Visual Tracking with Online Multiple Instance Learning

Ming-Hsuan Yang University of California at Merced mhyang@ucmerced.edu

Visual tracking has many practical applications (e.g., surveillance, HCI) and has been long studied in computer vision. Although there has been some success with building domain specific trackers (e.g., faces, humans, tracking generic objects has remained very challenging. A major challenge that is often not discussed in tracking literature is how to choose positive and negative examples when updating the discriminative appearance model. Most commonly this is done by taking the current tracker location as one positive example, and sampling the neighborhood around the tracker location for negatives. If the tracker location is not precise, however, the appearance model ends up getting updated with a sub-optimal positive example. Over time this can degrade the model, and can cause drift. On the other hand, if multiple positive examples are used (taken from a small neighborhood around the current tracker location), the model can become confused and its discriminative power can suffer (Fig. 1 (A-B)).

As the *exact* object locations are unknown, algorithms that deal with inherent ambiguities are likely to render better tracking results. The Multiple Instance Learning (MIL) has been proposed to address such problems in machine learning. The basic idea of this learning paradigm is that during training, examples are presented in sets (often called "bags"), and labels are provided for the bags rather than individual instances. If a bag is labeled positive it is assumed to contain at least one positive instance, otherwise the bag is negative. For example, in the context of object tracking, a positive bag could contain a few possible bounding boxes around each tracked object Therefore, the ambiguity is passed on to the learning algorithm, which now has to figure out which instance in each positive bag is the most "correct".

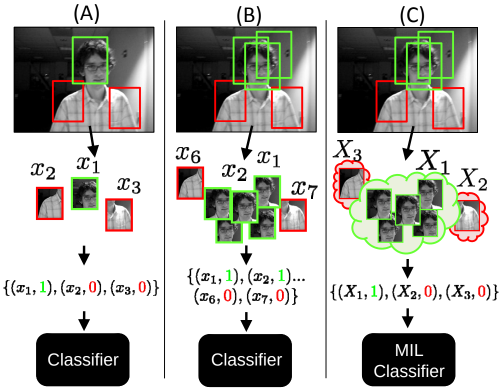
(A)
(B)
(C)
x_2
x_1
x_3
$\{(x_1, 1), (x_2, 0), (x_3, 0)\}$
Classifier
x_6
x_2
x_1
x_7
$\{(x_1, 1), (x_2, 1) \ldots (x_6, 0), (x_7, 0)\}$
Classifier
X_3
X_1
X_2
$\{(X_1, 1), (X_2, 0), (X_3, 0)\}$
MIL Classifier

Figure 1: **Updating a discriminative appearance model:** (A) Using a single positive image patch to update a traditional discriminative classifier. The positive image patch chosen does not capture the object perfectly. (B) Using several positive image patches to update a traditional discriminative classifier. This can confuse the classifier causing poor performance. (C) Using one positive bag consisting of several image patches to update a MIL classifier.

We present a MIL based appearance model for object tracking (Fig 1 (C)). As in the object tracking domain there is even more ambiguity than in object detection because the tracker has no human input and has to bootstrap itself, we expect the benefits of a MIL approach to be even more significant than in the object detection problem. In order to implement such a tracker, an online MIL algorithm is required. We present empirical results on challenging video sequences, which show that using an online MIL based appearance model can lead to more robust and stable real-time tracking than existing methods in the literature.