Exploiting Dimensional Reduction in Modelling of High Dimensional Distributions

Neil Lawrence
Machine Learning Group
School of Computer Science
University of Manchester, U.K.

8th December 2007

Outline

1. Motivation
   - High Dimensional Data
   - Examples
   - Hierarchical GP-LVM

2. Conclusions
   - Summary

Online Resources

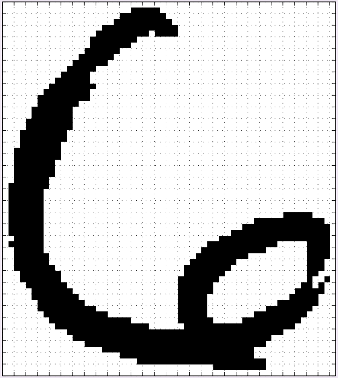
All source code and slides are available online

- This talk available from my home page (see talks link on left hand side).
- MATLAB examples in the 'oxford' toolbox (vrs 0.131), `demGplvmTalk`.
  - `http://www.cs.man.ac.uk/~neill/oxford/`.
- And the 'fgplvm' toolbox (vrs 0.15).
  - `http://www.cs.man.ac.uk/~neill/fgplvm/`.
- MATLAB commands used for examples given in `typewriter font`.

High Dimensional Data

USPS Data Set Handwritten Digit

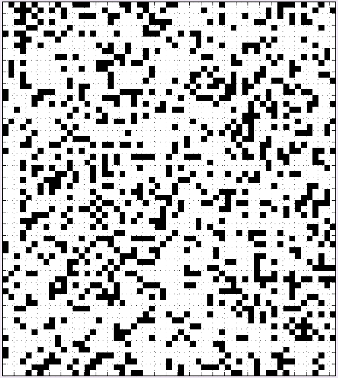
- 3648 Dimensions
  - 64 rows by 57 columns



High Dimensional Data

USPS Data Set Handwritten Digit

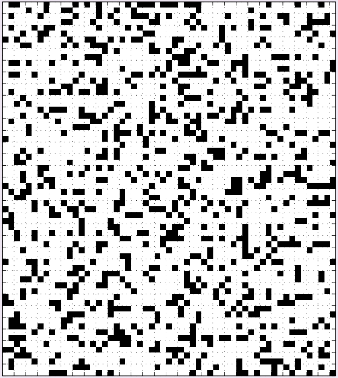
- 3648 Dimensions
  - 64 rows by 57 columns
- Space contains more than just this digit.



High Dimensional Data

USPS Data Set Handwritten Digit

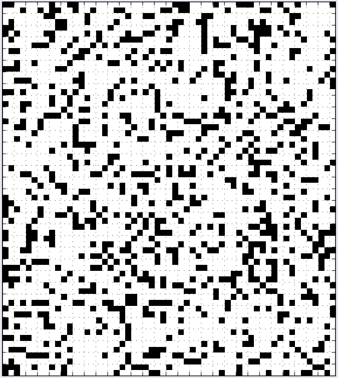
- 3648 Dimensions
  - 64 rows by 57 columns
- Space contains more than just this digit.
- Even if we sample every nanosecond from now until the end of the universe, you won't see the original six!



High Dimensional Data

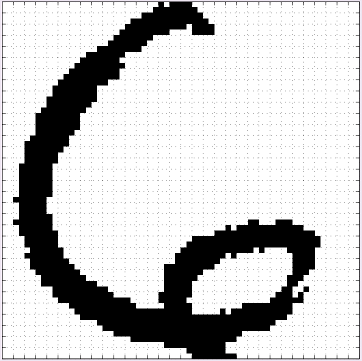
USPS Data Set Handwritten Digit

- 3648 Dimensions
  - 64 rows by 57 columns
- Space contains more than just this digit.
- Even if we sample every nanosecond from now until the end of the universe, you won't see the original six!



