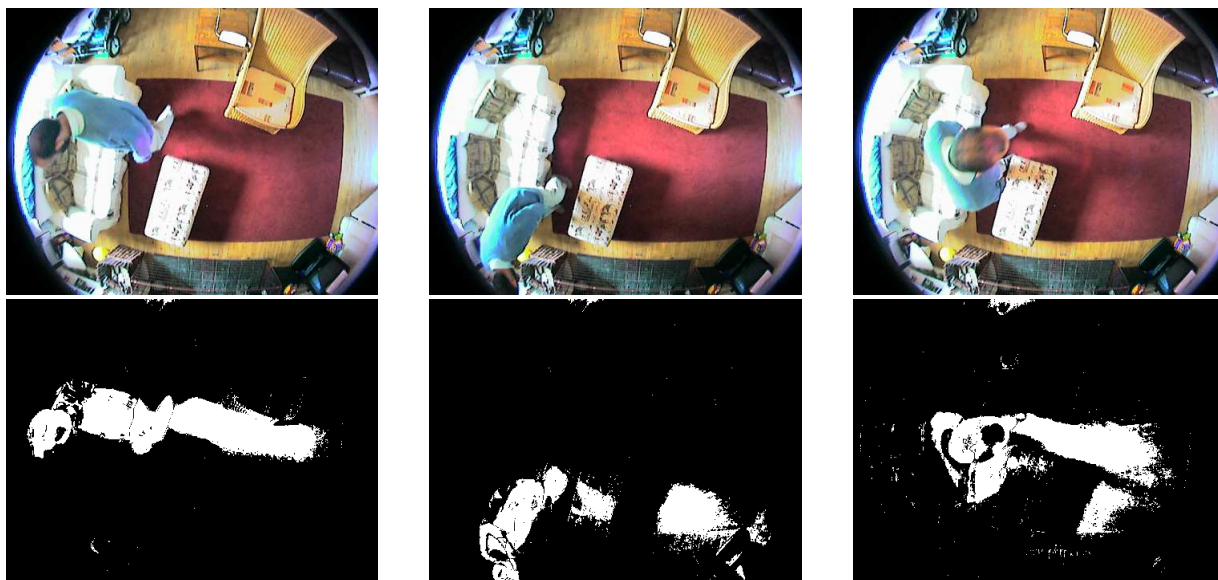


Fig. 3. Three examples of the adaptive background method providing a poor cue due to strong shadows. Top: input images. Bottom: corresponding adaptive background subtraction images.

