Boosting and Switching experts for object tracking

February 11, 2014

1 Introduction

There has been significant success with using boosting, specifically "online boosting" for object tracking in video []. While there has been significant success, no theoretical guarantees have been associated with this mkethod. It is therefor difficult to understand why the method fails when it does and how to improve it to reduce failures.

In this paper we construct an alternative algorithm to that of [], and give theoretical guarantees to it's performance.

Tracking algorithms are composed of two main parts: an appearance model and a dynamics model. In this paper we concentrate on the appearance model.

The appearance model is a *score function* that is used to identify the location of the tracked object. The score distribution we would like to have is a sharp peak surrounded by much lower scores. In order to have confidence in the location of the peak we want the difference between the max value and the values far from the max to be *statistically significant*. Note that statistical significance is used here as a stand in for "ground truth" which is not usually available.

We propose two complementary methods for combining score functions:

- **Boosting:** We use boosting to combine so-called *weak score functions* into a single *strong score function*

- **Sleeping experts:** We use an online algorithm for *switching among a small set of experts* to allow different scoring functions to be used at differen times, and to combine different trackers that are centered at different locations and/or use multiple resolutions. Note that the experts in this part are equivalent to the *strong* score functions of the boosting method.

The rest of the paper is organized as follows. We start by describing the method for boosting score functions. We then give the details of the boosting process and it's justification. We then describe details that are specific to the tracking process. Finally we describe how sleeping experts are used to combine different strong score functions.

2 Boosting significance

A score function F is mapping of a region of an image which we call the *appearance window* to a real number. Let $w(\vec{x})$ denote the appearance window (for example, a 20×20 grey-level matrix) centered at the location defined by the vector $\vec{x}$. The score that the score function F associates with the location $\vec{x}$ is denoted by $F(w(\vec{x}))$.

Given an image window, we quantify the performance of the score function as follows:

1. Find the location where the score is maximized, denote the location by $\vec{x}_{\text{max}}$ and the window at that location by $w_{\text{max}} = w(\vec{x}_{\text{max}})$. The maximal score is therefor $s_{\text{max}} \doteq F(w_{\text{max}})$.[1]

2. compute the score for all locations $\vec{x}$ in the image window which are sufficiently far from $\vec{x}_{\text{max}}$. We call the window that achieves the highest score in this set the *decoy* window because this window is the most likely one to distract the tracker

   $$s_{\text{decoy}} = \max\left\{F(w(\vec{x}))\,\|d(\vec{x},\vec{x}_{\text{max}}) > r\right\}$$

   The performance of the scoring function is measured by $s_{\text{max}} - s_{\text{decoy}}$. In order to see if this difference is significant we need to compare it to a "Null" distribution. We construct this null distribution as follows:

   - We estimate the standard-deviation of the score function F using the empirical variance of the score function F in the image window.
   - We make the assumption that the scores $F(w)$ are drawn IID from a normal distribution with some mean μ and the estimated standard deviation.
   - Suppose the total number of locations for which the score is calculated is N. The null distribution over the score difference $s_{\text{max}} - s_{\text{decoy}}$ is the distribution resulting from taking N IID samples from the normal distribution and then considering the difference between the largest and the second largest values. We need to find out what is the form of this distribution.

3 Definition of the visual regions

We now define the goal of the apearance model training algorithm.

We denote $w(\vec{x})$ the appearance window centered at the location $\vec{x} = (x, y)$. The point $\vec{x}$ is constrained to a region of the image that we call the **tracking window**. See Figure 3 for a depiction of the tracking region and the other regions defined below.

We score all windows whose center is inside the tracking region and define the window with the maximal score as the **max window** and denote it by w_{max}. We then define a region around the center of the max window which we call the **center region**. The center region corresponds, intuitively, to the desired shape of the peak. More on that below.

The training examples are defined as pairs of windows. The first window in each pair is the max window w_{max}. The second is a window $w(\vec{x})$ such that $\vec{x}$ is not in the center region. The score of a pair

[1] It might be better to use the maximum of an *average* over a small neigborhood of scores such as a 2×2 or a 3×3 square.

