Object Tracking Using Globally Coordinated Nonlinear Manifolds

Che-Bin Liu[1,2] Ruei-Sung Lin[1,3] Ming-Hsuan Yang[4] Narendra Ahuja[1] Stephen Levinson[1]
[1]University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA
[2]Epson R&D, Inc., Palo Alto, CA 94304, USA
[3]Motorola Labs, Schaumburg, IL 60196, USA
[4]Honda Research Institute, Mountain View, CA 94041, USA

Abstract

We present a dynamic inference algorithm in a globally parameterized nonlinear manifold and demonstrate it on the problem of visual tracking. An appearance manifold is usually nonlinear, embedded in a high dimensional space, and can be approximated by a mixture of locally linear models. Existing methods for nonlinear dimensionality reduction, which map an appearance manifold to a single low dimensional coordinate system, preserve only spatial relationships among manifold points and render low dimensional embeddings rather than mapping functions. In this paper, we parameterize the mixture of linear appearance subspaces of an object in a global coordinate system, and apply it to visual tracking using a Rao-Blackwellized particle filter. Experimental results demonstrate that the proposed approach performs well on object tracking problem in scenes with significant clutter and temporary occlusions which pose difficulties for other methods.

1. Introduction

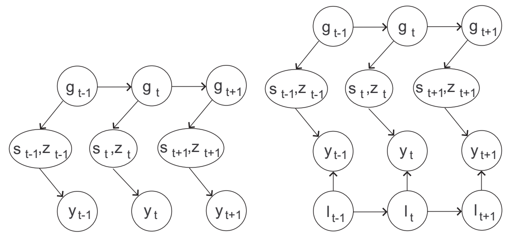
For many tracking problems, the intrinsic parameters of the tracked object gradually change over time. As a result, expanding trackers to take into account this additional information will greatly improve tracking performance. Take the appearance-based face tracking for example. It is well known that the appearance of the tracked face undergoes smooth variation across video frames. If the face location and its appearance variation are simultaneously tracked, the smoothness constraint on the appearance variation can be a strong cue to prune out implausible candidates and to maintain a reliable and consistent tracking result. This is especially helpful while tracking a face in a cluttered background where there are many false candidates (i.e. other faces) in the surroundings.

The intrinsic parameters of object images are low dimensional vectors embedded in the high dimensional image space. Usually, these parameters are embedded in a nonlinear manifold, and recovering these parameters from the images is a manifold learning problem. Manifold embedding methods, such as Isomap [8] and LLE [5], map high dimensional data to a low dimensional global coordinates while preserving local geometric relationships among data samples. Experiments have shown that these global coordinates represent the intrinsic parameters of the data. Nevertheless, the main drawback is that these methods find lower dimensional embeddings rather than mapping functions between sample points and manifolds; that is, these methods can not map out-of-sample data points to globally coordinated nonlinear manifolds. For applications such as object tracking, parameterizing the nonlinear manifold and acquiring mapping functions are both critical.

Some attempts have been made to parameterize nonlinear manifolds using mixtures of subspace models. Wang et. al. [9] find the mapping function using mixture of factor analyzers and use the Isomap result to learn the model by solving a regression problem. Their mapping function is therefore determined by the Isomap results. Teh and Roweis [7] first compute a mixture of subspace models and then align these locally coordinated mixture parameters within a global coordinate by imposing a locally linear constraint as used in LLE. This method produces an exact nonlinear mapping function, and it has shown impressive results for modeling static data. Our work extends this method to model a dynamic process, i.e., tracking intrinsic parameters on the nonlinear manifold of object appearance. It is a challenging task that has not been well exploited.

