Contents lists available at ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

Automated detection of unusual events on stairs

Jasper Snoek [a,*], Jesse Hoey [b], Liam Stewart [a], Richard S. Zemel [a], Alex Mihailidis [a,c]

[a] *Department of Computer Science, University of Toronto, 10 Kings College Road, Toronto, Ont., Canada M5S 3H5*
[b] *School of Computing, University of Dundee, Dundee, Scotland, UK*
[c] *Department of Occupational Science and Occupational Therapy, University of Toronto, 500 University Avenue, Toronto, Ont., Canada*

ARTICLE INFO

Article history:
Received 22 February 2007
Accepted 10 April 2008

Keywords:
Human motion analysis
Hidden Markov model
Gait analysis
Anomaly detection
Event recognition
Bayesian tracking
Particle filter
Automated video analysis
Machine learning
Biomedical analysis
Stairs

ABSTRACT

This paper presents a method for automatically detecting unusual human events on stairs from video data. The motivation is to provide a tool for biomedical researchers to rapidly find the events of interest within large quantities of video data. Our system identifies potential sequences containing anomalies, and reduces the amount of data that needs to be searched by a human. We compute two sets of features from a video of a person descending a stairwell. The first set of features are the foot positions and velocities. We track both feet using a mixed state particle filter with an appearance model based on histograms of oriented gradients. We compute expected (most likely) foot positions given the state of the filter at each frame. The second set of features are the parameters of the mean optical flow over a foreground region. Our final classification system inputs these two sets of features into a hidden Markov model (HMM) to analyse the spatio-temporal progression of the stair descent. A single HMM is trained on sequences of normal stair use, and a threshold on sequence likelihoods is used to detect unusual events in new data. We demonstrate our system on a data set with five people descending a set of stairs in a laboratory environment. We show how our system can successfully detect nearly all anomalous events, with a low false positive rate. We discuss limitations and suggest improvements to the system.

1. Introduction

Stairs have long been the subject of study for architects and designers [1], who attempt to build more ergonomic and safe stairs for different public and private situations. Increasingly, stairs have become a subject of interest for biomedical researchers, who realise that, even with perfect design, stairs are inherently difficult for humans to navigate and their use will always lead to accidents. The US Consumer Product Safety Commission estimates that in 2005 alone over one million people received hospital treatment in US due to stair related injuries [2]. Older adults are particularly susceptible to accidents on stairs due to their reduced mobility and weaker musculoskeletal systems. This is of special concern to the growing population of elderly people who wish to age in their homes. Falls in general are the leading cause of accidental mortality and morbidity among the elderly population [3,4] and stairs are a significant cause of falls [1]. In fact, in the United States, the Netherlands and the United Kingdom, steps and stairs are the single most dangerous element in the home [1].

Biomedical researchers study the ways in which adverse events happen on stairs, and aim to identify and predict the causes of these events. One of the major hurdles involved in such research is the gathering of real stair data. Aside from the ethical difficulties of recording stair usage in public or private spaces, there is a technical difficulty imposed by the rarity of adverse events. It is estimated that on public staircases, a slip, stumble, trip, or other loss of balance not resulting in a fall occurs once in 2222 stair uses, while minor accidents such as falls occur only once in 63,000 stair uses [5]. It is hypothesized that the labour intensive process of manually identifying unusual events in stair video data can be avoided with an automated system which is proposed herein.

It is assumed that the system will have access to a database of stair events on a particular set of stairs, where each stair event consists of a single person entering the stairwell and descending the stairs. A stair event (or descent) is considered to be of two types, normal and anomalous. In a normal stair event, the person descends the stairs with no problems, correctly placing their feet on steps without any loss of balance. We consider an *anomalous* event to be one in which the person misses a step at some point in the stair event. More obvious abnormal events, such as a person falling down the stairs, will not be considered. A person can miss a step either by completely overstepping, or by catching their heel on the nosing of a step and slipping off onto the next lower step (a slip). These are the most common anomalous events and account for a combined 65% of all “gait incidents” on stairs [1] (followed by stumbles (17%), balance loss(10%), other (8%)). The primary goal of our system is to filter a large database, removing a large fraction

[*] Corresponding author. Tel.: +1 416 946 8573.
E-mail address: jasper@cs.toronto.edu (J. Snoek).

0262-8856/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.imavis.2008.04.021

of stair events which are sure not to contain anomalies. The remaining data could be forwarded to a human for final analysis. Therefore, while our system should miss as few anomalous events as possible, we can afford a reasonable amount of false positive anomalous events. It is assumed that only a single person is descending the stairs at a time, a limitation that could be overcome with multiple target tracking (such as by using the Bramble system [6]). We only look at descents, as these present a significantly higher risk for adverse events (75% of stair falls causing injury occur during descent [7]), and are of most interest to biomedical researchers and stairwell designers.

This new system operates in five stages as shown in Fig. 2. Two types of features are computed for each frame of video: foot dynamics and overall body motion. The subject's foot positions are tracked using a mixed-state Bayesian sequential estimator with an appearance model based on histograms of oriented gradients (HOGs). Six features are derived from the locations of the feet: vertical and horizontal velocities of both feet and the vertical and horizontal distance between both feet. Body motion is computed as the mean value of the optical flow [8] over the foreground region obtained using an adaptive background subtraction technique. The resulting feature vector consisting of the tracked feet features and the mean flow features forms a time series over each stair descent. A hidden Markov model is then trained to model the statistical progression of these feature vectors over time in normal stair descents. A new stair descent is classified as normal or anomalous by computing its likelihood under the hidden Markov model and comparing it to a threshold.

2. Previous work

Relatively little work has been done to detect anomalous human motion in video. Lee and Mihailidis [9] detect the most severe anomalous motion (falls) by thresholding the diameter and velocity of a background subtracted silhouette. McKenna and Nait-Charif [10] detect deviations from models of normal activities in a home to detect unusual behaviors such as falls. Bauckhage et al. [11] estimate pose by encoding a background subtracted silhouette as a mapping onto a rectangular grid. A feature vector is generated from a concatenation of the grid representations at consecutive frames and a support vector machine is applied to perform binary classification of normal and anomalous sequences of poses. This innovative approach unfortunately requires error-free segmentations of people's silhouettes and suffers from an inability to generalize well to new subjects and gaits. The reason for this is that the class of anomalous poses and motions is simply too large and variable to model. Even with a significant amount of training data of anomalous gait it is generally easy to consider additional cases which are not represented within the training set. This is further highlighted by the fact that each person's individual gait is different and thus what is considered normal gait for one person is anomalous for another. This fact presents a major challenge to detecting anomalous events across multiple subjects. In previous work [12], we have shown that it is more challenging to detect anomalous events on stairs for a person who is not represented in the normal event training data. Medical studies have shown [13] that gait is unique across individuals. In psychological studies [14] people have been able to easily identify others by observing only their gait.

As such, a significant amount of work exists in attempting to identify people based on their gait. Niyogi and Adelson [15] detect individual gait by tracking the progression through time of skeletons fitted to background subtracted silhouettes. Little and Boyd [16] recognize people by computing periodic characteristic features of optical flow. While significant research exists into detecting individual gait [17–19] relatively little work exists on detecting anomalous gait, particularly on stairs.

There is some work on detecting anomalous behavior in video in the context of visual surveillance [20] or user modeling [21,22]. However, these approaches use coarse features such as positions and velocities of people within a scene and attempt to characterise trajectories. A larger body of computer vision research has looked into modeling the motion of the human body in fine detail. Periodic motion of walking figures is analysed in [23] by computing self-similarity of a segmented image region with itself over multiple time scales. The Fourier transform of the resulting correlations gives indications of the periodicity of the motion. The motion history (MHI) [24] is a descriptor of temporally localised image changes. However, these works do not attempt to recognise anomalous events and do not look at motion on stairs.

Little work has been done on characterising human motion on stairs. Notable exceptions are work done on motion capture data of people ascending and descending stairs, in which recovered joint angles are mapped to a subspace that can be used for synthesis [25]. However, this work does not use video and does not attempt recognition of unusual events. Human gait on staircases was analysed in [26] by fitting a skeletal model to the view-based human form, and then modeling the joint angles as a dynamical system. This was used to classify gaits such as walking, running and descending stairs, but no work was done on recognising unusual events within each of these motion types. An interesting study in [27] used a camera mounted above a side-by-side public stair-

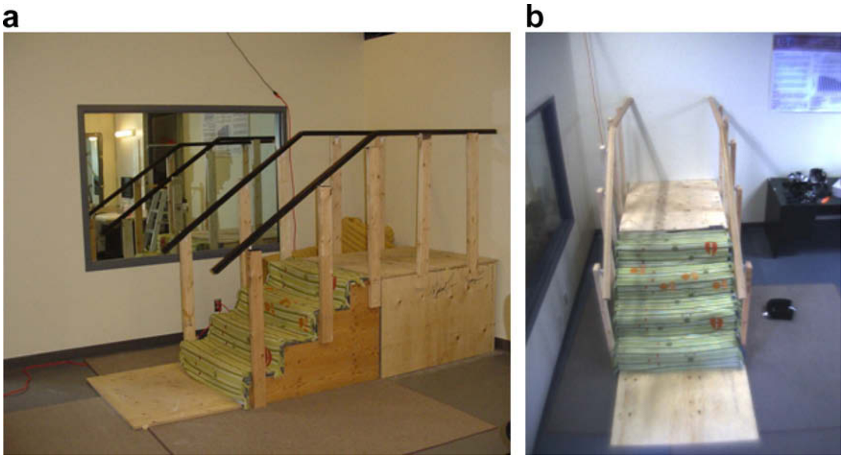

Fig. 1. The stairs (a) and the view from the overhead camera (b).

