Location-specific Transition Distributions for Tracking

Nathan Jacobs, Michael Dixon, and Robert Pless
Department of Computer Science and Engineering
Washington University in St. Louis St. Louis, MO, 63117
{jacobsn,msd2,pless}@cse.wustl.edu

Abstract

Surveillance and tracking systems often observe the same scene over extended time periods. When object motion is constrained by the scene (for instance, cars on roads, or pedestrians on sidewalks), it is advantageous to characterize and use scene-specific and location-specific priors to aid the tracking algorithm. This paper develops and demonstrates a method for creating priors for tracking that are conditioned on the current location of the object in the scene. These priors can be naturally incorporated in a number of tracking algorithms to make tracking more efficient and more accurate. We present a novel method to sample from these priors and show performance improvements (in both efficiency and accuracy) for two different tracking algorithms in two different problem domains.

1. Introduction

When a camera observes the same scene over a long period of time, and when objects within that scene have some consistency in the way they move, then a tracking algorithm can use these regularities to improve performance. This paper presents an initial approach to characterizing object motion within a scene, and using this as a prior to improve tracking.

It is well known that an accurate transition distribution is an important part of a tracking algorithm. Specifying a transition distribution requires domain knowledge or many training examples. In applications in which a camera observes the same scene for a long time many examples of object transitions are available. We use these previous observations to learn object transition distributions and we focus on the fact that these distributions often depend on the location of the object in the scene (See Figure 1).

In this work we show two improvements possible when using location-specific transition distributions: improved prediction accuracy and increased computational efficiency. The improved prediction accuracy is possible because the location-specific distributions can incorporate location-specific biases, such as a stop sign in a traffic scene. The computational efficiency gains are the result of the difference between the variances of the transition distributions in different image regions. These differences make it possible to expend far less computational effort because it is easy to track an object through an image region with a low variance transition distribution. In sampling-based tracking algorithms, the computational effort is directly related to the number of samples the algorithm uses. We modify a sampling-based tracking algorithm by replacing a standard sampling scheme that uses a fixed number of samples with one that adapts the number of samples based on the transition distribution and the measurement model.

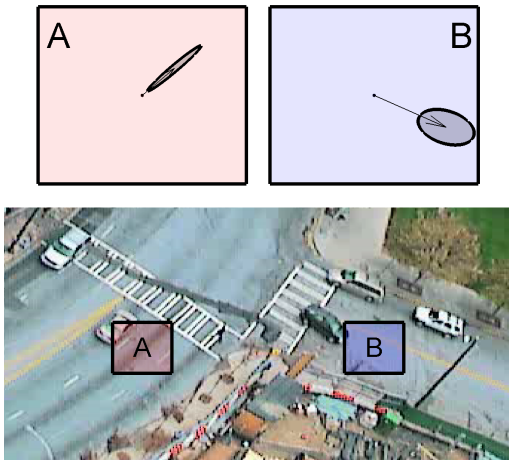


Figure 1. An illustration of location-specific transition distributions. The arrows and ellipses (top) represent the mean object translation and translation uncertainty conditioned upon the object starting in the corresponding box at bottom. In this work, we show how these distributions can be used to improve tracking algorithms.

We show results in two important domains: tracking vehicles for surveillance applications and tracking faces for human-computer interaction applications. In the vehicle tracking domain the location dependence of object transi-

tions is strong because motions are constrained by traffic laws. In the face-tracking domain the location-dependence is less strong but, as we will show, location-dependence is present and beneficial. As an example of the benefits, our results show that, when using a location-specific distribution, 60% fewer samples are required to obtain the same accuracy as when using a location-independent (global) distribution.

2. Related Work

Tracking is a rich domain, encompassing research in filtering, prediction, measurement models, and local alignment procedures. Our work builds on classical probabilistic tracking with a focus on learning priors from a scene and adapting the dynamics of the tracking algorithm. We believe the most related work is as follows.

There is a large body of work on learning motion patterns and motion priors from observed object transitions. Here we describe representative works [7, 4] from each of these areas. North and Blake [7] use an EM-algorithm to learn a transition distribution for tracking. The motion priors are shown to improve the accuracy of tracking and to enable classification of discrete object states. Hu et al. describe a system [4] for learning motion patterns in a scene from trajectories generated by an object tracker. The patterns are then used to classify individual trajectories as anomalous and to predict future object trajectories from partial trajectories.