The idea of simultaneously tracking location and appearance parameters has been proposed by [3]. In their work, they assume the appearance parameters are embedded on a linear manifold. Hence, their tracker is based on a linear Kalman filter. The linear property also enables them to apply Rao-Blackwellized particle filtering for efficient tracking. Our work extends such idea by tracking the appearance parameters on a nonlinear manifold. We present a dynamic inference algorithm for visual tracking in which

Proceedings of the 18th International Conference on Pattern Recognition (ICPR'06)
0-7695-2521-0/06 $20.00 © 2006 IEEE



(a) Our DBN. (b) Our complete graphical model for object tracking.

Figure 1. Our graphical models.

an observed data sequence is modeled by a continuous and smooth trajectory on a low dimensional nonlinear manifold. By imposing the temporal dependency constraints among the points on the mapped trajectory in a global coordinate system, we develop an efficient and effective inference algorithm within a dynamic Bayesian network (DBN). The particular structure of our model provides means for an efficient approximate inference algorithm.

2. Dynamic Inference using Global Coordinates

Teh and Roweis [7] present an algorithm that learns a mapping from observed data to global coordinates of a nonlinear manifold. With such mapping, we now present our dynamic model based on global coordinates. Our model is grounded on the assumption that the dynamics of an observed data sequence can be modeled well by points moving slowly along a smooth trajectory on the manifold. This property has been observed in many nonlinear embedding experiments [5, 6, 7, 8] and is validated by our experiments.

In our model, the temporal dependency between global coordinates is assumed to be Markovian, $P(g_t|g_{t-1},\ldots,g_1) = P(g_t|g_{t-1})$. That is, the global coordinates g_t are only determined by its previous global coordinates g_{t-1}. Figure 1(a) illustrates our model as a dynamic Bayesian network. Our graphical model is structurally similar to a linear dynamical system (LDS), but our mapping between observed data and global coordinates is through latent variables (s, z_s) and is therefore nonlinear. Since these latent variables are temporally independent, we marginalize them out at each time step and present an efficient approximate inference for this dynamical model.

Our model draws an interesting property when compared with the switching state-space models [2, 4, 9]. In these models, the state transition probability is assumed to be independent from the continuous latent variables, and it has been pointed out that this assumption contradicts to the actual underlying dynamic process [1]. On the contrary, our model does not have the state transition problem even though our model also contains a discrete state variable. This is because our model tracks the globally coordinated state variables where the actual dynamics are well preserved.

In dynamic inference, we aim to estimate the posterior distribution $P(g_t|y_{1:t})$ as time progresses. Although our model has more latent variables than a standard Bayes filter, latent variables (s, z_s) can be marginalized out at each time step so $P(g_t|y_{1:t})$ can be estimated recursively using a Bayes filter:

$$P(g_t|y_{1:t}) = kP(y_t|g_t)\int P(g_t|g_{t-1})P(g_{t-1}|y_{1:t-1})dg_{t-1}. \tag{1}$$

Given that g_t is assumed to be a slowly changing variable, we model the dynamics of g_t using Brownian motion: $P(g_t|g_{t-1}) \sim \mathcal{N}(g_{t-1}, Q)$. Furthermore, since we have the mapping from $P(g_t|y_t)$, we can also compute $P(y_t|g_t)$ using the Bayes rule: $P(y_t|g_t) = P(g_t|y_t)P(y_t)/P(g_t)$, and thus obtain a mapping function between y and g.

2.1 Approximate Inference

The recursive Bayes-estimated inference on our graphical model is intractable because $P(g|y)$ is a mixture of Gaussians. That is, the number of model parameters in the distribution $P(g_t|y_{1:t})$ will grow exponentially with time. Therefore, we propose an approximate inference to overcome this problem.