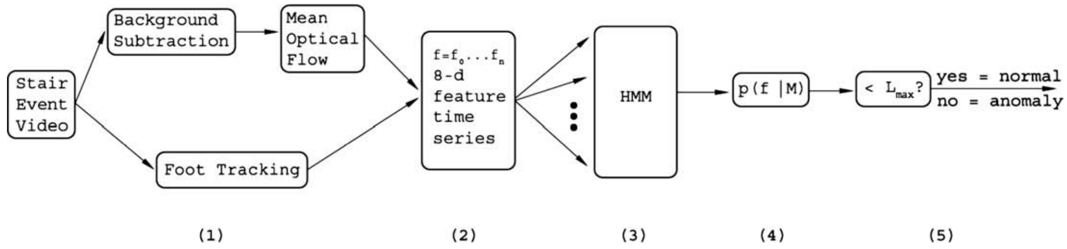

Fig. 2. Overview of system. (1) A person's silhouette in a video of a stair descent is extracted using adaptive background subtraction. Mean optical flow is computed over the silhouette to compute the overall body motion and the person's feet are tracked as they progress down the stairs. (2) An 8-feature time series is derived from the optical flow and foot motion. (3) The likelihood of the sequence is computed given a trained hidden Markov model (HMM) and (4) compared to a threshold L^*. (5) If the likelihood is below a certain threshold the descent is classified as anomalous and otherwise normal.

way and escalator to implement a prompting device that would encourage people to use the stairs if they were about to use the escalator. Background subtraction was used to determine if people were using the escalator or the stairs, and the work also addressed some issues of multiple people on the stairs. However, there is no recognition of unusual events.

Of particular relevance to this paper is the subset of gesture recognition work that uses hidden Markov models to classify gestures. Other vision-based gesture recognition approaches classify gestures using multiple HMMs, where a separate HMM is used to model each gesture (e.g. see [28–31]). Starner and Pentland showed that HMMs could be used to classify hand gestures in video of American sign language [28]. Kapoor and Picard used HMMs to detect head nods and shakes [32]. Brand et. al. derived coupled HMMs and used them to classify Tai Chi gestures from the tracked locations of hands in video [29]. In [31], Vogler et al. used parallel HMMs to classify American sign language gestures in 3D visual tracking data and different walking gaits from motion capture data. In all the above mentioned approaches separate HMMs were used to model each gesture to be classified. Each type of gesture is modeled by a single HMM trained on that type of gesture. Classification then proceeds by computing the likelihoods of the test data under all the HMMs and classifying according to the HMM which assigns the highest likelihood. Particle filters have been widely used in computer vision primarily for the purpose of approximating the Bayesian density of the position of an object throughout a sequence of images (see [33,34,6,35] for examples).

3. Stair event classification

In this section, we detail the algorithms and techniques used to model and classify the motion of people on stairs. Section 3.1 details a procedure to identify feet, model their motion and then track them as they progress down the stairs using a Bayesian Monte Carlo sampling method. Section 3.2 then outlines a simple background subtraction procedure used to identify a person on the stairs, and how the motion of the person is quantified by computing optical flow over the background subtracted region. Then in Section 3.3 we use the information gathered in a hidden Markov model to classify stair descents as either normal or anomalous.

3.1. Tracking feet

A person's feet are tracked using a probabilistic Bayesian sequential estimation technique. We estimate $P(x_t \mid y_{1:t})$, the distribution over the positions of the two feet, $x_t = \{x_t^l, x_t^r\}$, given a sequence of observed data (images) $y_{1:t} = \{y_1, \ldots y_t\}$. Assuming conditionally independent observations and first order Markovian state dynamics, we get the recursive Bayesian solution:

$$P(x_t|y_{1:t}) \propto P(y_t|x_t) \int_{x_{t-1}} P(x_t|x_{t-1}) P(x_{t-1}|y_{1:t-1}) \, \mathrm{d}x_{t-1} \qquad (1)$$

In order to solve this equation one must model the observation likelihood $P(y_t \mid x_t)$, the probability of observing the object given a configuration, and its dynamics $P(x_t \mid x_{t-1})$. In addition a suitable prior $P(x_1)$ must be specified to initialize the algorithm. We model the appearance of feet using a histogram based representation of their contours and derive an observation likelihood accordingly (Section 3.1.1). The dynamics of the feet is a mixed-state dynamical model (Section 3.1.2). Tracking is initialized by specifying $P(x_1)$ as a uniform distribution spatially over the top of the staircase. We approximate the recursion in Eq. (1) using an importance sampling technique (Section 3.1.3).

3.1.1. An appearance model of feet

The appearance of feet is modeled as a template histogram of oriented gradients (HOG) [36]. This technique works by splitting image windows into a grid of cells, and computing histograms of the spatial gradient orientations in the image over each cell. The combination of all the histograms for all cells provides a representation of the appearance of the feet. Spatial gradient information is a more robust measure than color since it can, for example, account for differently colored or shaped shoes. Using histograms of oriented gradients as the appearance representation provides us a model of the contours of feet that is invariant to slight changes in orientation, allows for non-rigid contours and provides a smooth observation likelihood. These properties allow a small set of hypotheses to represent the possible locations of an object. This is particularly important for tracking methods, such as particle filters, that maintain multiple hypotheses.

We compute the HOG over an image window as follows. First, the spatial gradients of the image are computed by convolving the image with a derivative of a Gaussian filter. The resulting gradients are normalized using L2 normalization followed by clipping values above a maximum [36].[1] Second, the image window over which we are trying to compute the HOG is split spatially into a 3×3 grid of cells. For each cell, N_h histogram bins are created, representing spatial gradient orientations from 0 to 2π, where the ith bin is for orientations in $\frac{2\pi i}{N_h - 1}$, $i = 0 \ldots N_\mathrm{h} - 1$. Each pixel then submits a vote for its gradient direction to the histogram for the cell it is located in. Votes in neighboring bins are bilinearly interpolated to reduce aliasing. The number N_h of histogram bins used for matching provides a control on the flexibility of changes in orientation of the object to be matched. Similarly, the number of grid cells used provides a control on the spatial flexibility of the match.

[1] Note that, in [36], the authors report that Gaussian smoothing in the gradient computation significantly decreases performance on their detection task. In contrast, we found that in our task Gaussian smoothing significantly improved foot detection. Presumably this is because the smoothing operation removes texture both on the stairs and on shoes.

In order to compute the likelihood of a foot being present in an image window given a foot location, x_t (either left or right foot), we need to compare the HOG of the image window, q^c, to a reference HOG, q^*. We use a Bhattacharyya distance measurement given by [37]:

$$D[q^*, q^c] = \left[1 - \sum_{n=1}^{N} \sqrt{q_n^*, q_n^c}\right]^{\frac{1}{2}} \tag{2}$$

Then, the observation likelihood is [38]:

$$P(y_t|x_t) \propto \mathrm{e}^{-\lambda D^2[q^*, q^c_{y_t}]} \tag{3}$$

where q_y^c is the HOG derived from y_t, and λ is a scaling parameter (we used $\lambda = 50$).

For our experiments histograms were computed over 21×21 pixel windows centered at the hypothesized foot location. We used $N_\mathrm{h} = 10$ histogram bins, and discarded low amplitude gradients. See Fig. 3 for an example HOG of a foot. Reference histograms for left and right feet were computed from an independent set of stair sequences, with manually annotated foot positions. HOGs were computed for each frame, and averaged to produce the final template histograms. Fig. 4 shows the log likelihood of a match to the right foot using our method for each pixel in an example image.

3.1.2. Dynamics

The dynamics of foot positions over a sequence is modeled by the density $P(x_t \mid x_{t-1})$. A model of the foot dynamics is particularly important when the appearance model fails. This occurs in our system during virtually every stair descent when the knees and thighs occlude the view of a foot during a step (see Figs. 9 and 8). An accurate model of the dynamics of the feet allows us to infer their position while they are not visible and predict where they will reappear. During a stair descent the feet exhibit a number of different motions (i.e. they are stationary while on a stair, accelerate during a step and decelerate as the foot reaches the next step), and a single dynamical model is not in general sufficient to capture these different modes. It is preferable to specify a separate dynamical model for each type of motion the feet can exhibit. Assuming N motion types, we index the dynamics by $i = 1 \ldots N$, such that we have N dynamical models. In fact, we are postulating a new set of discrete-valued hidden states z_t, where each state $z_t = i$ corresponds to the ith dynamics model. The state of the system is now mixed [39,40], comprised a continuous component, x_t, representing the left and right foot positions and a discrete component, z_t, representing the current dynamics mode. We represent this extended state as $X = \{x, z\}, x \in \mathbb{R}, z \in \{1, \ldots, N\}$. The process density is now given by $p(X_t|X_{t-1}) = p(x_t^\mathrm{r}, x_t^\mathrm{l}, z_t|x_{t-1}^\mathrm{r}, x_{t-1}^\mathrm{l}, z_{t-1})$, where we have used the fact that x represents the positions of both left and right feet, x^l and x^r, respectively.