The use of location-specific models of pixel intensities in static camera scenes, commonly referred to as background modeling, is well known. In addition, location-specific models of object shape [5] and spatio-temporal derivatives [8] have been used to improve object detection and anomalous motion detection.

In the context of mobile robot localization, Fox [3] chooses the number of particles to satisfy a bound on the Kullback-Leibler divergence between the sample set and a grid-based approximation of the true distribution. This approach works well in the two-dimensional robot localization problem but becomes expensive as the dimensionality of the state space grows. In the domain of tracking, Zhou et al. [11] adapt the number of samples based on the particle prediction error.

To our knowledge, ours is the first work to use location-specific object transition distributions in a static camera tracking application.

3. Object Tracking Background

We begin with a description of the sequential Bayesian tracking framework. Given an object with a sequence of states $\{x_k, k \in \mathbb{N}\}$ governed by a state transition function $x_k = f(x_{k-1}, v_{k-1})$ and a corresponding measurement sequence $\{z_k = h(x_k, n_k), k \in \mathbb{N}\}$, where v_{k-1}, n_k are noise terms, our goal is to estimate the current state x_t given only the current and previous measurements $\{z_1, \ldots, z_t\}$. In the Bayesian tracking framework, our goal is to construct a probability density function (pdf) $p(x_t|z_{1:t})$ to specify our belief in the current state x_t given only the previous measurements $z_{1:t}$. Assuming the existence of an initial state estimate $p(x_0|z_0) = p(x_0)$ the desired pdf can be estimated in a two stage process.

In the first stage, the current belief $p(x_{t-1}|z_{1:t-1})$ is propagated forward using the Chapman-Kolmogorov equation:

$$p(x_t|z_{1:t-1}) = \int p(x_t|x_{t-1})p(x_{t-1}|z_{1:t-1})dx$$

in which $p(x_t|x_{t-1})$ is a transition distribution that represents the uncertainty in the next state given the current state.

In the second stage, the predicted pdf $p(x_t|z_{1:t-1})$ is updated with the current measurement z_t using Bayes' rule to obtain the posterior pdf

$$p(x_t|z_{1:t}) \propto p(z_t|x_t)p(x_t|z_{1:t-1}).$$

The measurement models we use consist of a local alignment procedure, such as Lucas-Kanade [6, 1] or mean-shift tracking [2], that attempts to align the object representation, such as a bitmap template, to the current image. A common characteristic of these algorithms is the need to specify a starting location for the alignment procedure. We address this by sampling starting locations from the predicted pdf $p(x_t|z_{1:t-1})$. Section 5 discusses our approach to adapting the number of samples based on the shape of the basin of attraction of the alignment procedure.

This paper considers the use of location-specific transition distributions $p(x_t|x_{t-1})$ for tracking. We leave location-specific measurement models $p(z_t|x_t)$ for future work.

4. Location-Specific Transition Distributions

Many scenes have structure, such as roads and sidewalks, that significantly constrains the way objects move. In video of these types of scenes, the motion of objects is similarly constrained. Tracking algorithms can learn these constraints, in the form of statistical transition distributions, by watching the scene for a long time. This section formally describes, provides an example of, and discusses issues related to the learning and use of location-specific transition distributions.

The transition distributions used in most tracking applications are spatially uniform, the most common form being linear-Gaussian:

$$p(x_{t+1}|x_t) = N(x_{t+1}|Fx_t + \mu, \Sigma). \quad (1)$$

In this form, F is the process matrix, μ is a bias term, and Σ represents the uncertainty in the prediction. Often these parameters are manually specified, but given sufficient training examples it is easy to solve for the maximum-likelihood set of parameters.