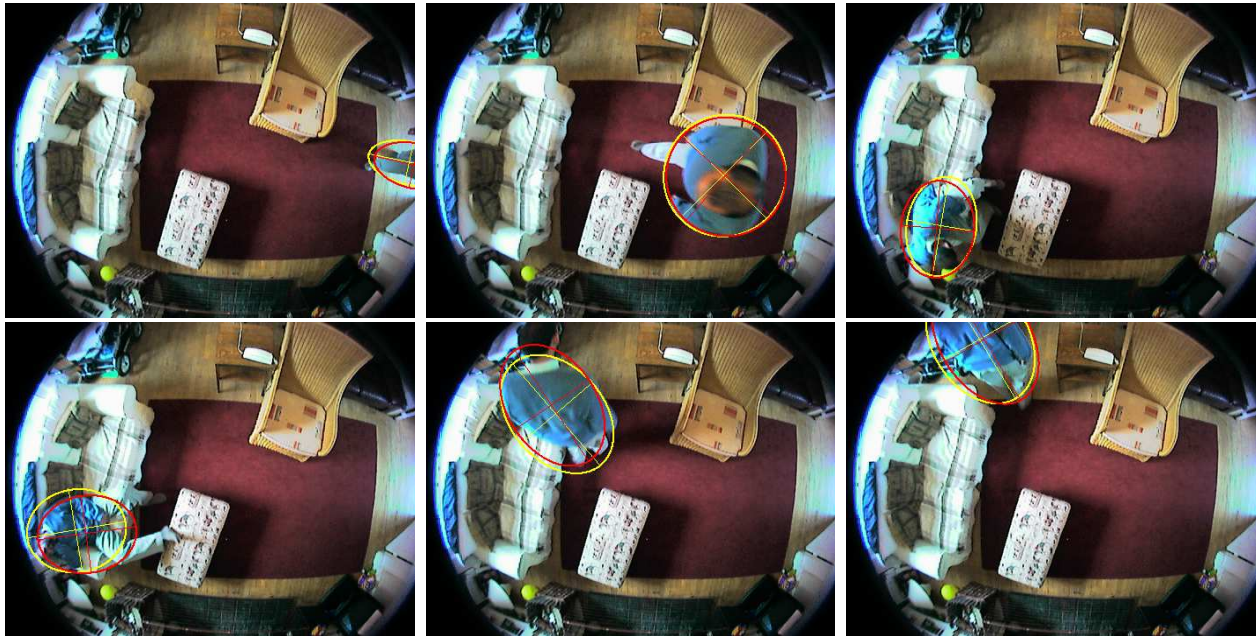


Fig. 4. Examples of ellipse estimates obtained by the tracking system. Two ellipses are displayed: the strongest particle and the mean.

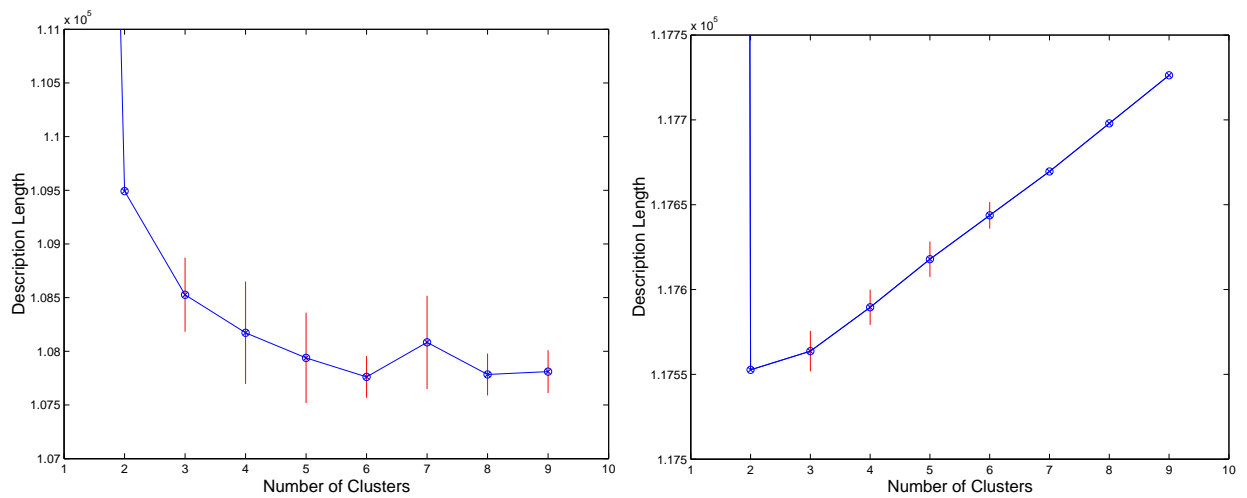


Fig. 5. The description length, \mathcal{C} , when learning inactivity zones using maximum likelihood (left) and the penalised likelihood (right). The values plotted are means obtained over ten runs for each model order. Error bars denote \pm one standard deviation.