Several conditional independence assumptions are used to simplify the dynamics. We assume that the dynamics modes specify the joint velocities of the two feet, but that each foot moves independently given this joint velocity. This assumption means that the motion of the two feet are correlated, but that their positions are not (e.g. the position of the left foot depends only on its position in the previous time step, and on the velocity, but not on the position of the right foot). This independence assumption will also reduce the complexity of our sampling approach detailed in the next section. To accomplish this, we first assume that the two parts of the state space, x and z, are only conditionally dependent within the same time-slice, such that

$$P(X_t|X_{t-1}) = P(x_t|z_t, X_{t-1})P(z_t|X_{t-1}) = p(x_t|z_t, x_{t-1})p(z_t|z_{t-1})$$

We then model the foot positions being coupled to the dynamics modes through the velocity of the feet $v = \{v_x^\mathrm{r}, v_y^\mathrm{r}, v_x^\mathrm{l}, v_y^\mathrm{l}\}$, a four dimensional vector giving horizontal and vertical *velocity* for each (left and right) foot. Writing the temporal dynamics over the discrete dynamics mode states, z, as a simple multinomial: $P(z_t = i \mid z_{t-1} = j) = T_{ij}$, the dynamics is written as

$$P(x_t, z_t = i|x_{t-1}, z_{t-1} = j) = \int_{v_t} p(x_t|x_{t-1}, v_t)p(v_t|z_t = i)T_{ij}\, \mathrm{d}v_t$$

where $p(v_t \mid z_t = i)$ is a Gaussian mixture consisting of N_v Gaussians, each with mixing proportion w_{ik}, mean μ_{ik} and covariance Σ_{ik}, for $k \in 1 \ldots N_v$

$$p(v_t|z_t = i) \sim \sum_{k=1}^{N_v} w_{ik} \mathscr{N}(\mu_{ik}, \Sigma_{ik})$$

Finally, we assume that each foot moves independently given the mode dynamics, v, such that the process dynamics is factored as

$$p(x_t|x_{t-1}, v_t) = p(x_t^\mathrm{l}|x_{t-1}^\mathrm{l}, v_t)p(x_t^\mathrm{r}|x_{t-1}^\mathrm{r}, v_t) = p(x_t^\mathrm{l}|x_{t-1}^\mathrm{l}, v_t^\mathrm{l})p(x_t^\mathrm{r}|x_{t-1}^\mathrm{r}, v_t^\mathrm{r})$$

where v_t^r (v_t^l) is the right (left) foot velocity at time t. The two time-slice Bayesian network (2TBN) for the model we use is shown in Fig. 5, and encapsulates the conditional independence assumptions used to simplify the dynamics.

Given the velocity, v_t, then the dynamics of each foot is a linear Gaussian model with additive noise, such that we have

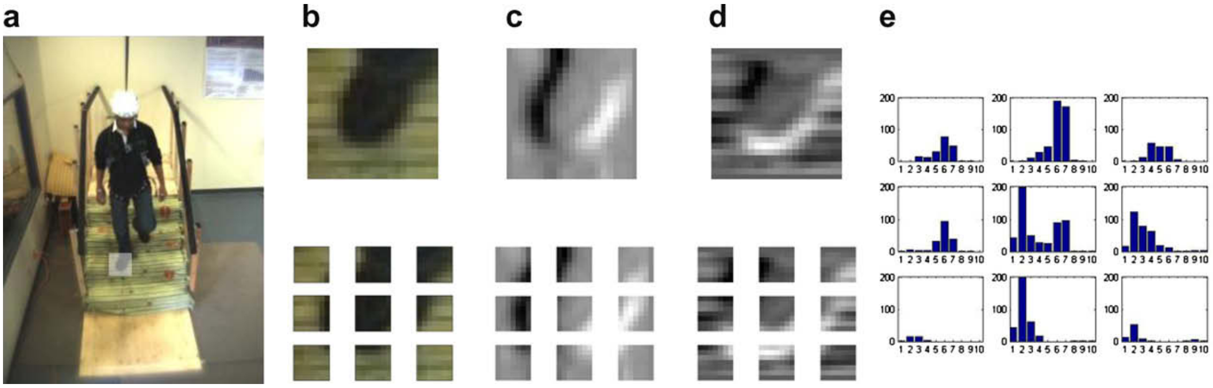

Fig. 3. An example of the histograms of oriented gradients computation over a foot where (a) is the original image with a 21×21 window highlighted over the right foot, (b) is the window around the foot and the corresponding 3×3 grid, (c) shows the horizontal gradients over the window, (d) shows the vertical gradients over the window and (e) shows the final histogram representation where the orientations of all pixels in each grid location are binned into the corresponding histogram. This shows the resulting nine histograms each with ten orientation bins (from 0 to π) along the x-axis and the number of pixels in each bin on the y-axis. Pixels with small amplitude gradients were discarded.

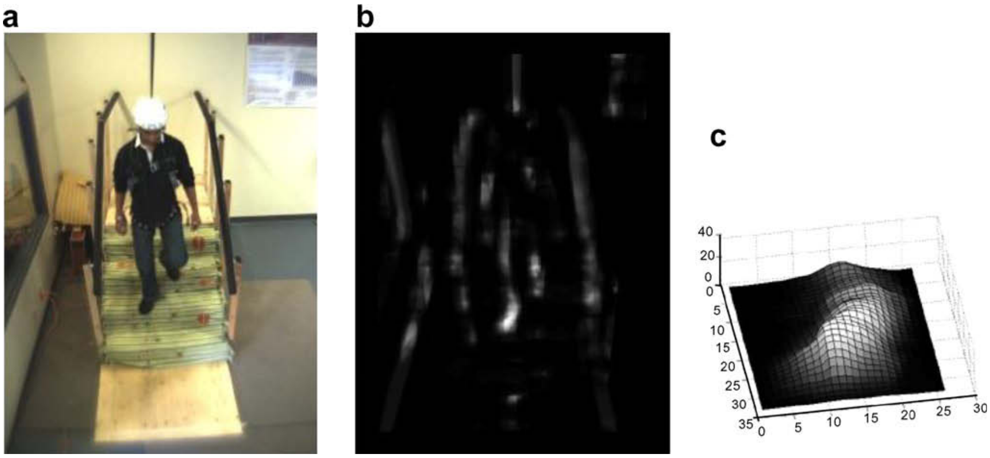

Fig. 4. This figure shows the log likelihood that the 21 × 21 image window centered at each pixel in the image matches the template HOG for the right foot (a) is the original image, in (b) each pixel's value is the log likelihood that the window centered at that pixel matches the template HOG for the right foot and (c) shows a 3D surface map of the log likelihoods over a small window centered on the right foot.

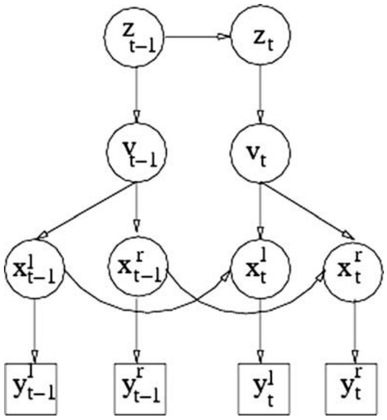

Fig. 5. A two time-slice dynamic Bayesian network (2TBN) dependency graph of the mixed-state dynamical system. The full dynamic Bayesian network is obtained by repeating this slice for the number of time steps in the data.

$$x^k_t = x^k_{t-1} + v^k_t + \lambda$$

where $k = \{l, r\}$ and λ is zero-mean Gaussian noise with fixed covariance C:

$$\lambda \sim \mathcal{N}(\lambda; 0, C)$$

To learn the parameters of this model, we first specify the fixed covariance C of the sub-process noise, λ. This restricts the learning to only the dynamics in the Markov chain over the z variable and the density $p(v \mid z)$.

The resulting relationship between the unobserved dynamical states $z_{1:t}$ and the observed dynamical motion $v_{1:t}$ is a hidden Markov model for which the algorithms for learning and inference are well understood [41] (see Section 3.3 for a description of HMMs). Given a training sequence of hand labeled foot positions $x^*_{1:t} = \{x^*_1, \ldots, x^*_t\}$, a corresponding sequence of dynamical motion $v^*_{1:t} = \{v^*_1, \ldots, v^*_t\}$ is computed using the first order finite differences, $v^*_t = x^*_t - x^*_{t-1}$. The parameters of the HMM, $\{T, w, \mu, \Sigma\}$, are those which maximize the likelihood of the foot position training data. A single 10-state hidden Markov model was used with one added regularization state with high covariance and low mixing proportion. The HMM was trained on 60 (30 normal and 30 anomalous) stair descents. This data was withheld from our experiments in Section 4.2. We used $C = diag(c)$ where $c = 5$ pixels. Fig. 6 shows the behavior of the dynamical model for an example stair descent.

3.1.3. Particle filtering

Now that we have an appearance model of feet and dynamical model of their motion on the stairs we can track their location through a sequence of images using Eq. (1). We approximate this equation using a variant of the Sampling Importance Resampling (SIR) algorithm [42] closely related to the CONDENSATION algorithm [33]. The SIR algorithm is a sequential Monte Carlo (particle filtering) method for estimating probability densities within Bayesian tracking (see [43] for a survey and tutorial of particle filtering methods).

In standard SIR filtering the state of the system X_t at time t is approximated using a finite set of samples $\{X^{(n)}_t, w^{(n)}_t, n = 1, \ldots, N\}$:

$$X^{(j)}_t \sim P(X_t | y_{1:t-1}) \approx \sum_n w^{(n)}_{t-1} P(X_t | X^{(n)}_{t-1}) \tag{4}$$

$$w^{(j)}_t \propto P(y_t | X^{(j)}_t) \tag{5}$$

where the weights w_t are normalized to sum to one, the state X_t refers to the configuration of the tracking target at frame t and $y_{1:t} = (y_1, \ldots, y_t)$ represents a sequence of observations. In Sections 3.1.1 and 3.1.2, we specified the observation density $P(y_t \mid X_t)$ and the state dynamics $P(X_t \mid X_{t-1})$. In our approach, we use a set of samples over the mixed-state $X = \{x, z\}$. Recall that $x = \{x^r, x^l\}$ is the continuous-valued positions of both feet, while z is a discrete-valued index of the dynamics mode. We propagate each sample using the mode dynamics $T_{ij} = p(z_t \mid z_{t-1})$ first, followed by a sampling of the velocity density $p(v \mid z)$, and finally, independent updates of the particles for each left and right foot. Fig. 7 shows the complete algorithm.

