Master Thesis

Visual Object Tracking Applying Online Ensemble of Multiple Trackers

Author:
Jorge Martinez Gomez

Supervisor:
Juan Carlos Niebles

A thesis submitted in fulfillment of the requirements for the degree of Master of Science

in the

Robotics and Intelligent Systems Research Group
Electrical and Electronics Engineering Department
Universidad del Norte

June 2015

UNIVERSIDAD DEL NORTE

Abstract

Electrical and Electronics Engineering Department
Robotics and Intelligent Systems Research Group
Computer Vision Laboratory

Master of Science

Visual Object Tracking Applying Online Ensemble of Multiple Trackers

by Jorge Martinez Gomez

The object tracking literature offers a large variety of tracking methods, which exhibit complementary properties in terms of their performance, best usage scenarios and failure modes. In this thesis, we introduce a new tracking algorithm based on an online ensemble of tracking algorithms. Our method runs multiple online trackers in parallel and fuses their outputs in an online fashion. The resulting tracker can leverage the strenghts and overcome failures of each individual tracker, producing more robust target tracking. We perform experiments on current object tracking benchmark and show how our ensemble consistently outperforms all trackers in the ensemble, and achieves state-of-the-art object tracking performance.

Acknowledgements

> No one is dumb who is curious. The people who don't ask questions remain clueless throughout their lives.
>
> Neil deGrasse Tyson

Phew!, I feel so relieved to fulfill this thesis without a sip of red bull or other similar beverages, except for coffee. This is a significant accomplishment in my life and it would be impossible without all the people who supported and believed in me.

I like to give my sincere thanks and extend my gratitude to my honorable supervisor Prof. Juan Carlos Niebles. I sincelery thank him for an exemplary guidance and encouragement. His endless support inspired me to make the right decisions in the most important moments and I am glad to work under his supervision. I also want to thank Bohyung Han from POSTECH and all thesis proposal reviewers for their supportive feedback.

I would like to thank my friends and specially to all VisionLab members for all the thoughtful and mind stimulating conversations we had, which helped me to think wisely and outside the box. I've enjoyed the companionship that you all have brought to me during these years.

Last, but not least, I would like to thank my parents, who taught me the value of hard work using themselves as an example. Giving me enormous support during my whole life, they steered me into the right path and served me as a guidance. I feel very thankful for having them as parents.

Contents

List of Figures

List of Tables

Dedicated to my family

Chapter 1

Introduction

Keywords: Computer vision, object tracking, object model, spatial clustering, inliers, outliers, ensemble.

The goal of visual object tracking is to estimate the state of a target in an image sequence. This is a difficult task, as the target object can be articulated or deformable, the scene illumination can change suddenly, background clutter may introduce distractions that result in tracker drifting, among others. In spite of the multiple challenges, there are many potential applications that make this capability attractive such as activity recognition, motion analysis, human surveillance and robotics.

Many approaches for object tracking have been proposed to cope with some of these challenges. While the state-of-the-art methods achieve relative success, there is still no single approach that is able to handle all challenging situations. For instance, tracking-by-detection methods may not be able to handle scale variations rigorously [76]. On the other hand, generative methods tend to suffer from model drifting and struggle to handle appearance variations [49].

In this thesis, we focus on "model free tracking" of arbitrary objects in videos, in which no prior knowledge other than the object location in the first frame is available. The online tracking benchmark recently proposed in [1] shows that each tracking algorithm performs best under particular circumstances. There is no single tracking algorithm that can perform well on all sequences in the benchmark. This indicates that each tracking challenge can be addressed better by a different algorithm. In other words, tracking strenghts may be distributed among the available trackers. This is the key observation that inspires our proposed method; we consider a tracking approach that combines the outputs of multiple trackers running in parallel via an online ensemble. This ensemble has the interesting property of leveraging the strenghts of individual trackers, while

overcoming the failure modes of each tracker. Since for a new and unseen sequence we do not know which tracker would perform best, our method computes a data-driven online ensemble that results in improved tracking performance when compared to the results of individual trackers.

In our method, we leverage the observation that only some of the trackers drift into non-target areas of the image in most cases while some of the trackers succeed by focusing on the correct target. Furthermore, our ensemble uses an appearance model that serves as an additional verification mechanism of the tracked region. Using these model components, we identify and exploit the successful trackers to steer failed trackers towards the correct target region. Effectively, our ensemble can correct failed trackers, which ultimately increases tracking performance.