In this work we extend Equation 1 by learning different parameters for different image regions. This gives a transition distribution with the following *locally* linear-Gaussian form:

$$p(x_{t+1}|x_t) = N(x_{t+1}|F(x_t)x_t + \mu(x_t), \Sigma(x_t)). \quad (2)$$

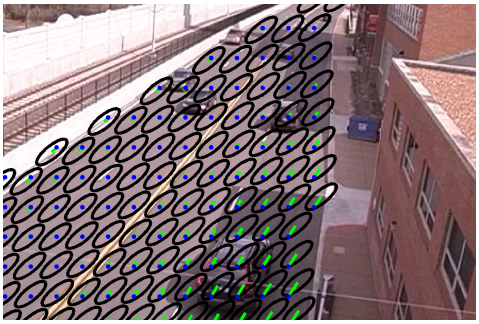
The process matrix $F(x_t)$, process bias $\mu(x_t)$, and the covariance matrix $\Sigma(x_t)$ are dependent on the current state. Although the process parameters can depend arbitrarily on the current object state, in this work we focus on the special case of depending exclusively on the object's image location. There are many possible ways of representing this conditional pdf, for efficiency and simplicity we divide the image into regularly-spaced bins and learn a parameter set for each bin. Our results demonstrate that this simple conditional pdf structure and training procedure works well in practice.

4.1. An Example: Interest Point Tracker

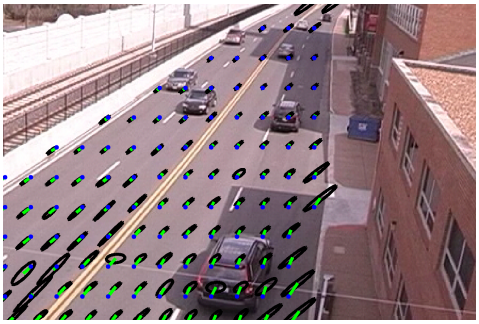
We begin with a concrete example of the structure and usefulness of location-specific transition distributions. Our interest point tracker works as follows: First locate interest points [9] in a temporal difference image, then predict the location of the interest points in the next frame, and finally locate the points using Lucas-Kanade [6, 1]. This gives a set of tracked interest points that correspond to regions of the image that have changed recently. Using these interest points we learn two transition distributions $p(x_{t+1}, y_{t+1}|x_t, y_t)$: one global (Equation 1) and one location-specific (Equation 2), which predict the next location given the current location.

The state space used in this tracker is the x, y position of the interest point. Figure 2 shows the mean translation and the covariance ellipses for the motion of an interest point in a video of an urban traffic scene. As the figure shows, the uncertainty in the next state distributions is much lower for the local model than the global model.

In Section 6.1, we use these transition distributions to demonstrate the accuracy improvements possible using these distributions. The results show that the local predictions are better than the global predictions, which are better than a static prediction (i.e., the tracker is initialized in the same location). In Section 6.2 we show how location-specific transition distributions are helpful in the case of tracking faces in an affine state space.



(a) Linear-Gaussian transition distribution



(b) Locally linear-Gaussian transition distribution

Figure 2. A visualization of the transition distribution for an interest point tracker. The dots (blue) represent the current location of the interest point. The lines (green) represent the mean translation of the interest point, and the ellipses represent the three standard deviation equi-probability ellipse of the distribution. (a) The global linear-Gaussian distribution has higher uncertainty. (b) The local distribution has lower uncertainty and more realistic translations.

4.2. Discussion

The accuracy and computational efficiency gains possible from location-specific transition distributions depend on a number of factors. The most significant factor is the amount of location dependence of the transition distribution. Other factors include the object state space and the measurement model.

As the dimensionality of the state space increases, the performance gains possible with location-specific distributions typically increase. Including higher-order terms such as velocity can reduce the magnitude of performance gains. In general, having a physically-accurate state space and dynamics model reduces the benefit of a location-specific transition distribution. However, in most scenes there will be structure that locally impacts the transition distribution that cannot be modeled without significant human effort or cannot be inferred easily from image data.

One benefit of location-specific models that we have left unexplored in this work is in object initialization, especially for state spaces with higher-order terms. For example, learning a location-specific state distribution can reduce the

problem of determining the velocity of an object when it appears. This improvement would be significant in scenes with many new objects that appear (e.g., traffic scenes).