We use $N = 300$ samples and a separate template appearance model for each foot. Our algorithm is initialized by selecting random sample sets with high spatial variance over the top of the stairs $\{X^{(n)}_t, w^{(n)}_t, n = 1, \ldots, N\}$ with equal normalized weights. The weight of each sample is then set to be the likelihood that the foot is observed at that sample $w_t(n) = P(y_t | X^{(n)}_t)$ and renormalize the weights. The configuration of each foot at time t is estimated as the mean configuration of its sample set $\{X^{(n)}_t, n = 1, \ldots, N\}$ weighted by the sample weights $w^{(n)}_t$. After resampling (with replacement) each sample is propagated independently through the dynamics by inferring from the hidden Markov model a new state of the dynamics z and velocity v given the previous states of the dynamics. The samples are then resampled and the procedure repeats for each frame in the video sequence. See Fig. 7 for details of the iterative sampling algorithm. Fig. 8 shows the initialization of the sample set and its progression through a

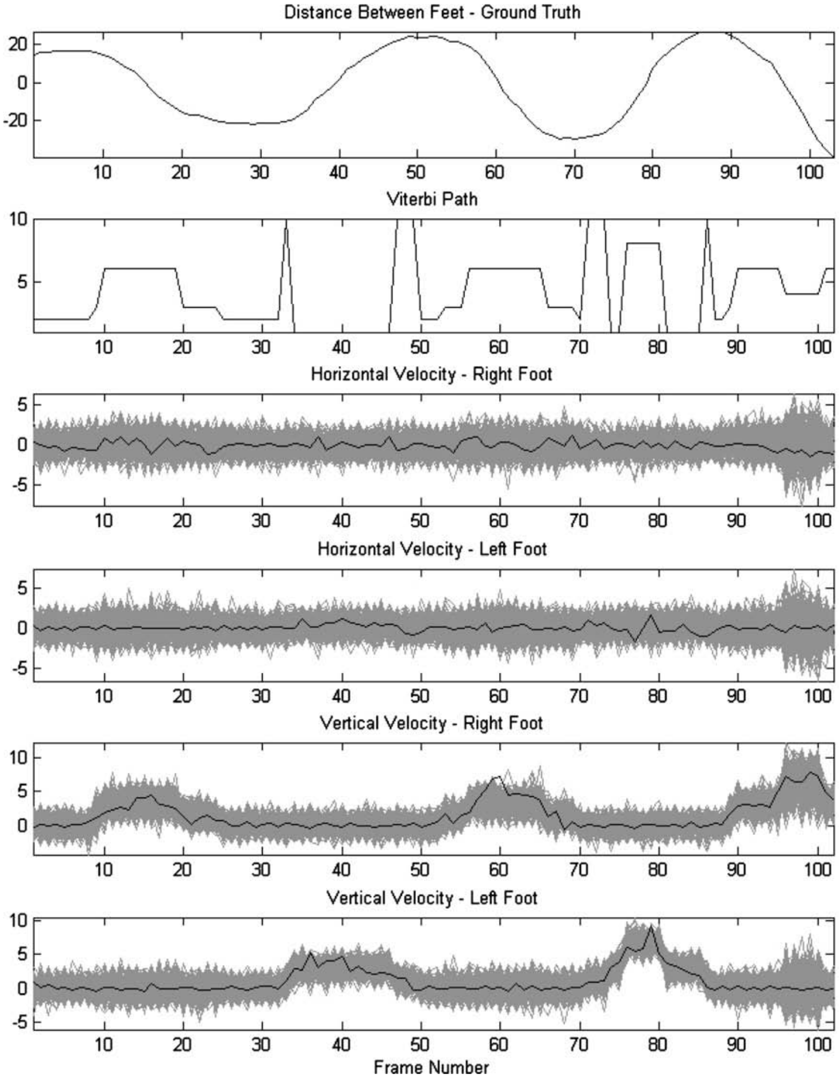

Fig. 6. This figure shows the dynamical state sequence $z_{1:t}$ and sampled translational motion $v_{1:t}$ from our dynamical model for an example stair descent. The top figure shows the ground truth distances between the feet for an example descent and the figure below it is the corresponding most likely (Viterbi) path of dynamical states z_t. The bottom four figures show the ground truth translational motion of the feet $v^*_{1:t}$ as a black line and the range of inferred motion $v_{1:t}$ at each state as sampled from the HMM (by taking 300 hundred samples at each timestep) in gray.

sequence in which there is complete occlusion of one foot, and where the multi-modal capability of the sampling technique is prominent. Fig. 9 shows the resulting tracked position of both feet through another sequence, again demonstrating the ability of the tracker to maintain a lock on the feet through occlusion.

3.2. Optical flow features

Optical flow is used as a secondary representation of the motion of a person, giving additional information about how the body of the person is moving overall. We first segment out the person from the background using an adaptive background subtraction technique, then we compute flow and take the mean over the foreground region. We found the mean flow was sufficient to give better recognition accuracy in our experiments, but that higher order moments of the flow field did not significantly improve results.

We use a simple adaptive background subtraction technique where we threshold the absolute difference between a new image at time t, $I_t(x,y)$ and a'reference image', $A(x,y)$, containing only the background. As this technique is very sensitive to changing background conditions, the reference image is updated after each frame by taking a weighted average of all previous images in a sequence, with a learning rate of α_b:

$$A_t(x,y) = (1 - \alpha_b) * A_{t-1}(x,y) + \alpha_b * I_t(x,y)$$

In our experiments, we used $\alpha_b = 0.8$ for the first 100 frames, then 0.0005 afterwards (set ad hoc to produce good background segmentation on our data). This technique suffers from a number of factors, the most significant of which are shadows and specularities. A number of methods exist which attempt to deal with these issues. These include using difference in depth from stereo information to segment the background [44], multi-component systems [45], and formulating probabilistic models of background pixels using mixtures of Gaussians [46,47]. We opted instead, due to its simplicity and good results, for a method where we perform a second round of background subtraction on the result of the initial background subtraction, but in the hue channel of the HSV color space. We found that this removed the majority of shadow pixels because the hue is less sensitive to changes in brightness than the RGB color space. Finally, we removed remaining noise by convolving with a gaussian kernel and finding the largest connected component. See Fig. 10 for an example.

Iterate

Given the old sample set $\{s_{t-1}^{(n)}, w_{t-1}^{(n)}, n = 1, ..., N\}$ at time $t-1$, construct a new sample set $\{s_t^{(n)}, w_t^{(n)}, n = 1, ..., N\}$ for time t. Construct the n^{th} of the N new samples as follows:

1. Select a sample $s_{t-1}^{\prime(n)} = (x_{t-1}^{\prime(n)}, z_t^{\prime(n)})$ by generating a random number j with probability proportional to $w_{t-1}^{(j)}$.

2. Predict by sampling from $P(X_t|X_{t-1})$ to choose $s_t^{(n)} = (x_t^{(n)}, z_t^{(n)})$:
(a) Sample transition probabilities $P(z_t^{(n)} = j|z_{t-1}^{\prime(n)}) = T_{ij}$ by generating from the transition density T_{ij} where $i = z_{t-1}^{\prime(n)}$ to find $j = z_t^{(n)}$.
(b) Sample the velocity from $p_j(v)$ for $j = z_t^{(n)}$ to give a velocity sample $v' = \{v'^l, v'^r\}$.
(c) Sample subprocess densities for each foot $p(x_t^{k(n)}|x_{t-1}^{\prime k(n)}, v'^k)$ according to $x_t^k = x_t^{\prime k(n)} + v'^k + \lambda$, where $k = l, r$.

3. Measure the observation likelihood by comparing the HOG centered at the sample's position to the reference histogram for the given foot and set the sample weight accordingly:
$w_t^{(n)} = P(y_t|X_t^f = s_t^{(n)})$

4. Normalise so that $\Sigma_n w_t^{(n)} = 1$.

Set the configuration of each foot as the mean configuration of the sample set $\{s_t^{(n)}, w_t^{(n)}, n = 1, ..., N\}$ where each sample $s_t^{(n)}$ is weighted by its corresponding weight $w_t^{(n)}$. Set the current state of the system as the mean posterior of both feet $X_t = \{X_t^l, X_t^r\}$

Fig. 7. The sampling algorithm for tracking both feet.

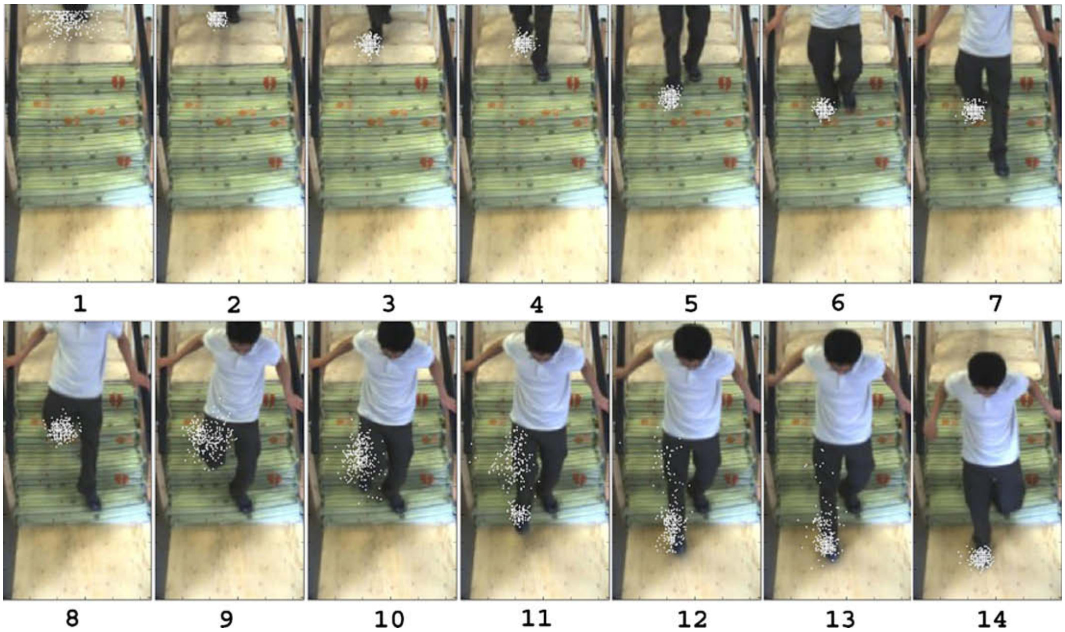

Fig. 8. A number of interesting frames from an example stair descent sequence where tracking feet is challenging due to occlusions and a large misstep. This figure shows the progression of the sample set used to track the right foot. The (*x*,*y*) location of each sample is shown as a white dot. The samples are initialized over a wide area in 1 but quickly converge on the foot in 2. In 8–10, the foot becomes completely occluded by the right knee and the sample set diverges. In 11, the sample distribution is briefly bimodal as the foot is found but eventually (12–14) all the samples return to the foot.