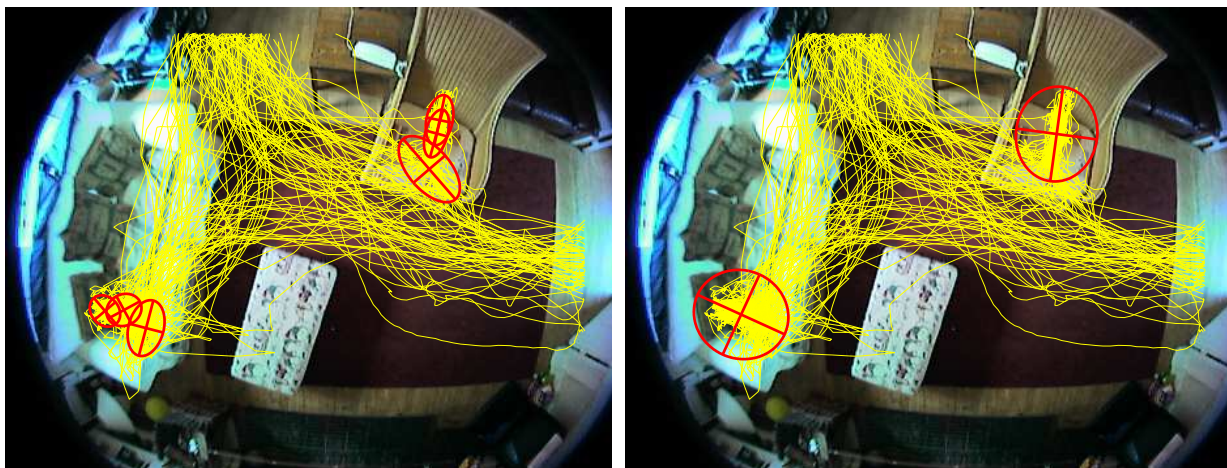


Fig. 6. Example results for learning inactivity zones. Left: ML with $K = 6$. Right: penalised likelihood with $K = 2$.

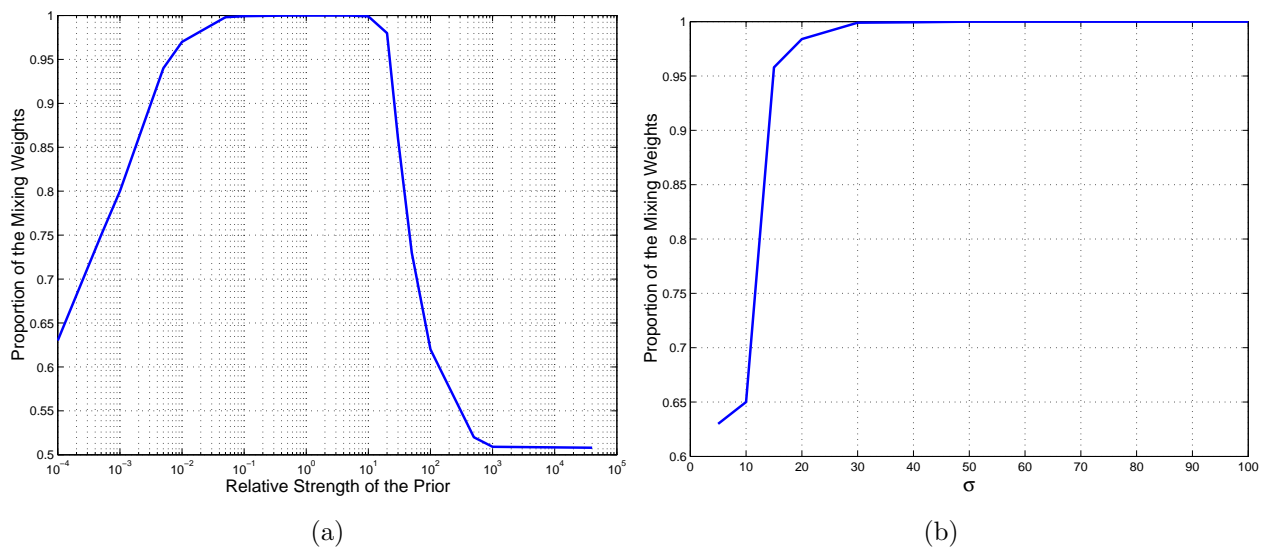


Fig. 7. The proportion of the mixing weights accounted for by the strongest two Gaussian components plotted against (a) the relative strength of the prior, ω/N , for $\sigma = 40$ and (b) the scale prior, σ , for $\omega = 0.2N$.

