Chapter 1

Problem Definition

1. Start and stop of tracks
   (a) Occlusions occur when object goes in front of tracked object
   (b) tracks can start/stop at edges of frame and when appearing behind a foreground object
2. objects move at a certain velocity,
3. objects can change scale and rotation
4. objects are rigid
5. objects are at the same distance from the camera. Objects aren't necessarily the same colour/texture, but they move in the same tracks
6. video is not like static images: motion blur is inherent in fast moving objects, making it look different than normal
7. realtime/anytime (improves, but can be stopped at any time)

1. how do you make an view-independent feature set for moving objects?
2. optimization: define a quickly converging optimization function

1.1 Potentially Useful Ideas

1. markov chain
   (a) viterbi algorithm
   (b) forward-backward algorithm
2. integral images
3. circulant structure (henriques)
4. bootstrapping
5. decision trees
6. entropy/infogain
7. quad trees $O(n^2)$ distances between each n point

high level ideas

1. recursion
2. optimization: define a quickly converging optimization function

1. k-d tree ”kd-trees are not much better than brute-force search when the number of descriptor dimensions exceeds 20”

Chapter 2

Related Work

2.1 Online RGB Trackers (vision papers)

Survey papers:

1. **Pang et al's Survey** [**?**] notes biases in comparisons: usually new papers list their method as the best (because of a specific methodology); however second best paper rankings are fairly robust. A meta-analysis concludes the following methods are competitive: Struck, MIL, TLD, VTD.

2. **an extensive PAMI Survey** [**?**] claims Struck is the best, and analyses specific failure cases and how it affects specific method

3. **Appearance Model Survey** [**?**]

4. A comprehensive list of papers can be found by searching papers that have referenced Struck; which is the most frequently used benchmark to compare against.

5. **Visual Tracking Benchmark**[**?**]

6. Wu et. al [**?**] (SCM, Struck, TLD, ASLA, CXT, VTD, VTS, CSK) concludes

    (a) **Background Information** background information is critical for effective tracking. It can be ex- ploited by using advanced learning

techniques to encode the background information in the discriminative model im- plicitly (e.g., Struck), or serving as the tracking context explicitly (e.g., CXT)

(b) **local models** are important for tracking as shown in the performance improvement of local sparse representation (e.g., ASLA and SCM) com- pared with the holistic sparse representation (e.g., MTT and L1APG). They are particularly useful when the appearance of target is partially changed, such as partial occlusion or deformation.

(c) **local models** motion model or dynamic model is cru- cial for object tracking, especially when the motion of target is large or abrupt. However, most of our evaluated tracker- s do not focus on this component. Good location predic- tion based on the dynamic model could reduce the search range and thus improve the tracking efficiency and robustness.

2.1.1 Pre-CVPR 2013 Trackers

Online RGB Trackers require no prior knowledge of the object, and only a bounding box of the target on the original frame. A survey of tracking methods [**?**] (SCM, Struck, TLD, ASLA, CXT, VTD, VTS, CSK), as well as papers following the survey [**?**], show the following methods are competitive for this problem:

1. **Struck** [**?**] uses a kernelized structured output SVM to directly learn displacement vectors. Gaussian kernel on 192 haar-like features.

2. **SCM** [**?**]

3. **TLD** [**?**] "We develop a novel learning method (P-N learning) which estimates the errors by a pair of experts: (i) P-expert estimates missed detections, and (ii) N-expert estimates false alarms. The learning process is modeled as a discrete dynamical system and the conditions under which the learning guarantees improvement are found. We describe our real-time implementation of the TLD framework and the P-N learning.

4. **APG-L1** [**?**] ”l1 norm related minimization model”.

5. **MIL** [**?**] Older well-known tracker using multiple instance learning.

6. **CXT** [**?**] uses background context

7. **ASLA** [**?**]

8. **Circulant** [**?**] uses fourier transformed graham matrix to improve The fastest tracker in [**?**].