Ideally, the nonlinear mapping between y and g should be one-to-one, so the distribution of $P(g|y)$ should be unimodal at any given time. Hence, at each time step, we dynamically approximate distribution $P(g|y)$, that consists of a mixtures of Gaussians, to a single Gaussian. We denote (μ_t, Σ_t) as the mean and the covariance matrix of the Gaussian that we use to approximate $P(g|y)$, and denote (μ_t^s, Σ_t^s) as the mean and the covariance matrix for Gaussian distribution $P(g|y, s)$ of a local linear model s. So the inference of global coordinate g conditioned on observation y can be written as:

$$P(g|y) = \sum_s \mathcal{N}(y; \mu_t^s, \Sigma_t^s)P(s|y). \tag{2}$$

To ensure $\mathcal{N}(\mu_t, \Sigma_t)$ is close to $P(g|y)$, we measure the weighted KL-distance:

$$(\mu_t, \Sigma_t) = \arg\min_{\mu,\Sigma} \sum_s P(s_t|y_t)KL(\mathcal{N}(\mu, \Sigma)||\mathcal{N}(\mu_t^s, \Sigma_t^s)). \tag{3}$$

And the solution of (μ_t, Σ_t) can be derived as

$$\begin{aligned}\mu_t &= \sum_s P(s_t|y_t)\mu_t^s, \\ \Sigma_t &= \sum_s P(s_t|y_t)(\Sigma_t^s + (\mu_t - \mu_t^s)(\mu_t - \mu_t^s)^T).\end{aligned} \tag{4}$$

0-7695-2521-0/06 $20.00 © 2006 IEEE

With $P(g|y)$ being approximated as a Gaussian, $P(g_t|y_{1:t})$ is also an Gaussian and the overall model performs like a Kalman filter. However, unlike a standard Kalman filter that has fixed measure function $P(y_t|g_t)$, our measurement function is dynamically changing.

2.2 Rao-Blackwellized Particle Filter

When applying our dynamical model to visual tracking problems, we need to estimate the object location and scale variables l_t in addition to their global coordinates g_t. Since the posterior distribution $P(g_t, l_t|Y_t)$, where Y_t is analogous to $y_{1:t}$, is unlikely to be modeled well by any analytical distribution, we approximate it with a particle filter. Figure 1(b) illustrates our complete dynamic inference model for visual tracking. If we directly draw samples from the state space of (g_t, l_t), a larger number of particles are required for good approximation. However, the analytic distribution of $P(y_t|g_t)$ described in the Section 2.1 allows us to integrate out part of the latent variables and reduce the number of dimensionality needed for sampling, which is the concept of the Rao-Blackwellized particle filter.

The likelihood function $P(g_t, l_t|Y_t)$ can be computed by

$$P(g_t, l_t|Y_t) = P(g_t|l_t, Y_t)P(l_t|Y_t), \tag{5}$$

where $P(g_t|l_t, Y_t)$ is an analytical distribution in our model and $P(l_t|Y_t)$ is approximated with a particle filter. To include the analytical distribution, $P(g_t|l_t, Y_t)$, the particle consists of three tuples $\{(s^{(i)}, w^{(i)}, \alpha^{(i)}(g))\}_{i=1}^N$, where $s^{(i)}$ is a sample of l, $w^{(i)}$ is its weight, and distribution $\alpha^{(i)}(g)$ corresponds to $P(g_t|l_t, Y_t)$. For a Rao-Blackwellized particle filter, the likelihood $P(g_t|l_t, Y_t)$ is approximated by

$$P(g_t, l_t|Y_t) \approx \sum_i w^{(i)}\delta(s^{(i)})\alpha^{(i)}(g). \tag{6}$$

According to the Bayes filter, we can estimate the object location by

$$P(l_t|Y_t) = \kappa \int_{g_t} (Y_t|g_t, l_t)\times \int_{l_{t-1}} \int_{g_{t-1}} P(g_t, l_t|g_{t-1}, l_{t-1})P(g_{t-1}, l_{t-1}|Y_{t-1}). \tag{7}$$

In our model, we assume the dynamics of g and l are independent:

$$P(g_t, l_t|g_{t-1}, l_{t-1}) = P(g_t|g_{t-1})P(l_t|l_{t-1}), \tag{8}$$

and each dynamics is a Brownian motion:

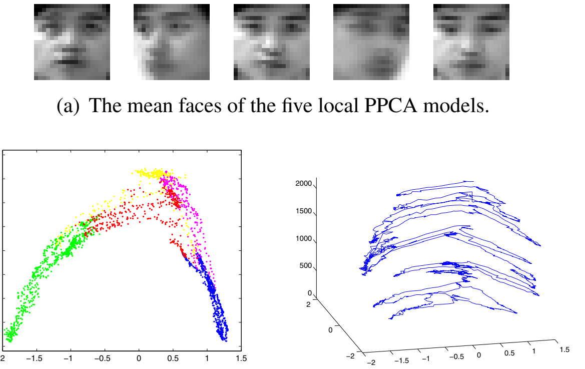
$$\begin{aligned} P(g_t|g_{t-1}) &\sim \mathcal{N}(g_{t-1}, Q) && (9)\\ P(l_t|l_{t-1}) &\sim \mathcal{N}(l_{t-1}, R) && (10) \end{aligned}$$

where Q and R are predefined covariance matrices.

With all the model parameters being defined, our Rao-Blackwellized particle filter is described as Algorithm 1.

Algorithm 1 A Rao-Blackwellized particle filter for visual tracking using globally coordinated nonlinear manifold

```
Starting with particles $\{(s_{t-1}^{(i)}, w_{t-1}^{(i)}, \alpha_{t-1}^{(i)}(g_{t-1}))\}_{i=1}^N$:
  1. Re-sample the particles from $\{s_{t-1}^{(i)}\}_{i=1}^N$ according to $\{w_{t-1}^{(i)}\}_{i=1}^N$, and denote the newly selected particles as $\{s_t^{(i)}\}_{i=1}^N$.
  2. Drift $\{s_t^{(i)}\}_{i=1}^N$ according to $P(l_t|l_{t-1})$.
  3. Update distributions $\alpha_{t-1}^{(i)}(g_{t-1})$ to $\{\alpha_t^{(i)}(g)\}$ according to $P(g_t|g_{t-1})$ and $P(y_t|s_t^{(i)}, g_t)$:
     $\alpha_t^{(i)}(g) = \kappa^{(i)} P(y_t|s_t^{(i)}, g_t) \int P(g_t|g_{t-1})\alpha_{t-1}^{(i)}(g_{t-1})dg_{t-1}.$   (11)
  4. Set $w_t^{(i)} = \kappa^{(i)}$.
```



(a) The mean faces of the five local PPCA models.

(b) 2D globally-coordinated mixture of five local PPCA models.

(c) X-Y-T view of the projected trajectory of the training video.

Figure 2. Manifold learning results.

3 Experiments

We test our globally-coordinated dynamical model by tracking human faces. To collect training images, we take a video of a face undergoing large pose variations in a clear, white background. We use a baseline appearance based tracker to crop out face images and scale these images to 19 by 19 pixels. The pose variations are large enough so that the baseline tracker loses track quickly. Therefore, we restart the tracker every 50 frames and obtain $2,200$ frames of face images. Note that the cropped face images do not align very well, compared to the hand-cropped training images used by most of the face-related research, which makes this sequence of training images even more challenging.

We learn a two-dimensional globally-coordinated non-

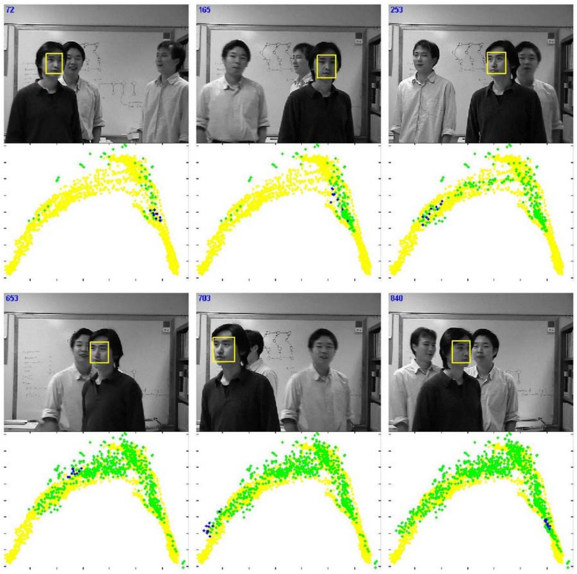


Figure 3. Tracking results.