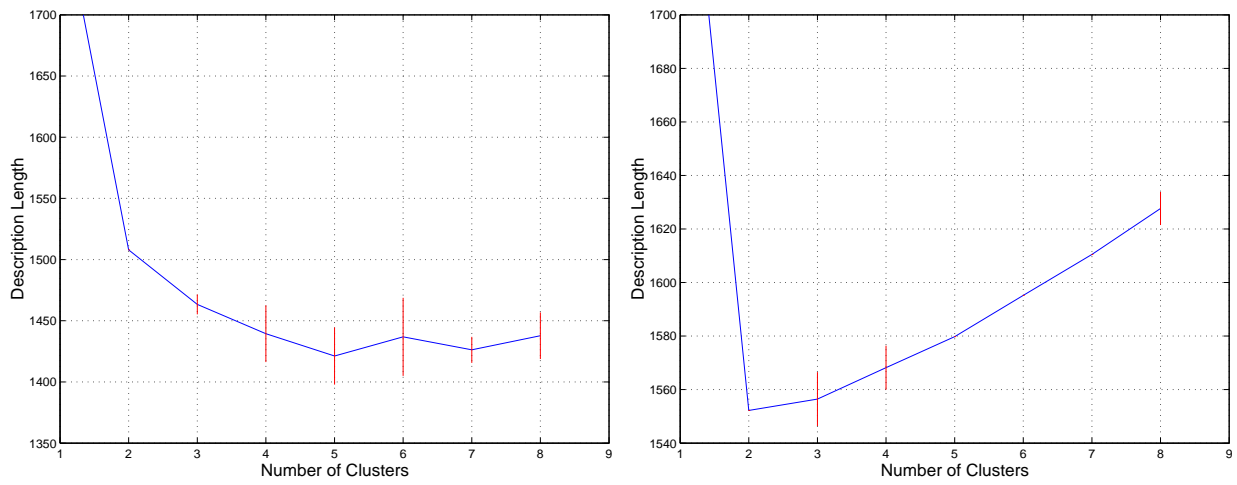


Fig. 8. The description length, C , when learning entry zones in 2D using maximum likelihood (left) and the penalised likelihood (right). The values plotted are means obtained over ten runs for each model order. Error bars denote \pm one standard deviation.

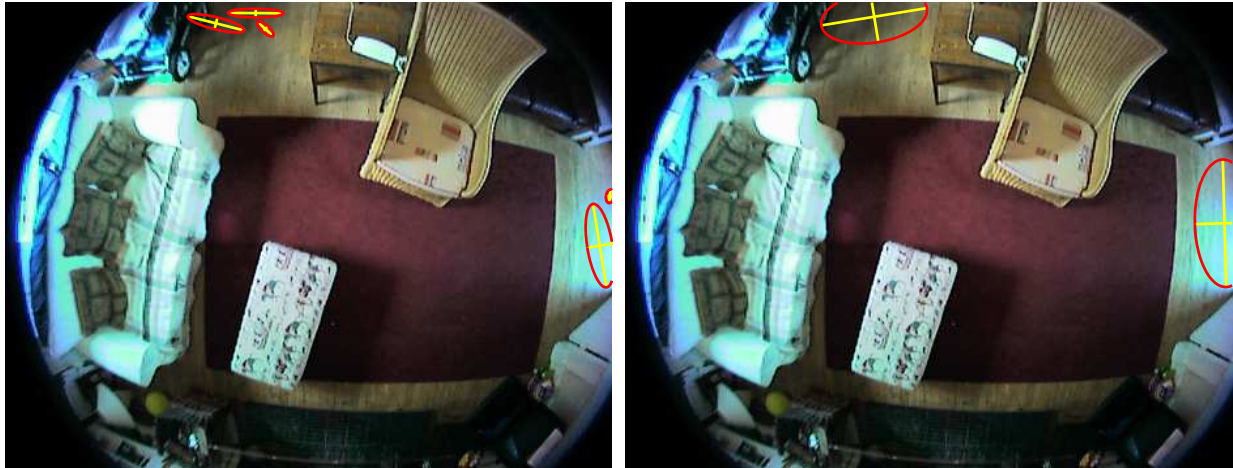


Fig. 9. Example results for learning entry zones as 2D regions Left: ML with $K = 5$. Right: penalised likelihood with $K = 2$.