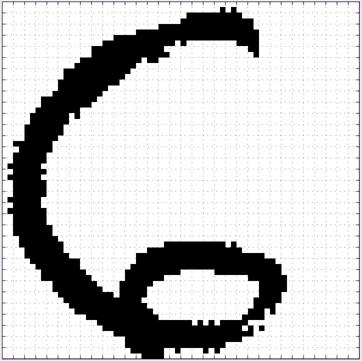
Simple Model of Digit

Rotate a 'Prototype'



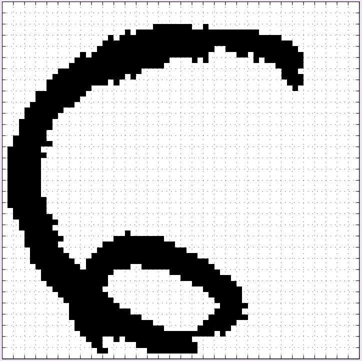
Simple Model of Digit

Rotate a 'Prototype'



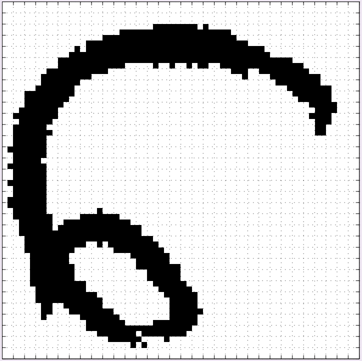
Simple Model of Digit

Rotate a 'Prototype'



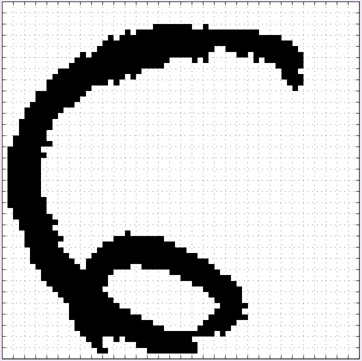
Simple Model of Digit

Rotate a 'Prototype'



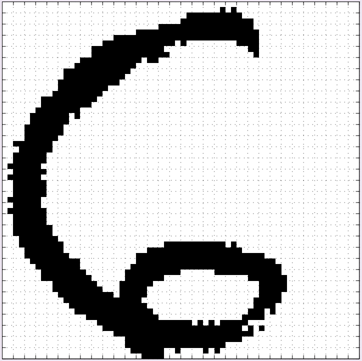
Simple Model of Digit

Rotate a 'Prototype'



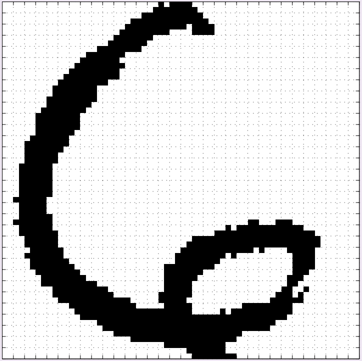
Simple Model of Digit

Rotate a 'Prototype'



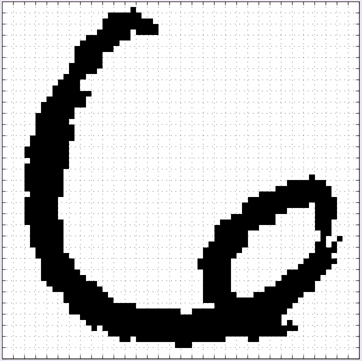
Simple Model of Digit

Rotate a 'Prototype'



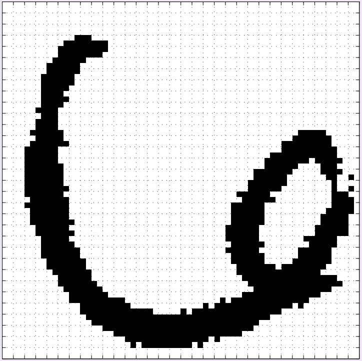
Simple Model of Digit

Rotate a 'Prototype'



Simple Model of Digit

Rotate a 'Prototype'



Low Dimensional Manifolds

Pure Rotation is too Simple

- In practice the data may undergo several distortions.
  - *e.g.* digits undergo 'thinning', translation and rotation.
- For data with 'structure':
  - we expect fewer distortions than dimensions;
  - we therefore expect the data to live on a lower dimensional manifold.
- Conclusion: deal with high dimensional data by looking for lower dimensional non-linear embedding.

GP-LVM

- *Probabilistic* non-linear generalisation of PCA.
- 'Kernelises' in opposite direction to Kernel PCA.

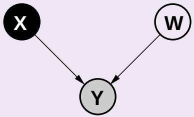
Notation — $\mathbf{X}$ and $\mathbf{Y}$ are *design matrices*

- Covariance given by $n^{-1}\mathbf{Y}^{\mathrm{T}}\mathbf{Y}$.
- Inner product matrix given by $\mathbf{Y}\mathbf{Y}^{\mathrm{T}}$.