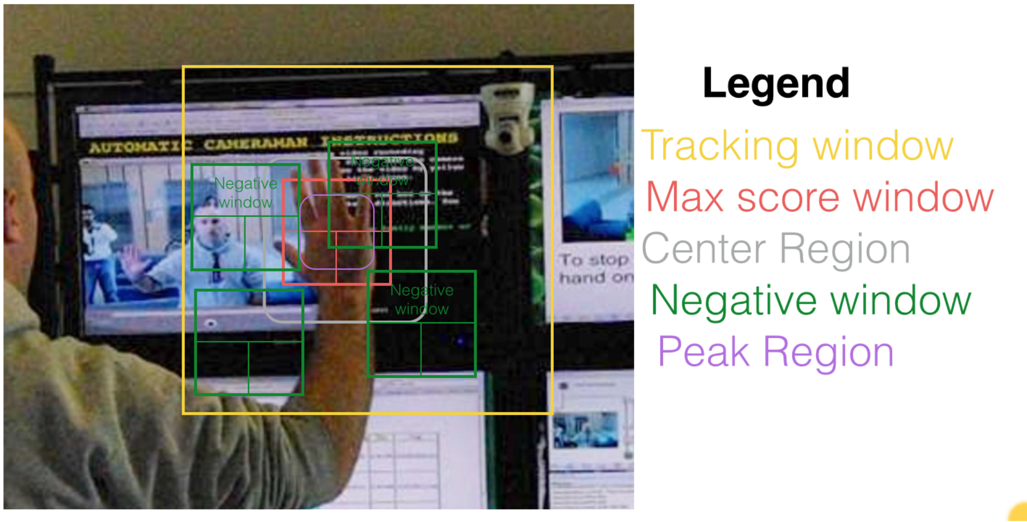

Figure 1: The windows involved in training the tracker

3.1 The effect of the size of the center region

Choosing the size of the center region is a tradeoff between accuracy and robustness. Setting it to be small will generate a tracker which can localize more accurately, possibly at the cost of failing to track more easily. On the other hand, setting the center window to be large, will generate a tracker that is less likely to loose tracking, but the location of the tracking point might slide from place to place on the target. A combination of sizes and resolution is probably the best way to create a tracker that is both robust and accurate.

We now come to the first novelty in our approach. In the standard framework for boosting the training set contains the *ground truth* label for each instance. However, in the tracking problem no such ground truth is available. Instead, we propose an approach which measures the quality of a base function using the *statistical significance* of the outputs of the function.

To define statistical significance we need to first define a *null hypothesis*. Our null hypothesis is that the values of $F(w)$ are independently drawn from a normal distribution with unknown, but fixed, mean and variance. We will be looking for a function F which defines a sinstatistically significant

whose value are statistically significant and which is *coherent* with the current scoring function.

We denote by $w(x, y)$ the window around the center point (x, y). We use the euclidean distance:

$$d((x_1, y_1), (x_2, y_2)) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

We define a "minimum distance" parameter d_{min} which will be around the size of the window w. Setting d_{min} smaller will make for a more accurate and less robust scoring function.

4 Boosting and Sanov

Let the input space be $X = \{-1,+1\}^d$, where d corresponds to the number of features or weak hypotheses. The label is $Y = \{-1,+1\}$. We are given a set of m training examples $(\mathbf{x}_i, y_i) \in X \times Y$, which defines the uniform empirical distribution U that is supported on the m training examples.

Using totalboost is equivalent to finding a distribution D over these m points such that for all $1 \le j \le d$, $\sum_{i=1}^m D(i) x_i^j y_i = 0$, and $RE(D||U)$ is minimized.

This looks similar to Sanov's theorem where D defines the "true" distribution, which is a distribution under which none of the x^j are correlated with y ($\sum_i x_i^j y_i \neq 0$) and U is the empirical distribution.

More precisely, Sanov's theorem states the following.

Let $\mathcal{H}$ be a finite domain, In our case $\mathcal{H} = \{-1,+1\}^{d+1}$. The "type" or "empirical distribution" of a sample of size n is a vector of length $|\mathcal{H}|$ where each entry is the count of the number of occurances of each element in $\mathcal{H}$. Define $\mathcal{P}_n$ to be the set of types of size n.

Let Q be a distribution over $\mathcal{H}$. Note that Q is a point in the $|\mathcal{H}|$-dimensional simplex $\Delta^{|\mathcal{H}|}$. Let E be a subset of $\Delta^{|\mathcal{H}|}$ that does not contain Q. Sanov's theorem gives an upper bound on the probability that the empirical distribution of a sample of size n is in E:

$$Q^n(E) \quad \le \sum_{P \in \mathcal{P}_n \cap E} 2^{-nD(P||Q)} \le |\mathcal{P}_n| 2^{-nD(P^*||Q)} \le (n+1)^{|\mathcal{H}|} 2^{-nD(P^*||Q)}$$

Where $P^* \mathrm{argmin}_{P \in E} D(P||Q)$.

The set E represents the null hypothesis, which is the hypothesis that states that the performance of the weak rule is not any better than a random guess. In classification tasks this corresponds to predicting the correct label with probability 1/2. In the case of the scoring functions it means that difference between the maximal score and the decoy score is consistent with that of considering the difference between the largest and second largest elements in a sample of size N.