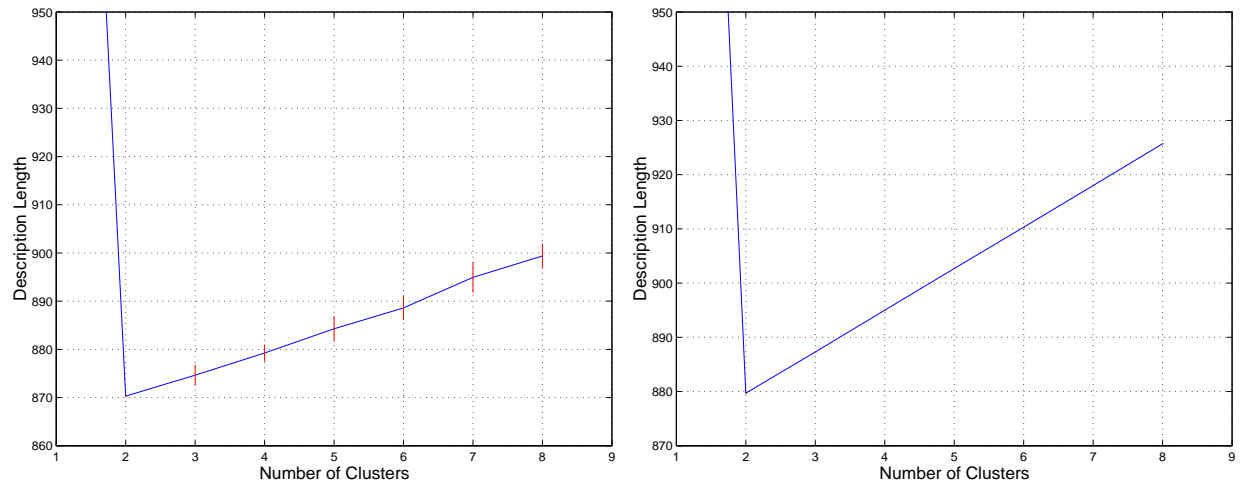


Fig. 10. The description length, \mathcal{C} , when learning entry zones on a 1D contour using maximum likelihood (left) and the penalised likelihood (right). The values plotted are means obtained over ten runs for each model order. Error bars denote \pm one standard deviation.

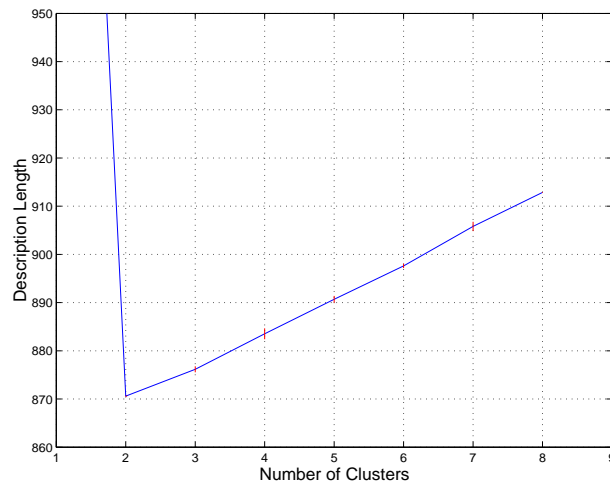


Fig. 11. The description length, \mathcal{C} , when learning entry zones on a 1D contour using penalised likelihood with $\sigma = 20$. The values plotted are means obtained over ten runs for each model order. Error bars denote \pm one standard deviation.