Linear Latent Variable Model

Dual Probabilistic PCA

- Define *linear-Gaussian relationship* between latent variables and data.

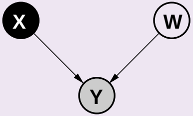


$$p\left(\mathbf{Y}|\mathbf{X},\mathbf{W}\right)=\prod_{i=1}^{n}N\left(\mathbf{y}_{i,:}|\mathbf{W}\mathbf{x}_{i,:},\sigma^{2}\mathbf{I}\right)$$

Linear Latent Variable Model

Dual Probabilistic PCA

- Define *linear-Gaussian relationship* between latent variables and data.
- **Novel** Latent variable approach:

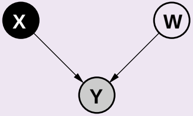


$$p(\mathbf{Y}|\mathbf{X},\mathbf{W}) = \prod_{i=1}^{n} N\left(\mathbf{y}_{i,:}|\mathbf{W}\mathbf{x}_{i,:},\sigma^2\mathbf{I}\right)$$

Linear Latent Variable Model

Dual Probabilistic PCA

- Define *linear-Gaussian relationship* between latent variables and data.
- **Novel** Latent variable approach:
  - Define Gaussian prior over *parameters*, $\mathbf{W}$.



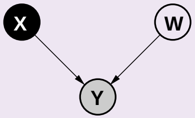
$$p(\mathbf{Y}|\mathbf{X},\mathbf{W}) = \prod_{i=1}^{n} N\left(\mathbf{y}_{i,:}|\mathbf{W}\mathbf{x}_{i,:},\sigma^2\mathbf{I}\right)$$

$$p(\mathbf{W}) = \prod_{i=1}^{d} N\left(\mathbf{w}_{i,:}|\mathbf{0},\mathbf{I}\right)$$

Linear Latent Variable Model

Dual Probabilistic PCA

- Define *linear-Gaussian relationship* between latent variables and data.
- **Novel** Latent variable approach:
  - Define Gaussian prior over *parameters*, $\mathbf{W}$.
  - Integrate out *parameters*.



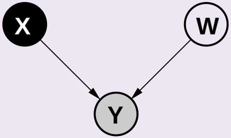
$$p\left(\mathbf{Y}|\mathbf{X},\mathbf{W}\right) = \prod_{i=1}^{n} N\left(\mathbf{y}_{i,:}|\mathbf{W}\mathbf{x}_{i,:},\sigma^2\mathbf{I}\right)$$

$$p\left(\mathbf{W}\right) = \prod_{i=1}^{d} N\left(\mathbf{w}_{i,:}|\mathbf{0},\mathbf{I}\right)$$

$$p\left(\mathbf{Y}|\mathbf{X}\right) = \prod_{j=1}^{d} N\left(\mathbf{y}_{:,j}|\mathbf{0},\mathbf{X}\mathbf{X}^{\top}+\sigma^2\mathbf{I}\right)$$

Linear Latent Variable Model II

Dual Probabilistic PCA Max. Likelihood Soln [Lawrence, 2004]



$$p\left(\mathbf{Y}|\mathbf{X}\right) = \prod_{j=1}^{d} N\left(\mathbf{y}_{:,j}|\mathbf{0}, \mathbf{X}\mathbf{X}^{\top} + \sigma^2\mathbf{I}\right)$$

Linear Latent Variable Model II

Dual Probabilistic PCA Max. Likelihood Soln [Lawrence, 2004]

$$p(\mathbf{Y}|\mathbf{X}) = \prod_{j=1}^{d} N\left(\mathbf{y}_{:,j}|\mathbf{0}, \mathbf{K}\right), \quad \mathbf{K} = \mathbf{X}\mathbf{X}^{\top} + \sigma^2\mathbf{I}$$

$$\log p(\mathbf{Y}|\mathbf{X}) = -\frac{d}{2}\log|\mathbf{K}| - \frac{1}{2}\mathrm{tr}\left(\mathbf{K}^{-1}\mathbf{Y}\mathbf{Y}^{\top}\right) + \text{const.}$$

If $\mathbf{U}_q'$ are first q principal eigenvectors of $d^{-1}\mathbf{Y}\mathbf{Y}^{\top}$ and the corresponding eigenvalues are Λ_q,

$$\mathbf{X} = \mathbf{U'}_q\mathbf{L}\mathbf{V}^{\top}, \quad \mathbf{L} = \left(\Lambda_q - \sigma^2\mathbf{I}\right)^{\frac{1}{2}}$$

where $\mathbf{V}$ is an arbitrary rotation matrix.