We compute optical flow over the foreground region using the method of Black and Anandan [8] as it is robust to multiple motions and outliers in the optical flow estimation. In this paper, we use only the mean component of the flow by averaging flow vectors over the foreground region. See Fig. 11 for examples of optical flow. Computing dense, robust, flow in order to estimate mean flow only is not strictly necessary and simpler techniques could replace this method.

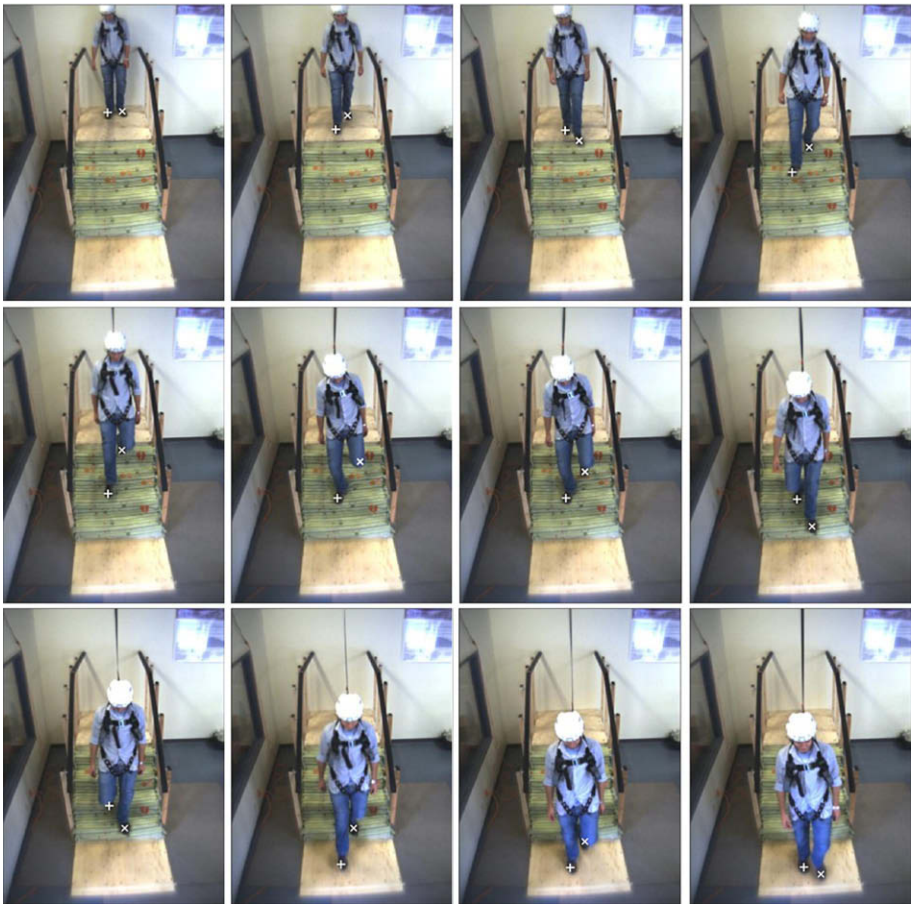

Fig. 9. A number of frames from another example stair descent where tracking feet is challenging. The resulting foot locations as tracked by our tracker are shown. The tracked location of the left foot is shown as a white "X" and the tracker location of the right foot is shown as a white "+".

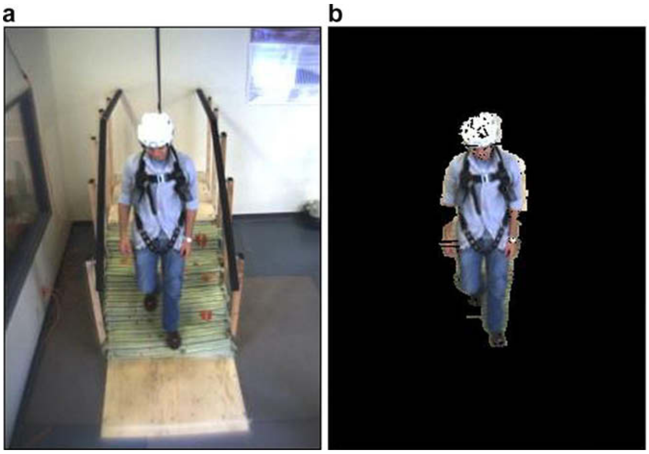

Fig. 10. An example of our background subtraction (a) is the original image and (b) is the resulting background subtracted image. *Note:* The tether attaching from the person to the overhead track has been completely removed as background. Since the tether is very thin, it is smoothed into the background as result of the Gaussian smoothing step before using connected components.

3.3. Event classification

Once we have extracted the foot positions and mean optical flow features from an image sequence containing a stair descent we can proceed to classify the descent as being normal or anomalous. Note that this procedure is distinct from the foot tracking method described in Section 3.1.2. We use the estimated mean positions of the feet as computed with the method in Fig. 7 as input features for event classification. We use eight features at each time step during a stair descent: the 2 flow features from Section 3.2 and 6 features representing the configuration of the feet. These 6 features, derived from the tracked foot positions (Section 3.1), are the horizontal and vertical instantaneous velocity of each foot and the horizontal and vertical distance between the two feet (see Fig. 12). The resulting feature vector forms a time series $\{y_{1:t}, y = \mathbb{R}^8\}$ representing a stair descent.

We use a single hidden Markov model (HMM), a probabilistic temporal model $\{\mathscr{S}, \mathbf{Y}, R, B\}$, where $\mathscr{S}$ is a finite set of states, $\mathbf{Y}$ is a continuous observation space, $R : \mathscr{S} \rightarrow \mathscr{S}$ is a transition function giving the probability of transitioning from state s at time t to state s' at time $t+1$, $T(s, s') = Pr(s' \mid s)$ and $B : \mathscr{S} \rightarrow \mathbf{Y}$ is an observation function giving the probability of observing observation feature vector $\mathbf{y}$ given state $s : B(s, \mathbf{y}) = Pr(\mathbf{y} \mid s)$. The observation function for a continuous space is parameterised using a full-covariance Gaussian mixture model:

$$P(\mathbf{y}|s = i) = \sum_j m_{ji} \mathscr{N}(\mathbf{y}; \tau_{ij}, \Lambda_{ij}) \tag{6}$$

where set of observation vectors y are generated by j Gaussians each with mixing proportion m_j, mean τ_j and covariance Λ_j. Training an HMM consists of finding the parameters $\{R, B\}$ that maximize the likelihood of a set of training data. This is done using a variant of the Expectation Maximization algorithm [48] known as the *forward–backward algorithm* or *Baum-Welch algorithm* [49]. See [41,50] for an in-depth explanation of HMMs.

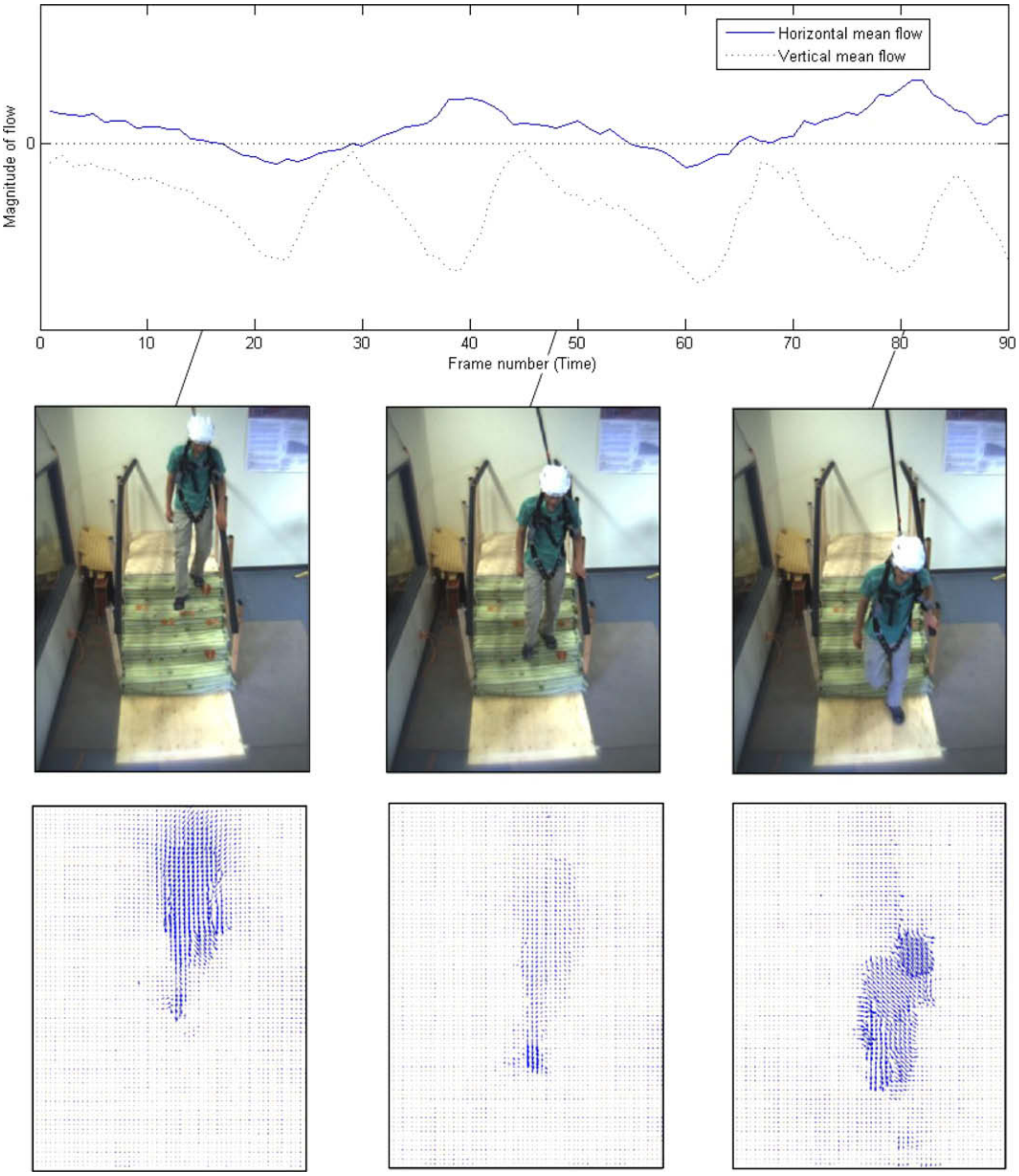

Fig. 11. Optical flow computation for an example stair descent. From top to bottom are the mean vertical and horizontal optical flow, original images and resulting optical flow.