2.1.2 Post-CVPR 2013 Trackers

After Wu et. al's survey [**?**], the following notable methods were also published:

1. **Self-paced learning** [**?**] ”we show that an accurate appearance model is considerably more effective than a strong motion model”.

2. **MEEM** [**?**] claims state-of-the-art over Struck, SCM, MIL.

3. **Xiang** [**?**]

4. **Occlusion and motion reasoning for long-term tracking** [**?**] ”Struck fails in the presence of long-term occlusions as well as severe viewpoint changes of the object. In this paper we propose a principled way to combine occlusion and motion reasoning with a tracking-by-detection approach.”

5. Color tracker [**?**] - the **best**.

2.2 Online RGB-D Trackers (vision papers)

Online RGB-D Trackers no prior knowledge of the object. Song et. al's survey [**?**], show **incorporating depth into tracking beats the state of the art.** It shows **Struck** [**?**] and VTD are competitive.

Gaussian Process Regression [**?**] beats struck and Song et. al's survey benchmarks.

2.3 Trackers (robotics papers)

1. RSS 14: Anytime Tracking [**?**]

2. RSS 14: DART Pose Estimation [**?**] (relevant?)

3. ICRA: Small Obstacle Discovery over Images [**?**]: small object segmentation using RGB, since depth info doesn't give much.

4. ICRA 14: Tracking ping pong balls [**?**] This paper proposes a way to observe and estimate ball's spin in real-time, and achieve an accurate prediction. Based on the fact that a spinning ball's motion can be separated into global movement and spinning respect to its center, we construct an integrated vision system to observe the two motions separately. With a pan-tilt vision system, the spinning motion is observed through recognizing the position of the brand on the ball and restoring the 3D pose of the ball. Then the spin state is estimated with the method of plane fitting on current and historical observations. With both position and spin information, accurate state estimation and trajectory prediction are realized via Extended Kalman Filter(EKF).

5. ICRA 14: road scene segmentation [**?**] "we first produce initial object hypotheses by clustering the sparse 3D point cloud. The image pixels registered to the clustered 3D points are taken as samples to learn each object's prior knowledge. The priors are represented by Gaussian Mixture Models (GMMs) of color and 3D location information only, requiring no high-level features. We further formulate the segmentation problem within a Conditional Random Field (CRF) framework, which incorporates the learned prior models, together with hard constraints placed on the registered pixels and pairwise spatial constraints to achieve final results. "

6. ICRA 14: Learning latent structure for activity recognition [**?**]. Good overview of basics.

2.4 Temporal Speech Data

Structured prediction is used in temporal data such as speech recognition. This has yet to fully permeate object tracking in videos.

RGB-D videos provide a unique set of data to images. Objects are more easily segmented based on depth data (without transformation) alone. This can be seen in LIDAR [**?**].

It is often the case that

2.5 RGB-D Datasets

1. Princeton RGB-D [**?**]
2. Wu et al. [**?**]
3. Bigbird [**?**] (relevant?)

2.6 RGB-D Features

1. Learning rich features from rgb-d images for object detection and segmentation [**?**]
2. Segmentation using RGB-D data [**?**]

2.7 Early Tracking

Tracking pre-2005ish.

CONDENSATIONConditional Density Propagation for Visual Tracking: "The problem of tracking curves in dense visual clutter is challenging. Kalman filtering is inadequate because it is based on Gaussian densities which, being unimodal, cannot represent simultaneous alternative hypotheses. The CONDENSATION algorithm uses "factored sampling", previously applied to the interpretation of static images, in which the probability distribution of possible interpretations is represented by a randomly generated set. CONDENSATION uses learned dynamical models, together with visual observations, to propagate the random set over time.

The result is highly robust tracking of agile motion. Notwithstanding the use of stochastic methods, the algorithm runs in near real-time."

http://www.cse.psu.edu/ rcollins/CollinsVLPR2012Lecture.pdf http://www.cse.psu.edu/ rcollins/-CollinsVLPR2009Lecture.pdf

Chapter 3

Experiments