linear manifold using these $2,200$ time-varying, 361-dimensional image vectors with a mixture of five 13-dimensional probabilistic PCA (PPCA) models. The learned two-dimensional global coordinates of the face images are depicted in Figure 2(b). The color of each point denotes the assigned cluster label of the mixture of PPCA models. It shows that the property of local linearity in input images is well captured in the mixture model. The five PPCA models (from left to right in Figure 2(a)) also correspond to faces that look upward, right, forward, left, and downward. Figure 2(c) validates the continuous projected trajectory of training video.

We test our tracker by tracking a target face in a cluttered background where multiple faces are present in the scene. All faces in the images undergo obvious pose variations. We initialize a tracking window of the target face in the first frame of the test video, and use 500 particles to track target face in the remaining image frames. Although multiple similar objects moving nearby makes tracking a challenging problem, our tracker tracks the target face very well even with changes of facial expressions. Figure 3 shows several snapshots of our tracking process and the 2D projections of the training images (yellow), the tracked faces from the first frame (green), and the most recent ten tracked faces (blue). More experimental results with temporary occlusion are available on our web site (`http://vision.ai.uiuc.edu/~cbliu/`). The tracker currently runs about 1 frame/second with non-optimized Matlab code using a 2.4GHz PC.

Our graphical model for visual tracking provides extra robustness for trackers because we track object position as well as its pose (appearance coefficients as 2D coordinates of the nonlinear manifold). The current pose estimation becomes a strong prior while tracking the next frame and prevents our tracker from being distracted by other objects with similar appearances. Some other trackers may perform well, but they do not simultaneously provide positional and pose information at each frame.

4. Conclusions

In this paper, we present a dynamic inference algorithm for nonlinear appearance based object tracking. We also apply a Rao-Blackwellized particle filter to facilitate efficient object tracking. Our dynamical model captures continuous motion as a continuous low-dimensional trajectory. By tracking appearance on a nonlinear low-dimensional manifold in addition to the object's location and scale, the tracking performance is more robust in the presence of other similar objects.

Our main goal is to develop a generative model capable of describing the process of appearance variation over time. That is, we aim at tracking and understanding simultaneously. The tracker understands the underlying process that causes appearance changes and uses this information for robust tracking. Our experiments validate our proposed model and that the inference procedure is correct.

References

[1] A. Agarwal and B. Triggs. Tracking articulated motion using a mixture of autoregressive models. In *Proc. ECCV*, 2004.

[2] Z. Ghahramani and G. E. Hinton. Variational learning for switching state-space models. *Neural Computation*, 12:831–864, 2000.

[3] Z. Khan, T. Balch, and F. Dellaert. A Rao-Blackwellized particle filter for eigentracking. In *Proc. IEEE CVPR*, volume 2, pages 980–986, 2004.

[4] V. Pavlovic, J. M. Rehg, and J. MacCormick. Learning switching linear models of human motion. In *Neural Information Processing Systems*, volume 13, 2000.

[5] S. Roweis and L. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, Dec 2000.

[6] S. Roweis, L. Saul, and G. E. Hinton. Global coordination of local linear models. In *Neural Information Processing Systems*, volume 14, pages 889–896, 2001.

[7] Y. W. Teh and S. Roweis. Automatic alignment of local representations. In *Neural Information Processing Systems*, volume 15, pages 841–848, 2002.

[8] J. B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, Dec 2000.

[9] Q. Wang, G. Xu, and H. Ai. Learning object intrinsic structure for robust visual tracking. In *Proc. IEEE CVPR*, volume 2, pages 227–233, 2003.