We train the single HMM to recognise only normal walks, and then we use it to compute the likelihood of a new sequence given this normal walk model. If the sequence is, in fact, an anomalous walk, we expect the likelihood to fall below some threshold. The last step is therefore to specify this threshold. We examine a method for doing this in the next section. We train the HMM using the expectation maximization algorithm, as implemented in the BNT toolbox [51]. We used 10 hidden states with full-covariance Gaussian mixture emissions, and initialised the EM algorithm randomly. A single extra state with high covariance and low prior, mixing proportion and transition probabilities was added for regularization. The standard forward algorithm is used to evaluate the likelihood of a new sequence given a trained HMM and the likelihood is then normalized by the sequence length. The number of hidden states was chosen by evaluating the overall performance of the method when generalizing to new test subjects using different settings (see Section 4.2).

3.4. Classification methodology

This section details the classification methodology for two separate classification experiments. Due to the variability of gait across subjects it is important to validate the system by demonstrating its ability to generalize to new people. Two distinct classification experiments were conducted to test: (1) how well the system can classify new descents for people who are included in the training set and (2) how well the system can classify descents for people who are not included in the training set. While (2) is most relevant to the final system, comparing (1) and (2) will reveal the system's sensitivity to individual gait. This paper will refer to (1) as *weak generalization* and (2) as *strong generalization*. For the purposes of this system it is preferable to have false positive anomalous stair descents (normal sequences classified as being anomalous) rather than missed anomalous descents. A cost function is incorporated in the classification procedure to give a higher cost for missing anomalous descents.

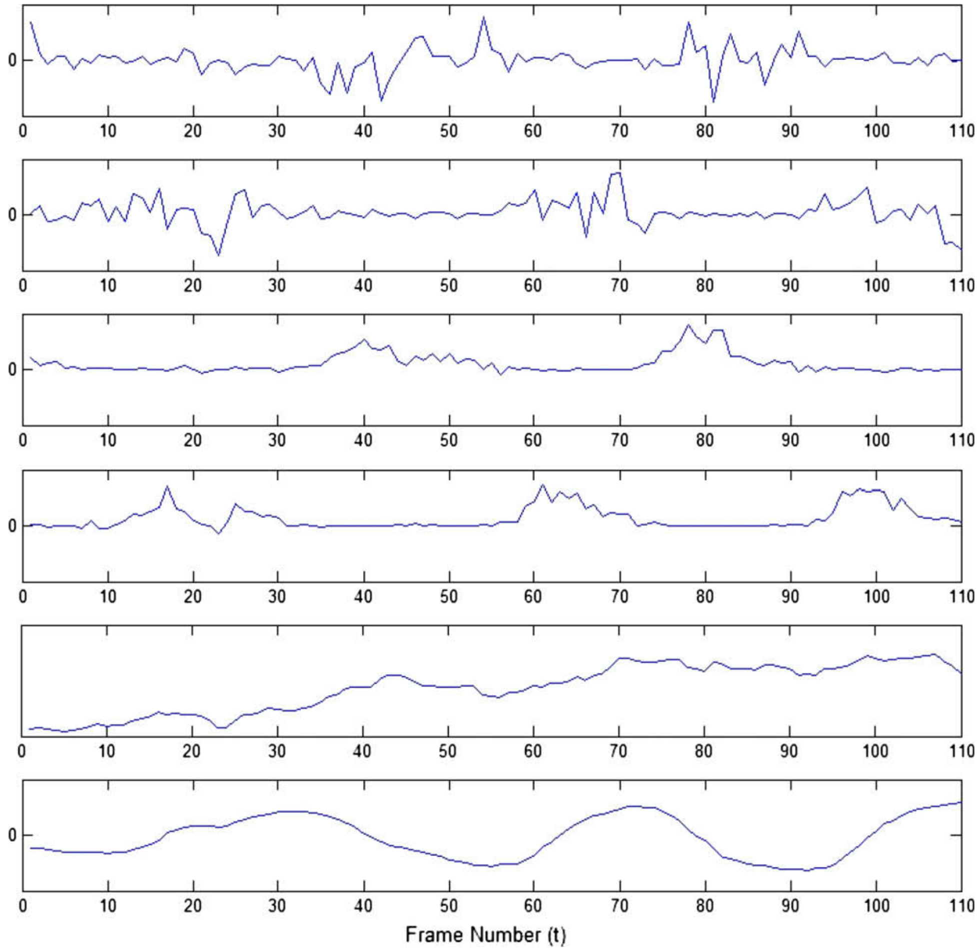

Fig. 12. Features representing the configuration of the feet. From top to bottom are the horizontal velocity of the right foot, the horizontal velocity of the left foot, the vertical velocity of the right foot, the vertical velocity of the left foot, the horizontal distance between the two feet and the vertical distance between the two feet.

For all experiments the *likelihood*, l^i, of a data sequence, $y^i_{1:t}$, refers to the log likelihood of the sequence given a hidden Markov model, θ, (as computed using the standard forward algorithm) normalized by the sequence length t:

$$l^i = \frac{1}{t}\log p(s^i_{1:t}|\theta)$$

The first step is to separate the available data into training and test sets, ensuring only that the training set contains at least some normal and some anomalous sequences. An HMM is then trained on only the training data set, and a likelihood threshold is computed that will discriminate between normal and anomalous sequences. This training and threshold determination is done using a leave-one-out cross validation procedure as follows. Each sequence, y^i, in the training data is left-out in turn, and the likelihood, l^i, of this left-out sequence is computed given a HMM trained on all remaining training data (note, however, that the HMM is only trained on the normal sequences in this training set). If the number of anomalous and normal training sequences are given by N_a and N_r, respectively, then we compute a cost for the training set that trades off the number of misses (anomalous sequences classified as normal) with the number of false positives (normal sequences classified as anomalous):

$$Cost(L) = \frac{1}{N_r}\sum_{i=1}^{N_r}\delta(l^i, L) + \frac{C_m}{N_a}\sum_{j=1}^{N_a}(1 - \delta(l^j, L)) \qquad (7)$$

where L is the likelihood threshold and $\delta(p, q)$ is the discriminant (threshold) function:

$$\delta(p,q) = \left\{ \begin{matrix} 1 & \text{if } p < q \\ 0 & \text{otherwise} \end{matrix} \right\}$$

The constant C_m gives the relative cost of a missed anomalous sequence over a false positive. In our experiments, we use $C_m = 3$: a miss is three times worse than a false positive. The threshold, L^*, is then chosen to be the value which minimizes the cost function[2]

$$L^* = \arg\min_L Cost(L)$$

This methodology is applied to both the strong and weak generalisation experiments. For the strong generalisation, the test data is all the data from a single subject. The training data is all the data from all other subjects. Each of the strong generalization experiments is repeated five times to give an average over different initialisations. The mean overall classification result of the five classification runs is reported as the overall classification rate. The mean normal and anomalous sequence classification rates over all five classifications are also reported.

When testing weak generalisation, we need to split the data into training and test sets. We do so by removing one normal sequence and half the anomalous sequences. We repeat this procedure for each normal sequence, removing a random sample of half the anomalous sequences each time. Again, this entire leave-one-out procedure is repeated five times, and the classification rate

[2] We break ties by taking the largest value of L which reflects our preference to obtain false positive anomalies rather than missed anomalies in the classification.

we report is an average over all the normal sequences and sets of anomalous sequences over all repetitions.

4. Experiments

In this section, the system is validated through conducting experiments on a carefully collected unbiased data set. We display classification results of the system and compare them to a couple of simple baseline techniques.

4.1. Data collection

In order to validate the system a data set was collected consisting of video sequences of people simulating stair descents. All subjects were adults between the ages of 18 and 30, weighed under 225lbs (a limitation of the safety harness used), had no recent musculoskeletal injuries and had no prior knowledge of the workings of the system. The subjects descended the stairs in three sets of events. In the first set, they descended the steps normally. In the second they missed a step completely (an *overstep*). In the third set they slipped off one step onto the next with one foot (a *slip*). These subtle events are the among the most common anomalous events on stairs [1]. More drastic events, such as trips and falls, are significantly less challenging to detect. Although the visual appearance of an untrained subject simulating an event on the stairs may differ from a real event, we believe that qualitatively the two events deviate similarly from normal events. Each subject was recorded by simulating each type of stair descent twenty times on an experimental staircase with four stairs in our laboratory (see Fig. 1(b)). A Point Grey Research Dragonfly 2^{TM} camera was mounted on the ceiling at a perpendicular distance of approximately 300 cm from the nosing of the center step in the stairway. The view from the camera is shown in Fig. 1. Image sequences were recorded at 30 Hz. with a resolution of 320 × 240 pixels. In order to prevent any injuries each subject wore a helmet, knee pads and a safety harness tethered to an overhead track (see Fig. 13). The experimental protocol was reviewed and approved by the University of Toronto Research Ethics Board.

