Particle dynamics and multi-channel feature dictionaries for robust visual tracking

Srikrishna Karanam
karans3@rpi.edu
Yang Li
yangli625@gmail.com
Richard J. Radke
rjradke@ecse.rpi.edu

Department of Electrical, Computer, and Systems Engineering
Rensselaer Polytechnic Institute
110 8th St.
Troy, NY USA

Introduction: Recent advances in the application of compressive sensing to traditional computer vision problems such as face recognition [5] inspired several visual tracking approaches based on sparse representations. The core idea of these approaches is to build an appearance model of the object using several pre-defined templates. The problem of tracking the object is then cast as finding a sparse approximation in the subspace spanned by the templates. In [3], Mei and Ling introduced the l_1 tracker, demonstrating impressive tracking results. Given an appearance model $\mathbf{A} = [\mathbf{t}_1 \cdots \mathbf{t}_n] \in \mathbb{R}^{m \times n}$ of an object formed using a set of templates $\mathbf{t}_i \in \mathbb{R}^m, i = 1, \ldots, n$, they express a tracking result $\mathbf{y} \in \mathbb{R}^m$ as $\mathbf{y} = \mathbf{A}\mathbf{x} + \boldsymbol{\varepsilon}$, where $\mathbf{x} \in \mathbb{R}^n$ is the sparse coefficient vector that is to be recovered, and $\boldsymbol{\varepsilon} \in \mathbb{R}^m$ is used to account for partial occlusions. The l_1 tracking algorithm hypothesizes that $\mathbf{x}$ and $\boldsymbol{\varepsilon}$ are sparse for a good tracking candidate, and recovers them by solving an l_1 regularized least squares problem. Subsequently, the candidate with the least projection error in the template subspace is chosen as a tracking target, and a Bayesian state inference model in a particle filter framework is used to track the object over time.

Contributions: In spite of the impressive progress achieved by the l_1 tracker and its recent variants, several issues remain that often lead to tracking failures. First, most related methods employ only a fixed-variance Gaussian distribution to represent the state transition model. Such a fixed-variance state transition model can cause significant drift errors in the approximation of the particle filter, resulting in severe tracking failures. In our work, we propose to mitigate this problem by adaptively learning the variance from past states using a dynamic state transition model. Specifically, we employ an autoregressive model in conjunction with block Hankel matrices to continuously learn the dynamics from past data. Second, most existing approaches use extremely low resolution image intensity features as part of the appearance model. Such features do not capture sufficient visual information required to reliably track the object and avoid drift. To mitigate this problem, we propose a three-channel appearance dictionary comprised of image intensity information, normalized image gradient magnitudes, and histograms of oriented gradients to construct an appearance model of the object. Finally, tracking algorithms typically employ a fixed number of particles to approximate the posterior distribution (e.g., 600 in [3], 400 in [6]). In this work, we demonstrate that many particles are not necessary to reliably track an object, given the initial location. Specifically, we propose to adapt the number of particles required during the state estimation process using the Kullback-Leibler (KL) distance measure [2].

Approach overview: We formulate visual tracking as a sparse representation problem in a particle filtering framework. Given the initial location of the target to be tracked, we warp the image into a 64×64 pixel template, thereby representing the position of the target in each frame using a four-dimensional state vector $\mathbf{s_t} \in \mathbb{R}^4$. By perturbing the initial location by a few pixels (typically, 1–3), we form m such templates. In our experiments, we set $m = 10$. We then construct three appearance dictionaries using these templates: an intensity channel dictionary $\mathbf{A}^1 = [\mathbf{t}_1^1 \cdots \mathbf{t}_m^1]$, a normalized gradient magnitude dictionary $\mathbf{A}^2 = [\mathbf{t}_1^2 \cdots \mathbf{t}_m^2]$, and a Histogram of Oriented Gradients (HOG) [1] dictionary $\mathbf{A}^3 = [\mathbf{t}_1^3 \cdots \mathbf{t}_m^3]$, where each dictionary $\mathbf{A}^j \in \mathbb{R}^{d_j \times m}$. Now, given a potential target particle $\mathbf{y}$, we compute its intensity feature vector $\mathbf{y}^1$, normalized gradient magnitude vector $\mathbf{y}^2$, and HOG vector $\mathbf{y}^3$. In each feature channel, we hypothesize that a good target candidate can be represented as a sparse linear combination of the dictionary templates, and recover the sparse vector by solving the following convex optimization problem:

$$(\mathbf{x}^{j*}, \boldsymbol{\varepsilon}^{j*}) = \arg\min_{\mathbf{x}^j, \boldsymbol{\varepsilon}^j} \|\mathbf{x}^j\|_1 + \|\boldsymbol{\varepsilon}^j\|_1 \text{ s.t. } \mathbf{y}^j = \mathbf{A}^j\mathbf{x}^j + \boldsymbol{\varepsilon}^j, j = 1,2,3 \quad (1)$$

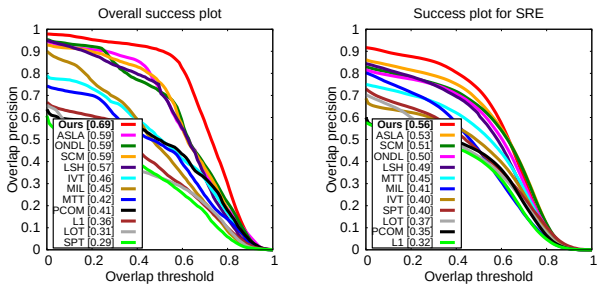
where $\boldsymbol{\varepsilon}^j$ is used to account for error and partial occlusion, and $\mathbf{x}^j$ is the sparse coefficient vector we wish to recover.

Tracking in a particle filtering framework proceeds by generating several hypotheses, and testing each for its likelihood. In our context, each hypothesis is a candidate particle, represented by its state vector. We search for potential candidate particles using an adaptive state transition model incorporating dynamic information. Our key insight is that learning from the dynamics of past state vectors can lead to improved and efficient search for new candidate particles, leading to both accuracy and speed benefits. Formally, if $\mathbf{s}_t \in \mathbb{R}^4$ is the current state vector, we estimate the elements of the next state vector $\mathbf{s}_{t+1}$ using the following transition model:

$$s_{t+1}(i) = s_t(i) + r(i)\sigma_{t+1}(i) \quad (2)$$

where $s_{t+1}(i)$ is the i^{th} element of $\mathbf{s}_{t+1}$, $r(i) \sim \mathcal{N}(0,1)$ is a normally distributed random number, and $\sigma_{t+1}(i)$ is the i^{th} element of the variance vector $\sigma_{t+1} \in \mathbb{R}^4$ that we estimate on-the-fly using the dynamics of the past states.

Results: We tested our algorithm on 25 publicly available video sequences[1] that represent several challenging aspects in visual tracking: illumination variation, scale variation, occlusion, non-rigid object deformation, motion blur, fast motion, in-plane rotation, out-of-plane rotation, out-of-view, background clutter, and low resolution.



(a) (b)

Figure 1: Success plots averaged over all the 25 test sequences. In both plots, the area under the curve (AUC) is reported in the legend.

The overall success plot and the success plot for the spatial robustness evaluation (SRE) test [4] are shown in Figure 1. We see that our approach results in a significant 10% overall improvement and an improvement of 3% for the SRE test when compared to the state of the art.

[1] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

[2] Dieter Fox. Adapting the sample size in particle filters through KLD-sampling. *IJRR*, 22(12):985–1003, 2003.

[3] Xue Mei and Haibin Ling. Robust visual tracking using l_1 minimization. In *ICCV*, 2009.

[4] Yi Wu, Jongwoo Lim, and Ming-Hsuan Yang. Online object tracking: A benchmark. In *CVPR*, 2013.

[5] Allen Y Yang, Zihan Zhou, Arvind Ganesh, S Shankar Sastry, and Yi Ma. Fast l_1-minimization algorithms for robust face recognition. *IEEE T-IP*, 22(8):3234–3246, 2013.

[6] Tianzhu Zhang, Bernard Ghanem, Si Liu, and Narendra Ahuja. Robust visual tracking via structured multi-task sparse learning. *IJCV*, 101(2):367–383, 2013.

[1] Videos demonstrating our tracking performance can be found in the supplementary material.