One concern in learning a location-specific model compared to a global model is the increased number of training examples required to avoid overfitting. In applications such as surveillance, where there are many training examples the need for additional training examples is less of a concern. In our shorter videos, usually 10 minutes long, we address this by including a small number of examples from other parts on the image in the training set along with the location-specific examples. This reduces the chance that a few incorrect examples in one bin will result in a very inaccurate distribution.

In most tracking algorithms, the computational efficiency gains from using more specific transition distributions come from modifying the number of samples the algorithm uses based on the uncertainty in the next object state. In the next section we describe a deterministic sampling technique that determines the number and location of samples based on the transition distribution and measurement model.

5. A Deterministic Method for Varying the Number of Samples

The number of samples used in a tracker has a direct impact on its computational efficiency. Tracking algorithms that use a fixed number of samples often expend excess effort when tracking is easy (the transition distribution is narrow) and insufficient effort when tracking is hard (the transition distribution is broad). We show an automatic technique for determining the samples to use, both the number and location, given the transition distribution. This technique benefits from the fact that location-specific transition distributions are more accurate than global distributions on many scenes.

The correct number of samples also depends on the state space and the measurement model—in our case, a local alignment procedure. The alignment procedures we use start from an initial point in state space and attempt to converge to a local minima. Given a local minima, the set of starting points for which the alignment procedure converges to the local minima is called a basin of attraction. We label two points as indistinguishable if they are in the same basin of attraction. The key observation is that samples do not need to be so close together that they are indistinguishable.

We approximate the basin of attraction by using the local alignment procedure on example objects from the scene. We start with a correctly-aligned template, perturb it by sampling from the transition distribution, and then run the alignment procedure. We then determine if the resultant template is correctly aligned by comparing it to the original template (e.g., by comparing the image space distance between the templates). Given this set of points we fit a zero-mean Gaussian to the correctly-aligned starting points. Our approximation of the basin of attraction is the largest equi-probibilitity ellipse of the Gaussian such that 98% of the samples inside the ellipse converged correctly.

Using the basin of attraction ellipsoid we generate well-spaced samples deterministically. The basic idea, illustrated in Figure 3, is to tile the high-probability region of the transition distribution with the capture ellipsoid.

The procedure for generating samples deterministically is as follows:

1. Transform the state space so that the capture ellipsoid becomes a unit circle.
2. Generate evenly-spaced samples on an axis aligned with the data distribution. Discard sample points that are outside the three standard deviation ellipse of the transformed transition distribution.
3. Undo the transformation to obtain well-spaced samples in the original state space.

This sampling scheme incorporates both the uncertainty in the transition distribution and the ability of the local alignment procedure to find the correct position to reduce the number of samples.

6. Experimental Results

The results demonstrate that using location-specific transition distributions is useful in several ways for improving the performance and accuracy of tracking algorithms.

6.1. Prediction Accuracy

In this section, we compare the accuracy of different transition distributions using the interest point tracker described in Section 4.1. We conducted these experiments on video from several urban traffic scenes[1], each 10 fps and approximately 10 minutes long. The model parameters were learned from the first five minutes of the video, and the different distributions were tested over the remainder. In this experiment, the starting point of the Lucas-Kanade procedure used was the mean of the distribution.

The results in Figure 4 show that the location-specific transition distribution (Equation 2) is more accurate than a global distribution (Equation 1). The location-specific transition distribution enables the tracker to handle larger object motions, caused by faster moving objects and/or lower frame rates. With the global distribution the predicted starting point is often outside the basin of attraction of the correct position.

[1] Videos are courtesy of the Federal Highway Administration's Next Generation Simulation program: http://ngsim.camsys.com/.