Fig. 12. Example results for learning entry zones on a 1D contour. Left: ML with $K = 2$. Right: penalised likelihood with $K = 2$. Note that the 1D closed contour used was close to but not identical with the image border.

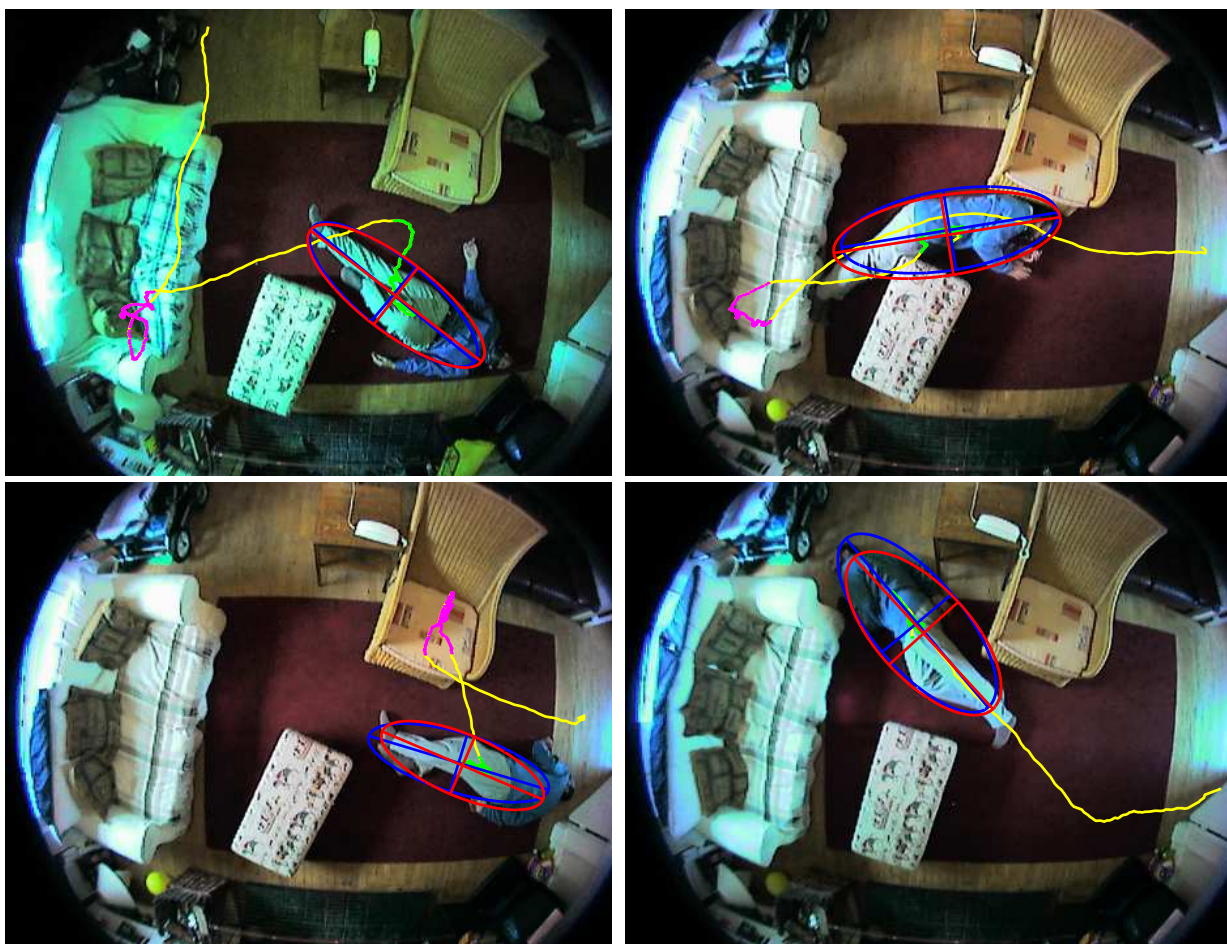


Fig. 13. Examples of segmented trajectories and detected inactivity.