The start and end of each descent sequence from each subject was manually annotated. The start of each descent was recorded as the frame where either foot starts moving at the top of the stairs. The end of each descent sequence was recorded as the first frame where both feet have cleared the stairs and either touched the ground or left the view of the camera. The subjects were informed which type of descent to perform and the descents were annotated as normal or anomalous accordingly. From the data taken, subject 1 recorded one extra anomalous event (41 anomalous events) and subject 2 recorded one less normal event (19 normal events).

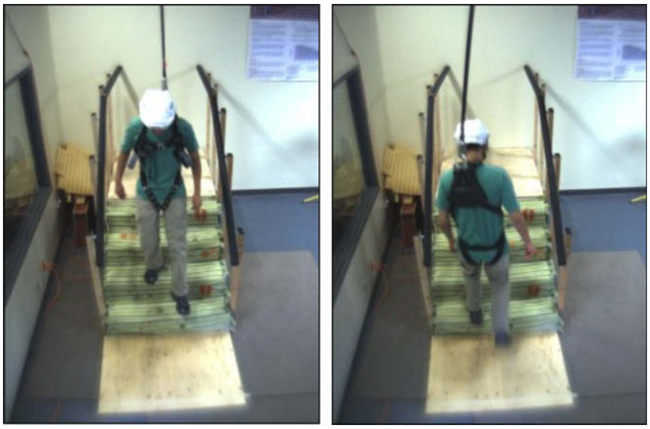

Fig. 13. This figure shows the safety gear used during data collection. Each subject wore a white hockey helmet and a safety harness tethered to an overhead track. The safety harness is connected at the subject's back to a tether. The tether connects to an overhead track. The harness and tether arrest the subject's motion in the event of a fall. The subject is wearing kneepads under his trousers.

In addition to the five person data set, another data set was recorded consisting of four subjects simulating the three types of stair descents (normal, misstep and slip) without safety equipment on the same experimental staircase. Each subject simulated each type of descent at least 15 times. Ground truth foot positions were obtained for 60 (30 normal and 30 anomalous) of these simulated descents and these were used to train the dynamical model of the foot tracker (Section 3.1.2) and used to compute the template histogram appearance models of the feet (Section 3.1.1).

4.2. Results

For all experiments we compiled a time series data set by computing the optical flow features and foot-tracking features as described in Section 3. Thus the results presented in this section reflect not only the accuracy of our HMM based classification but also that of the foot-tracking and optical flow techniques. In general, the optical flow computation was found to be very robust. The foot tracker correctly tracked the feet with high accuracy in virtually every normal stair descent but occasionally failed while tracking anomalous descents. It is far more challenging to track feet in anomalous descents because the dynamics of the feet are unpredictable and the feet move briefly at very high velocity (see Figs. 9 and 8 for examples). Even though most anomalous descents are tracked accurately, this is not of critical importance since the HMM should classify any descent as anomalous in the event that tracking fails.

Three separate types of experiments were performed. The first type of experiment tested weak generalization. This refers to testing the ability to be able to classify new examples of stair descents from a person who is represented in the training set. These experiments are labeled Ti-WEAK, where $i \in \{1, 5\}$ is the number of subjects in the training set. In the second type of experiment, STRONG, we test strong generalization across people. For this experiment we train the HMM and set the threshold using data from all subjects except one. The classification rate is then evaluated on the data from the remaining subject. It is this experiment that is the most relevant to our final system: we want to be able to flag an anomalous event occurring for a never before seen person descending the stairs. However, it is also the most difficult: the types of normal motion exhibited by the unseen person will not be modeled by the HMMs and will more often be flagged as anomalous. See Section 3.4 for a more in-depth explanation of the strong and weak generalization classification procedures. The classification results are presented in Table 1. Fig. 14 demonstrates how the cost parameter (C_m) of the likelihood threshold cost function (Eq. (7)) can be used to trade off correctly classified anomalies with false positive anomalies on the strong classification experiment.

4.3. Baselines

This method was compared to two baselines. First, classification was done based on a threshold on the number of zero-crossings of the distance between both feet in the vertical direction. This is counting the number of distinct times the two feet pass each other while descending the stairs and is related to the number of steps taken during the sequence. Sequences containing slips or oversteps should contain less distinct steps. All sequences were smoothed using a simple Gaussian filter in order to remove noise resulting from the tracking process. A threshold of five was selected, which is the threshold which gave the maximum overall classification

Table 1
HMM classification results

Exp. type	Test subj.	Missed anom. (%)	False +ve anom. (%)	Overall correct (%)
T1-WEAK	1	0.18	3.75	98.08
	2	11.3	17.89	85.53
	3	0	5.0	97.5
	4	3.5	10.0	93.25
	5	2.06	3.75	97.94
	Avg.	3.41	8.08	94.46
STRONG	1	0.08	1.0	99.02
	2	10.0	10.53	89.83
	3	1.0	13.00	95.00
	4	5.56	5.00	94.44
	5	12.5	5.00	90.00
	Avg.	5.83	6.96	93.49
T5-WEAK	1,2,3,4,5	7.28	10.1	92.72

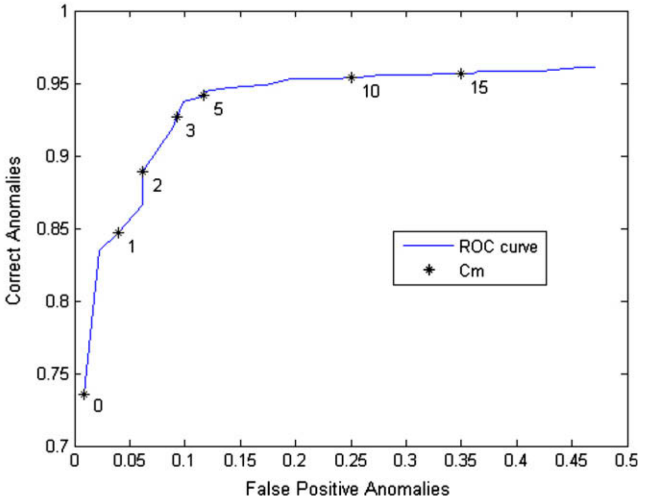

Fig. 14. This figure demonstrates how the cost parameter (C_m) of the likelihood threshold cost function (Eq. (7)) can be used to trade off correctly classified anomalies (hits) with false positive anomalies. This parameter imposes a bias in the classification procedure by adding a relative cost for missing anomalies. Plotted on this ROC curve are the percentage of correct anomalies vs. the percentage of false positive anomalies (the number of incorrectly classified normal sequences) for different settings of the cost parameter on the average strong classification result for all subjects. For the results in Table 1 $C_m = 3$ was used.

performance across all subjects. Sequences with less than or equal to five zero-crossings were classified as anomalous. The results are shown in Table 2.

The second technique is to count the zero-crossings of the derivative of the distance between the two feet in a stair descent sequence. These zero-crossings indicate when a new step starts and a previous one ends. Again the sequences were smoothed using a Gaussian filter and a threshold of five was used, which is the threshold which gave the maximum overall classification performance across all subjects. Sequences with less than or equal to five zero-crossings were classified as anomalous. See Table 3 for results.

5. Discussion

The results presented in Table 1 and Fig. 14 lead to some interesting observations. The strong classification rates for subjects 1, 3 and 4 are very promising at 99.02%, 95% and 94.44%, respectively. For subject 1 four of the five strong classification runs were 100% correct. Subjects 3 and 4 similarly had very good results. The classification accuracy for subject 2 was somewhat lower, but was actually better for the strong generalization (89.83%) than the weak generalization (85.53%). It was observed that this subject exhibited very subtle anomalous events. Particularly some of the heel slip events were difficult even for the authors to visually identify. It is believed that the subtlety of these events resulted in high likelihoods of being normal given the HMM. This can be observed in the ROC curve in Fig. 14 where even with a very strong bias for correctly classifying anomalous events approximately 4% of these events were missed.

It is possible that temporally segmenting stair events into individual stair sequences, and then attempting to identify anomalous single steps on stairs may improve this result, and forms an avenue for future research.

One issue which should be discussed is the inclusion of safety equipment in our experiments. While the use of safety equipment was necessary to protect the test subjects, it may have altered the results of our experiments. The knee pads and helmet did not likely affect our results. However, the tether between the safety harness and the overhead track noticeably altered the optical flow estimation and background subtraction. We conducted the same experiments with four test subjects from our lab, without using any safety equipment. For these subjects the average weak generalization (T1-WEAK) classification rate was 92.07% (compared to 94.46% in our results section) and the average strong generalization

Table 2
Results of classifying sequences based on the number of zero-crossings of the vertical distances between feet in each sequence (counting the number of times the feet cross)

Test subj.	Missed anom. (%)	False +ve anom. (%)	Overall (%)
1	20.0	0	86.67
2	2.5	26.32	89.83
3	20.0	5.0	93.33
4	47.5	0	68.33
5	20.0	0	86.67
Avg.	22.0	6.26	84.97

Table 3
Results of classifying sequences based on the number of zero-crossings of the derivative of the vertical distances between feet in each sequence (counting the number of steps)

Test subj.	Missed anom. (%)	False +ve anom. (%)	Overall (%)
1	17.5	20.0	81.67
2	20.0	47.37	71.19
3	22.5	10.0	81.67
4	7.5	10.0	91.67
5	27.5	15.0	76.67
Avg.	19.0	20.47	80.57

(`Strong`) classification rate was 93.42% (compared to 93.49%). The difference between these results and those presented in our results section implies that the use of safety equipment does not make a significant difference in the results.

There are a number of limitations to this system. The need to track feet as they descend the stairs presents a number of issues. First, classification is very difficult if the feet are occluded during a stair descent. This could be the case, for example, if multiple people descend the stairs simultaneously. Also, the system will flag as anomalous any sequence where the appearance model of the foot tracker fails. While this appearance model appears robust in our experiments, it is possible to think of situations where it will not accurately model the feet. This will happen whenever a subject's feet do not share the round contours of typical shoes (e.g. with bare feet, especially pointy shoes, or oddly shaped shoes such as fluffy slippers) or when the contours are not easily recognizable because the shoes share the same color as the stairs. In a situation where subjects will likely descend the stairs on bare feet it would be pertinent to create a separate appearance model to model bare feet. The system would then have to detect wether the subject is wearing shoes or is on bare feet.