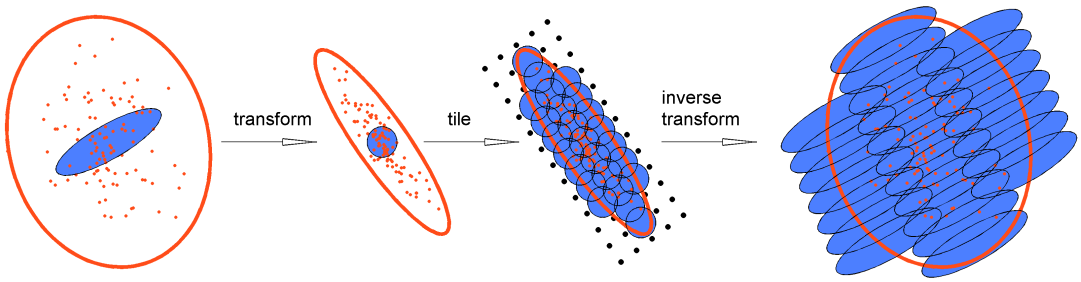
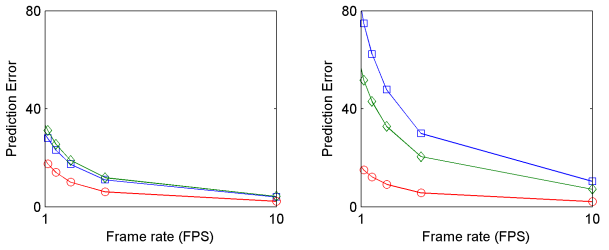


Figure 3. An illustration of the deterministic process, described in Section 5, for generating samples (the centers of the small ellipses in the rightmost sub-figure). In each sub-figure, the small (blue) ellipses represent the basin of attraction of a sample and the large (red) ellipse is the three standard deviation ellipse of the transition distribution, and the (red) dots are samples from the data distribution.



(a) Urban Intersection (b) Highway

Figure 4. Plots of prediction error, average distance between predicted interest point location and actual location, computed from two different videos at differing frame rates for three methods of prediction: no prediction, global-model prediction, and location-specific model prediction (square, diamond, and circle). These figures show that location-specific models improve the prediction accuracy significantly, especially at lower frame rates.

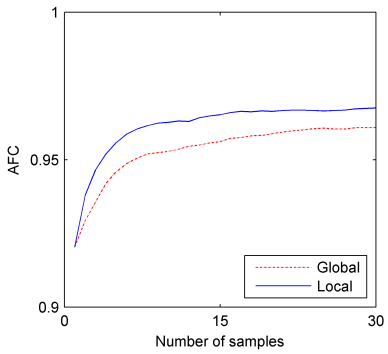


Figure 5. A plot of the average frequency of convergence (AFC) when using different numbers of starting samples. Increasing the number of starting samples improves the chances of converging to the correct position at the cost of increased computation. When samples are drawn from the location-specific distributions rather than global distributions, the same accuracy can be achieved with fewer samples.

6.2. Impact of Changing the Number of Samples

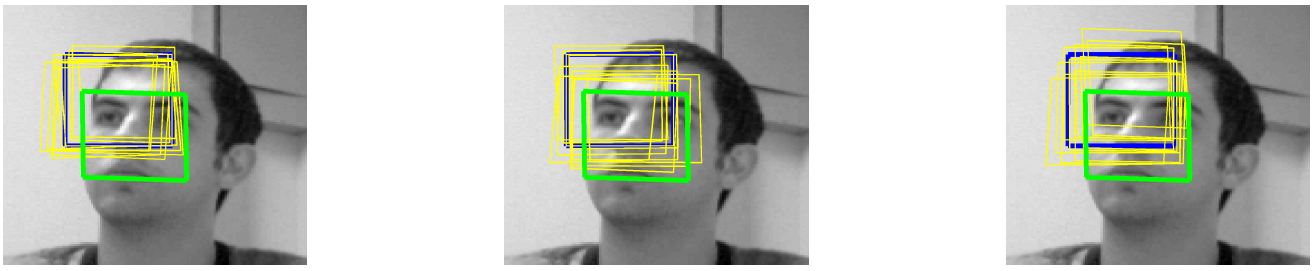
In this section, we explore the relationship between the number of initial samples and the accuracy of the tracker. Our results demonstrate that using initial samples drawn from location-specific distributions rather than global distributions reduces the number of tracking failures by approximately 20%. We also show that our deterministic adaptive sampling algorithm (described in Section 5) further improves the tracking accuracy. Figure 6 shows samples generated for different transition distributions and sampling schemes.

In these experiments, we compare the performance of the two transition distributions on the task of face tracking using an affine state space. We evaluate each the distributions on a video of a person using a computer. To create a ground truth for our evaluation, we captured and tracked [1] the video at 60 fps, hand-verified the high-frame-rate tracking results, and then downsampled the video to 10 fps.