Linear Latent Variable Model II

Probabilistic PCA Max. Likelihood Soln [Tipping and Bishop, 1999]

$$p\left(\mathbf{Y}|\mathbf{W}\right) = \prod_{i=1}^{n} N\left(\mathbf{y}_{i,:}|\mathbf{0}, \mathbf{C}\right), \quad \mathbf{C} = \mathbf{W}\mathbf{W}^{\top} + \sigma^2\mathbf{I}$$

$$\log p\left(\mathbf{Y}|\mathbf{W}\right) = -\frac{n}{2}\log|\mathbf{C}| - \frac{1}{2}\mathrm{tr}\left(\mathbf{C}^{-1}\mathbf{Y}^{\top}\mathbf{Y}\right) + \text{const.}$$

If $\mathbf{U}_q$ are first q principal eigenvectors of $n^{-1}\mathbf{Y}^{\top}\mathbf{Y}$ and the corresponding eigenvalues are Λ_q,

$$\mathbf{W} = \mathbf{U}_q\mathbf{L}\mathbf{V}^{\top}, \quad \mathbf{L} = \left(\Lambda_q - \sigma^2\mathbf{I}\right)^{\frac{1}{2}}$$

where $\mathbf{V}$ is an arbitrary rotation matrix.

Equivalence of Formulations

The Eigenvalue Problems are equivalent

- Solution for Probabilistic PCA (solves for the mapping)

$$\mathbf{Y}^{\top}\mathbf{Y}\mathbf{U}_q = \mathbf{U}_q\Lambda_q \qquad \mathbf{W} = \mathbf{U}_q\mathbf{L}\mathbf{V}^{\top}$$

- Solution for Dual Probabilistic PCA (solves for the latent positions)

$$\mathbf{Y}\mathbf{Y}^{\top}\mathbf{U}'_q = \mathbf{U}'_q\Lambda_q \qquad \mathbf{X} = \mathbf{U}'_q\mathbf{L}\mathbf{V}^{\top}$$

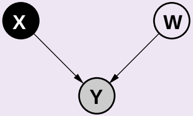
- Equivalence is from

$$\mathbf{U}_q = \mathbf{Y}^{\top}\mathbf{U}'_q\Lambda_q^{-\frac{1}{2}}$$

Non-Linear Latent Variable Model

Dual Probabilistic PCA

- Define *linear-Gaussian relationship* between latent variables and data.
- **Novel** Latent variable approach:
  - Define Gaussian prior over *parameteters*, $\mathbf{W}$.
  - Integrate out *parameters*.



$$p(\mathbf{Y}|\mathbf{X},\mathbf{W}) = \prod_{i=1}^{n} N\left(\mathbf{y}_{i,:}|\mathbf{W}\mathbf{x}_{i,:},\sigma^2\mathbf{I}\right)$$

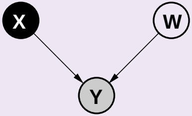
$$p(\mathbf{W}) = \prod_{i=1}^{d} N\left(\mathbf{w}_{i,:}|\mathbf{0},\mathbf{I}\right)$$

$$p(\mathbf{Y}|\mathbf{X}) = \prod_{j=1}^{d} N\left(\mathbf{y}_{:,j}|\mathbf{0},\mathbf{X}\mathbf{X}^{\top}+\sigma^2\mathbf{I}\right)$$

Non-Linear Latent Variable Model

Dual Probabilistic PCA

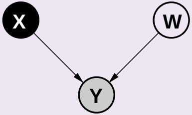
- Inspection of the marginal likelihood shows ...



$$p\left(\mathbf{Y}|\mathbf{X}\right) = \prod_{j=1}^{d} N\left(\mathbf{y}_{:,j}|\mathbf{0}, \mathbf{X}\mathbf{X}^{\top} + \sigma^2\mathbf{I}\right)$$

Dual Probabilistic PCA

- Inspection of the marginal likelihood shows ...
  - The covariance matrix is a covariance function.