The framework of our system is such that in the event that any component fails during a descent, the sequence will be flagged as anomalous (since the HMM will assign the descent a very low likelihood of being normal). This is desirable as we have a preference for generating false positive anomalous sequences as opposed to missed anomalous sequences.

Future work will focus on extending the current system so that it runs in real-time, detecting anomalous events as they occur. This will necessitate the use of a less computationally intensive optical flow estimation technique and a faster C++ based implementation of the particle filter (the current implementation is in Matlab). Classification can be performed by computing the likelihood under the classification HMM at each time step and comparing it to a threshold. Such a real-time system could be used, for example, in the home of an elderly person as part of a system to notify medical staff in the event of an emergency.

Acknowledgements

The authors thank the members of the Intelligent Assistive Technology and Systems Lab at the University of Toronto for their various contributions and David Fleet for helpful discussions. This research was funded in part by the Canadian Institutes of Health Research Institute of Aging Pilot Project Grant Program.

References

[1] J.A. Templer, The Staircase: Studies of Hazards, Falls, and Safer Design, MIT Press, Cambridge, Mass, 1994.

[2] C.P.S. Commission, National Injury Surveillance System – Online, Washington, DC, 2005.

[3] National Safety Council, Injury facts, Itasca, Illinois, 2003.

[4] National Safety Council, Accident facts, Itasca, Illinois, 1998.

[5] J. Archea, B. Collins, F. Stahl, Guidelines for stair safety, National Bureau of Standards Building Science Series, 120, Washington, DC, 1979.

[6] M. Isard, J. MacCormick, BraMBLe: A bayesian Multiple-Blob tracker, in: ICCV, 2001, pp. 34–41.

[7] T. Masud, R.O. Morris, Epidemiology of falls, Age Ageing 30 (Suppl. 4) (2001) 3–7.

[8] M.J. Black, P. Anandan, The robust estimation of multiple motions: parametric and piecewise-smooth flow fields, Computer Vision and Image Understanding 63 (1) (1996) 75–104.

[9] T. Lee, A. Mihailidis, An intelligent emergency response system: preliminary development and testing of automated fall detection, Journal of Telemedicine and Telecare 11 (4) (2005) 194–198.

[10] J. McKenna, H. Nait-Charif, Summarising Contextual Activity and Detecting Unusual Inactivity in a Supportive Home Environment, Springer-Verlag, 2004.

[11] C. Bauckhage, J.K. Tsotsos, F.E. Bunn, Detecting abnormal gait, in: CRV, 2005.

[12] J. Snoek, J. Hoey, L. Stewart, R.S. Zemel, Automated detection of unusual events on stairs, in: CRV, IEEE Computer Society, Quebec City, QC, 2006.

[13] M. Murray, Gait as a total pattern of movement, American Journal of Physical Medicine 46 (1) (1967) 290–332.

[14] J. Cutting, L. Kozlowski, Recognizing friends by their walk: gait perception without familiarity cues, Bulletin of the Physchonomic Society 0 (5) (1977) 353–356.

[15] A. Niyogi, Sourabh, H. Adelson, Edward, Analyzing and recognizing walking figures in XYT, in: IEEE Conference on Computer Vision and Pattern Recognition, no. 223, 1993, pp. 469–474.

[16] J. Little, J. Boyd, Recognizing people by their gait: the shape of motion (1996).

[17] T.B. Moeslund, E. Granum, A survey of computer vision-based human motion capture, Computer Vision and Image Understanding 81 (3) (2001) 231–268.

[18] Z. Liu, P. Grother, The humanid gait challenge problem: data sets, performance, and analysis, IEEE Transactions on Pattern Analysis and Machine Intelligence 27 (2) (2005) 162–177. member-Sudeep Sarkar and Member-P. Jonathon Phillips and Member-Isidro Robledo Vega and Fellow-Kevin W. Bowyer.

[19] D.M. Gavrila, The visual analysis of human movement: a survey, Computer Vision and Image Understanding 73 (1) (1999) 82–98.

[20] J. Fernyhough, A. Cohn, D. Hogg, Constructing qualitative event models automatically from video input, Image and Vision Computing 18 (2000) 81–103.

[21] M. Brand, V. Kettnaker, Discovery and segmentation of activities in video, IEEE Transactions on Pattern Analysis and Machine Intelligence 22 (8) (2000) 844–851.

[22] L. Liao, D. Fox, H. Kautz, Learning and inferring transportation routines, in: Proceedings of AAAI-04, 2004.

[23] R. Cutler, L. Davis, Robust real-time periodic motion detection: analysis and applications, IEEE Transactions on Pattern Analysis and Machine Intelligence 13 (2000) 129–155.

[24] A.F. Bobick, J.W. Davis, The recognition of human movement using temporal templates, IEEE Transactions on Pattern Analysis and Machine Intelligence 23 (2001) 257–267.

[25] M.A.O. Vasilescu, Human motion signatures: analysis, synthesis, recognition, ICPR'02, vol. 3, IEEE Computer Society, Washington, DC, USA, 2002, p. 30456.

[26] A. Bissacco, A. Chiuso, Y. Ma, S. Soatto, Recognition of human gaits, in: Conference on Computer Vision and Pattern Recognition, vol. 2, 2001, pp. 52–58.

[27] J.A. Hyman, Computer vision based people tracking for motivating behavior in public spaces, Master's thesis, Massachusetts Institute of Technology, 2003.

[28] T. Starner, A. Pentland, Visual recognition of american sign language using hidden markov models, in: International Workshop on Automatic Face and Gesture Recognition, 1995, pp. 189–194.

[29] M. Brand, N. Oliver, A. Pentland, Coupled hidden markov models for complex action recognition, in: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition 1997, 1997, pp. 994–999.

[30] L. Campbell, D. Becker, A. Azarbayejani, A. Bobick, A. Pentland, Invariant features for 3-d gesture recognition, in: Proceedings of FG'96, 1996, pp. 157–162.

[31] C. Vogler, H. Sun, D.N. Metaxas, A framework for motion recognition with applications to american sign language and gait recognition, in: Workshop on Human Motion, 2000, pp. 33–38.

[32] A. Kapoor, R. Picard, A real-time head nod and shake detector, in: Proceedings from the Workshop on Perspective User Interfaces, 2001.

[33] M. Isard, A. Blake, Condensation – conditional density propagation for visual tracking, International Journal of Computer Vision 29 (1998) 5–28.

[34] Z. Khan, T. Balch, F. Dellaert, A rao-blackwellized particle filter for eigentracking, in: CVPR, 2004.

[35] K. Okuma, A. Taleghani, N. de Freitas, J. Little, D. Lowe, A boosted particle filter: multitarget detection and tracking, in: ECCV, Springer, 2004, pp. 28–39.

[36] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: C. Schmid, S. Soatto, C. Tomasi (Eds.), International Conference on Computer Vision and Pattern Recognition, vol. 2, INRIA Rhône-Alpes, ZIRST-655, av. de l'Europe, Montbonnot-38334, 2005, pp. 886–893.

[37] D. Comaniciu, V. Ramesh, P. Meer, Real-time tracking of non-rigid objects using mean shift, in: CVPR, 2000, pp. 142–151.

[38] P. Perez, C. Hue, J. Vermaak, M. Gangnet, Color-based probabilistic tracking, in: ECCV'02: Proceedings of the Seventh European Conference on Computer Vision. Part I, Springer-Verlag, London, UK, 2002, pp. 661–675.

[39] M. Isard, A. Blake, A mixed-state CONDENSATION tracker with automatic model-switching, in: ICCV, 1998, pp. 107–112.

[40] B. North, A. Blake, M. Isard, J. Rittscher, Learning and classification of complex dynamics, IEEE Transactions on Pattern Analysis and Machine Intelligence 22 (9) (2000) 1016–1034.

[41] L.R. Rabiner, A tutorial on hidden markov models and selected applications in speech recognition, in: Readings in Speech Recognition, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1990, pp. 267–296.

[42] N.J. Gordon, D.J. Salmond, A.F.M. Smith, A novel approach to nonlinear and non-gaussian bayesian state estimation, IEEE Proceedings of Radar and Signal Processing 140 (1993) 107–113.

[43] S. Arulampalam, S. Maskell, N. Gordon, T. Clapp, A tutorial on particle filters for on-line non-linear/non-gaussian bayesian tracking, IEEE Transactions on Signal Processing 50 (2) (2002) 174–188.

[44] Y.A. Ivanov, A.F. Bobick, J. Liu, Fast lighting independent background subtraction, International Journal of Computer Vision 37 (2) (2000) 199–207.

[45] K. Toyama, J. Krumm, B. Brumitt, B. Meyers, Wallflower: Principles and practice of background maintenance, in: ICCV, issue No. 1, 1999, pp. 255–261.

[46] C. Stauffer, Adaptive background mixture models for real-time tracking, in: IEEE Conference on Computer Vision and Pattern Recognition, 1999, pp. 246–252.

[47] C.R. Wren, A. Azarbayejani, T. Darrell, A. Pentland, Pfinder: real-time tracking of the human body, IEEE Transactions on Pattern Analysis and Machine Intelligence 19 (7) (1997) 780–785.

[48] A.P. Dempster, N.M. Laird, D.B. Rubin, Maximum likelihood from incomplete data via the em algorithm, Journal of the Royal Statistical Society, Series B 39 (1977) 138.

[49] L. Baum, An inequality and associated maximization technique in statistical estimation of probabilistic functions of markov processes, Inequalities 3 (1972) 1–8.

[50] Y. Bengio, Markovian models for sequential data, Tech. rep., University of Montreal, 1996.

[51] K.P. Murphy, The bayes net toolbox for MATLAB, in: Computing Science and Statistics, vol. 33.