We performed our evaluations as follows. For each frame we initialized the tracker using the output of a face detector [10] and then searched for the best alignment of this template in the subsequent frame. In the case of multiple initial samples, we solved for the locally optimal template alignment for each sample independently and then selected the alignment with the lowest root mean squared error. We then compared the output of the tracker to the ground truth position to determine if the algorithm converged to the correct location.

In the first experiment, we compare the accuracy of the global and location-specific transition distributions for varying number of initial samples. As a quantitative measure of tracking performance, we compute the average frequency of convergence (AFC), defined as the number of frames that converged to the correct location divided by the total number of frames evaluated. In Figure 5, we show the AFC of the two sampling methods for varying numbers of starting samples, ranging from 1 to 30.

Our results show that when samples are drawn from the location-specific distributions rather than global distributions, there is a measurable improvement in tracking performance. For a fixed number of samples, the use of location-specific transition distributions results in up to 20% fewer



(a) global (b) location-specific (random) (c) location-specific (deterministic)

Figure 6. Examples of different sampling methods. The blue box shows the position of the face in the previous frame. The green box shows the ground truth position in the current frame. The yellow boxes show the multiple initial guesses used to initialize the tracker. In this frame, the face is located at a position where large movements were more common. Notice that samples chosen from the global distribution (a) are more tightly centered around the previous position, while the samples drawn from the location-specific distribution (b,c) are more spread out, improving the likelihood of capturing large movements.

tracking failures. Alternatively, we see that using location-specific distributions provides the same accuracy as the global distribution while using far fewer samples. As an example, to obtain an accuracy of 96% using the location-specific distribution requires seven samples but when using the global distribution 25 samples are required.

As a second experiment, we used the dataset and tracker from the first experiment but replaced the random sampler with the deterministic adaptive sampling scheme described in Section 5. The training set for learning the basin of attraction contains 70,000 starting states generated from faces detected on 700 different frames. The algorithm uses the learned basin of attraction and adapts the number of samples based on the location-specific variance of the transition distribution. On average, the algorithm generated 3.9 samples per frame. The result is an accuracy improvement of 9% over the random sampling strategy that used four samples per frame using the same transition distribution. The accuracy improvement is the result of allocating more samples to periods when the transition distribution has higher variance.

7. Conclusion

In video of scenes with structure that constrains typical object motions, location-specific transition distributions can be used to improve tracking algorithms. We show accuracy and performance improvements possible when using such distributions. We also describe a deterministic sampling method that varies the number of samples based on the transition distribution and the measurement model.

Acknowledgement

The authors gratefully acknowledge the support of the NSF through Career award IIS-0546383.

References

[1] S. Baker and I. Matthews. Lucas-kanade 20 years on: A unifying framework. *International Journal of Computer Vision*, 56(3):221–255, 2004.

[2] D. Comaniciu, V. Ramesh, and P. Meer. Real-time tracking of non-rigid objects using mean shift. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 142–151, 2000.

[3] D. Fox. Kld-sampling: Adaptive particle filters and mobile robot localization. 2001.

[4] W. Hu, X. Xiao, Z. Fu, D. Xie, T. Tan, and S. Maybank. A system for learning statistical motion patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(9):1450–1464, 2006.

[5] N. Jacobs and R. Pless. Shape background modeling : The shape of things that came. In *Proc. IEEE Workshop on Motion and Video Computing (WMVC)*, Austin, Tx, Feb. 2007.

[6] B. D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *Proc. International Joint Conference on Artificial Intelligence*, pages 674–679, 1981.

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

[8] R. Pless, J. Larson, S. Siebers, and B. Westover. Evaluation of local models of dynamic backgrounds. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 73–78, 2003.

[9] J. Shi and C. Tomasi. Good features to track. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Seattle, June 1994.

[10] P. Viola and M. J. Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137–154, May 2004.

[11] S. K. Zhou, R. Chellappa, and B. Moghaddam. Visual tracking and recognition using appearance-adaptive models in particle filters. *IEEE Transactions on Image Processing*, 13:1491–1506, 2004.