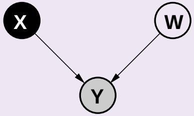
$$p\left(\mathbf{Y}|\mathbf{X}\right) = \prod_{j=1}^{d} N\left(\mathbf{y}_{:,j}|\mathbf{0},\mathbf{K}\right)$$

$$\mathbf{K} = \mathbf{X}\mathbf{X}^{\mathrm{T}} + \sigma^2\mathbf{I}$$

Non-Linear Latent Variable Model

Dual Probabilistic PCA

- Inspection of the marginal likelihood shows ...
  - The covariance matrix is a covariance function.
  - We recognise it as the 'linear kernel'.



$$p(\mathbf{Y}|\mathbf{X}) = \prod_{j=1}^{d} N\left(\mathbf{y}_{:,j}|\mathbf{0}, \mathbf{K}\right)$$

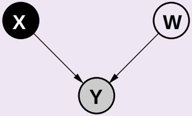
$$\mathbf{K} = \mathbf{X}\mathbf{X}^{\top} + \sigma^2\mathbf{I}$$

This is a product of Gaussian processes with linear kernels.

Non-Linear Latent Variable Model

Dual Probabilistic PCA

- Inspection of the marginal likelihood shows ...
  - The covariance matrix is a covariance function.
  - We recognise it as the 'linear kernel'.
  - We call this the Gaussian Process Latent Variable model (GP-LVM).



$$p(\mathbf{Y}|\mathbf{X}) = \prod_{j=1}^{d} N\left(\mathbf{y}_{:,j}|\mathbf{0}, \mathbf{K}\right)$$

$$\mathbf{K} = ?$$

Replace linear kernel with non-linear kernel for non-linear model.

Non-linear Latent Variable Models

RBF Kernel

- The RBF kernel has the form $k_{ij} = k\left(\mathbf{x}_{i,:}, \mathbf{x}_{j,:}\right)$, where

$$k\left(\mathbf{x}_{i,:}, \mathbf{x}_{j,:}\right) = \alpha \exp\left(-\frac{\left(\mathbf{x}_{i,:} - \mathbf{x}_{j,:}\right)^{\top}\left(\mathbf{x}_{i,:} - \mathbf{x}_{j,:}\right)}{2l^2}\right).$$

- No longer possible to optimise wrt $\mathbf{X}$ via an eigenvalue problem.
- Instead find gradients with respect to $\mathbf{X}, \alpha, l$ and σ^2 and optimise using conjugate gradients.

Stick Man

Generalization with less Data than Dimensions

- Powerful uncertainly handling of GPs leads to suprising properties.
- Non-linear models can be used where there are fewer data points than dimensions *without overfitting*.
- Example: Modelling a stick man in 102 dimensions with 55 data points!

Stick Man II

`demStick1`

Figure: The latent space for the stick man motion capture data.

Stick Man II

demStick1

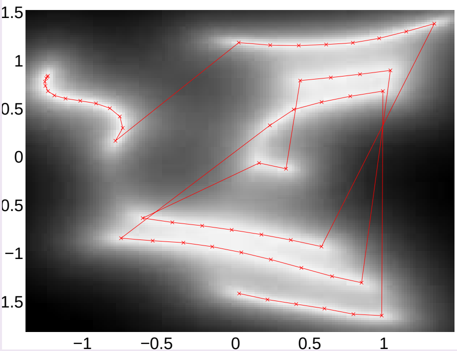
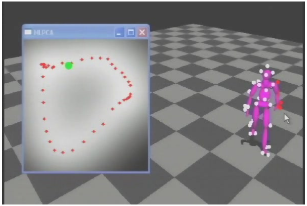
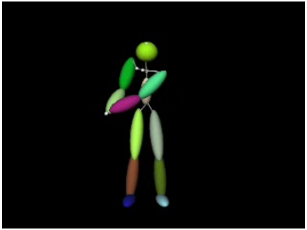


Figure: The latent space for the stick man motion capture data.