1.1 Goals

Our main goal is to design and implement a visual object tracking system that performs an ensemble of multiple tracking algorithms. We achieve this goal via several milestones. First, we aim at selecting a set of standarized sequences, necessary for the analysis, development and tests of the proposed system. Also, it is important to select a group of trackers that can store their state on each frame and be reinitialized. Then, we seek to develop an ensemble of multiple tracking algorithms. Finally, we plan to test and verify our approach using performance quantification.

1.2 Contributions

The main contribution of this thesis is an ensemble tracking framework that builds on top of the output of available online tracking algorithms. Running in parallel several trackers, the outcome corresponds to an online fused tracking result that leverages each tracker best features. Our method does not use prior knowledge about the nature of the trackers in the pool. The fused tracking output is obtained by considering appearance and spatial relations among tracker outputs. In order to cope with trackers weaknesses, our ensemble identifies successful trackers in a data-driven fashion and uses them to steer failed trackers by restarting them asynchronously. This helps to avoid sequence dependent parameters and overtuning.

1.3 Thesis overview

The remainder of this thesis is organized as follows. We first briefly review the state-of-the-art of tracking algorithms in Chapter 2 and then present our online ensemble tracking algorithm in Chapter 3. Chapter 4 illustrates quantitative and qualitative results of our tracker on a standard benchmarking dataset. Finally, we conclude the thesis in Chapter 5.

Chapter 2

Related Work

In this chapter, we provide a review of state-of-the-art tracking methods. Numerous approaches for object tracking have been proposed before. These methods differ from each other based on the way the authors solve common questions, such as: Which object representation is suitable for tracking the target? How motion, appearance, or shape of the object are modeled? The answers for these questions and some others depend on the context/environment where tracking is performed and the visual features of the object that needs to be tracked. This chapter focuses on reviewing methodologies for object tracking in general and not for specific objects, such as humans or faces.

We cover several important aspects that should be addressed in object tracking systems. In section 2.1 we describe common object representations and image features. Section 2.2 summarizes general schemes for detecting objects in an scenario. In Section 2.3, we categorize and describe existing tracking methods. Section 2.4 shows existing datasets used for object tracking and explains challenges which are present in different video sequences. Finally, we show in Section 2.5 recent tracking methods that perform ensemble of multiple trackers or features.

2.1 Object representation and visual features

An object can be considered as an entity of interest for further analysis. For instance, birds in the sky, pedestrians or vehicles on a road, ships on the sea, are set of objects of interest for tracking in different contexts. These elements can be represented by their shape or appearance. According to the object to be tracked, an object representation may be preffered. For tracking small objects, point representation is usually appropiate [2, 3]. In the case of tracking objects whose shapes are approximated to rectangle and

ellipses, shape representation is more appropiate [4]. To track objects with complex shapes, such as humans or excavators machines, contour or silhouette-based representation is appropiate [5].

In order to track objects, selecting the right features plays a critical role. Researchers usually choose visual features manually depending on the application domain. The problem of automatic feature selection has received significant attention in the pattern recognition community. In general, a feature is a property of an image which we are interested in. The most important property of a visual feature is its discriminative power so that the object of interest could be easily distinguished from others. The most common visual features selections are color [6, 7], edges [8, 9], displacement vectors [10, 11], corners [12] and textures [13–15]. Among all features, color is one of the most widely used for tracking. However, color features are sensitive to illumination variation. To tackle this problem, in scenarios where this issue is inevitable, other features are incorporated to model the appearance of an object.

2.2 Moving Object Detection

Detection can be defined as finding instances of objects in images or videos. Some tracking approaches require an object detection method as initialization only. Some others, make use of an object detector over all frames in ofrder to track objects. A large number of methodologies have been proposed for object tracking, focusing first on the task of object detection. Most of them apply combinations among different methodologies, making it very difficult to create a uniform classification of existing approaches. This section classifies different approaches available for object detection from videos.

2.2.1 Background Subtraction

Background subtraction is a commonly used technique for object segmentation in static camera scenarios [16]. This task consists in detecting moving regions by subtracting the current image pixel-by-pixel from a reference background image. The pixels above some threshold are classified as foreground, which means pixels belonging to the object of interest (Figure 2.1). The background image is created averaging images over time in an initialization period, and is updated with new images to adapt to dynamic scene changes. Also, the foreground map is followed by morphological operations such as closing and erosion, which eliminate small-sized blobs.

Although background subtraction techniques extracts well most of the relevant pixels, this method is sensitive to changes when some background and foreground pixels have similar value.