Applications

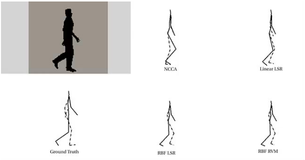


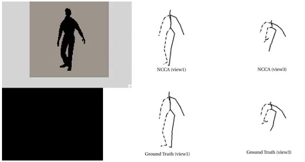
Facilitating animation through modelling human motion with the GP-LVM [Grochow et al., 2004]



Tracking using models of human motion learnt with the GP-LVM [Urtasun et al., 2005, 2006]

Applications





Hierarchical GP-LVM

Stacking Gaussian Processes

- Regressive dynamics provides a simple hierarchy.
  - The input space of the GP is governed by another GP.
- By stacking GPs we can consider more complex hierarchies.
- Ideally we should marginalise latent spaces
  - In practice we seek MAP solutions.

Two Correlated Subjects

demHighFive1

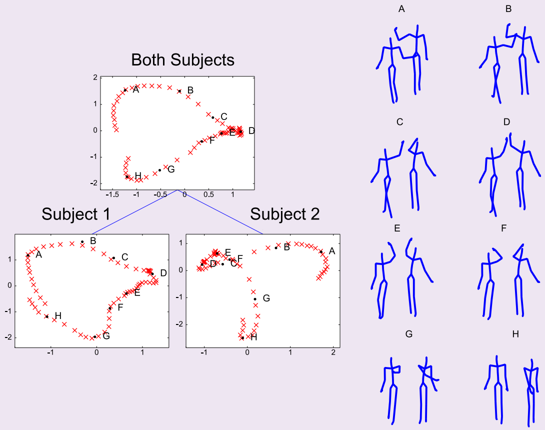


Figure: Hierarchical model of a 'high five'.

Within Subject Hierarchy

Decomposition of Body

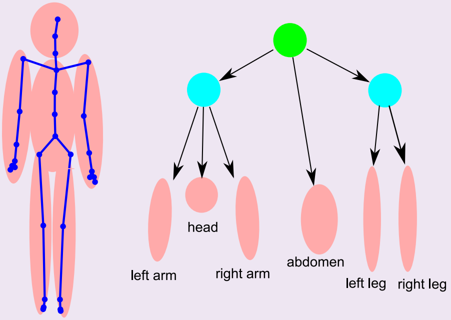


Figure: Decomposition of a subject.

demRunWalk1

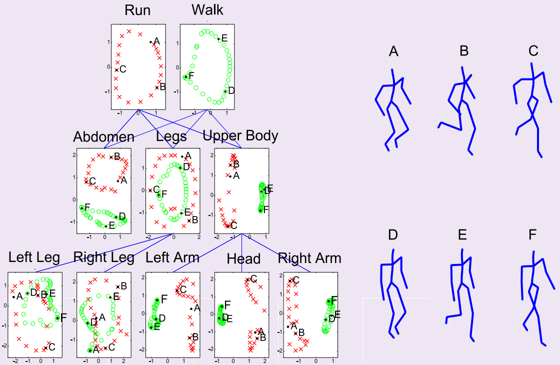


Figure: Hierarchical model of a walk and a run.

Summary

- GP-LVM is a Probabilistic Non-Linear Generalisation of PCA.
- Works Effectively as a Probabilistic Model in High Dimensional Spaces.
- Also need conditional independencies to have a truly general model.
    - Here they were hard coded.
    - They should be learned!!

References

K. Grochow, S. L. Martin, A. Hertzmann, and Z. Popovic. Style-based inverse kinematics. In *ACM Transactions on Graphics (SIGGRAPH 2004)*, pages 522–531, 2004. doi: 10.1145/1186562.1015755.

N. D. Lawrence. Gaussian process models for visualisation of high dimensional data. In S. Thrun, L. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems*, volume 16, pages 329–336, Cambridge, MA, 2004. MIT Press.

M. E. Tipping and C. M. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society, B*, 6(3): 611–622, 1999.

R. Urtasun, D. J. Fleet, A. Hertzmann, and P. Fua. Priors for people tracking from small training sets. In *IEEE International Conference on Computer Vision (ICCV)*, pages 403–410, Bejing, China, 17–21 Oct. 2005. IEEE Computer Society Press.

R. Urtasun, D. J. Fleet, and P. Fua. 3D people tracking with Gaussian process dynamical models. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pages 238–245, New York, U.S.A., 17–22 Jun. 2006. IEEE Computer Society Press.