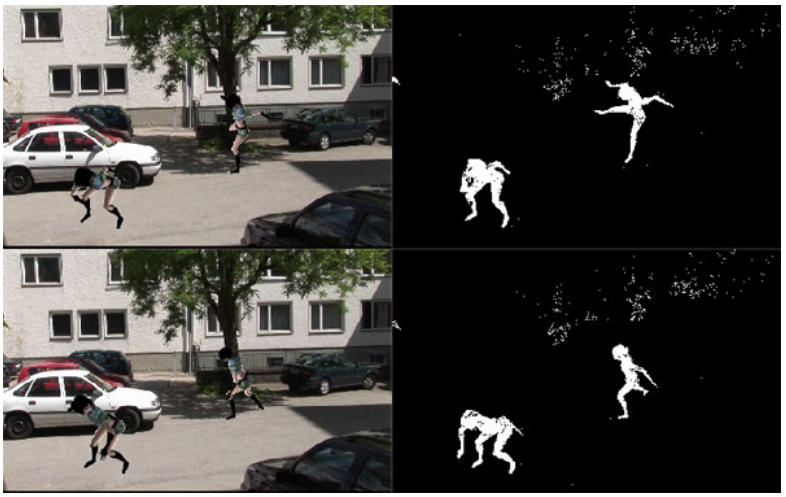


Figure 2.1: Object detection using Gaussian Mixture Models for background subtraction. Foreground pixels are drawn in white. Figure reproduced from [17].

2.2.2 Temporal differencing

In temporal differencing, objects are detected by taking pixel-by-pixel difference of consecutive frames in a video sequence. This method is most common for object detection in scenarios where camera is moving. Unlike static camera scenarios, the background is changing rapidly. Therefore it is not appropiate to create a background model. Instead, the moving object is detected by taking the difference between frames $t-1$ and t.

This method is highly adaptive to dynamic changes in the scene as most recent frames are involved in the process. However, it fails detecting new small regions as moving objects, known as ghost regions. Detection will not be correct either for objects which remain static in a video.

A two-frame differencing method is presented in [18], where the pixels that satisfy the following equation are marked as foreground.

$$|I_t(x,y), I_{t-1}(x,y)| > Th$$

Other methods were developed in order to overcome drastic changes of two frame differencing. For instance, a three-frame differencing method [19] and a hybrid method which combines three-frame differencing with an adaptive background subtraction model [20].

2.2.3 Statistical Approaches

Statistical characteristics of pixels have been used, in order to overcome shortcomings between frames of basic background subtraction methods. The approaches consist in keeping and updating a statistical model for the pixels that belong to the background model. Foreground pixels are identified by comparing each pixel's statistics with that of the background model. These methods are becoming more popular due to its reliability in scenes that contain noise, illumination changes and shadows. Some approaches apply Hidden Markov Models (HMM). These methods [21, 22] represent the intensity variation of a pixel in an image sequence as discrete states.

The statistical method proposed in [17] describes an adaptive background model for real-time tracking. Every pixel is modeled by a mixture of Gaussians which are udpated online using incoming image data. Then, the Gaussians distributions of the mixture model for each pixel are evaluated in order to detect whether a pixel belongs to foreground or background.

2.2.4 Point detectors

Point detectors are used to find interesting points in objects which have an expressive texture in their respective localities. An interest point should have invariance to changes in illumination and camera viewpoint. One important detector uses optical flow (KLT) approach [23]. This method makes use of the flow vectors of moving objects over time to detect moving blobs in an image. In this approach the apparent velocity and direction of every pixel in the frame must be computed. Some other methods are SIFT [24] and Harris [12] corners detectors (Figure 2.2).

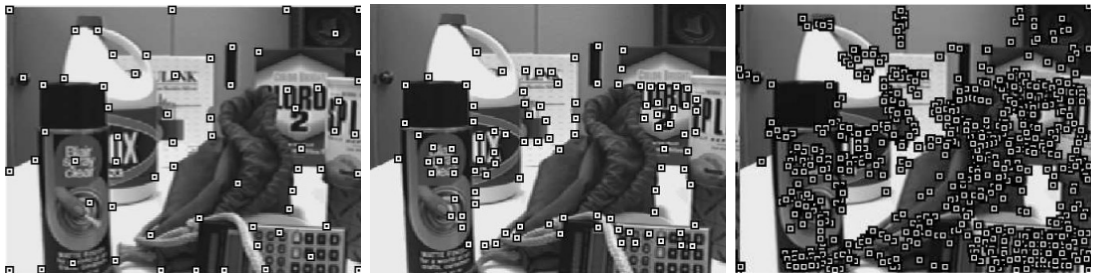


Figure 2.2: Interest points for object detection. Left - Harris, center - KLT, right - SIFT. Figure reproduced from [25].

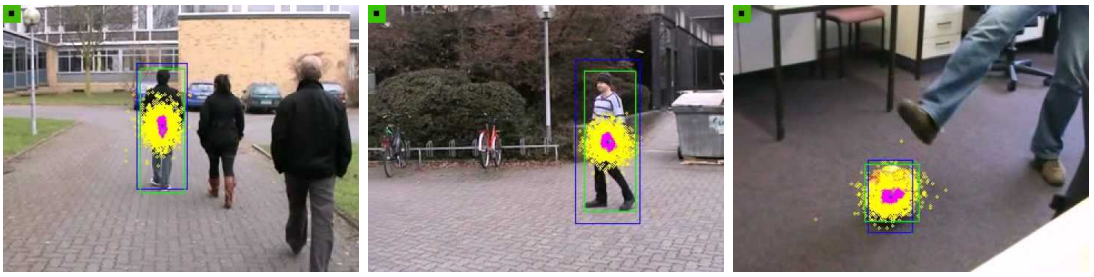


Figure 2.3: Object tracking using particle filter. Each particle represents one possible location of the object being tracked. A set of particles have more weight at locations where the probability is very high. Figure reproduced from [22].

2.3 Object Tracking

The goal of an object tracker is to generate an object trajectory in a video. This path consists of the object position across the time. Additionally, a tracker may provides other information, such as scale, orientation, area, or shape of an object. This section classifies and covers popular approaches for object tracking on each category.

2.3.1 Point Tracking

Tracking can be formulated as the correspondence of objects represented by points across frames. This category can be divided into two subcategories:

Deterministic Methods. These approaches for point correspondence define a cost of associating each object in frame $t-1$ to a single object in frame t using motion constraints, such as proximity, velocity, rigidity and motion. Minimization of the correspondence cost is formulated as a combinatorial optimization problem. A solution, which consists in one-to-one correspondence among all possible associations, can be obtained by optimal assignment methods. For instance Hungarian Algorithm [26] or greedy search methods.

Statistical methods for Point Tracking. Statistical correspondence methods solve tracking problems whose measurements obtained from video sensors contain noise, or where object motion can undergo random perturbations. These approaches take measurements and model uncertainties into account. During object state estimation, these methods apply state space approach to modeling properties such as position, velocity and acceleration. In single object state estimation, the optimal state of an object is given by a Kalman Filter [27, 28], assuming measurement noise have a Gaussian distribution. In the general case, where noise is not assumed as Gaussian, estimation can be performed using particle filters (Figure 2.3) [22, 29].

2.3.2 Kernel Tracking

In this type of tracking, the object motion is computed using representations of a primitive object region, from one frame to the next. These algorithms differ in terms of appearance representation, the number of objects to be tracked, and the method used for object motion estimation.

Density-based tracking: According to [30], the object is modeled with one or more probability density functions, such as Gaussian, mixture of Gaussian, Parzen windows or histograms, that describe the probability of object appearance. Mean-shift is an approach of density-based tracking. This method shifts a data point to the average of data points in its neighborhood, using fixed color distribution. A similar approach is CAMSHIFT [31] that handles dynamically changing color distribution by adapting the search window size and computing color distribution in the window.

Template-based tracking: These approaches apply templates of the object to calculate appearance probability on every frame of the video sequence. The most common is *Template matching* [32] that searches accross the image, a region similar to the object template, defined in previous frames. A similarity measure is calculated using normalized cross correlation. A limitation of this method is its high computational cost due to brute force search. To reduce this cost, some methods limit the object search to a neighborhood 2.4.

Instead of templates, other object representations can be used for tracking. For example, color histograms or mixture models can be computed using the appearance of pixels inside a rectangular or ellipsoidal region. To reduce computational complexity, the similarity between object model and the hypothesized position, is computed evaluating the ratio between color means of object model and position [82]. The position with highest ratio is selected as current object location.

Tracking by detection [34] systems generally perform target object appearance learning. These methods are closely related to object detection and has encouraged some successful real-time tracking algorithms [33, 35]. However, many tracking algorithms employ static appearance models that are defined manually or trained at the first frame only [10, 36–39], these methods are often unable to deal with significant appearance changes. In order to cope with this problem, an adaptive appearance model that changes during the tracking process as the appearance of the object changes gets better results [40–42].

Boosting has been used in a wide field of machine learning tasks and applied to computer vision problems. Many tracking algorithms are based on the boosting framework [43] and is related to the work on Online Adaboost [44–46], multi-class boost [47] and MILBoost

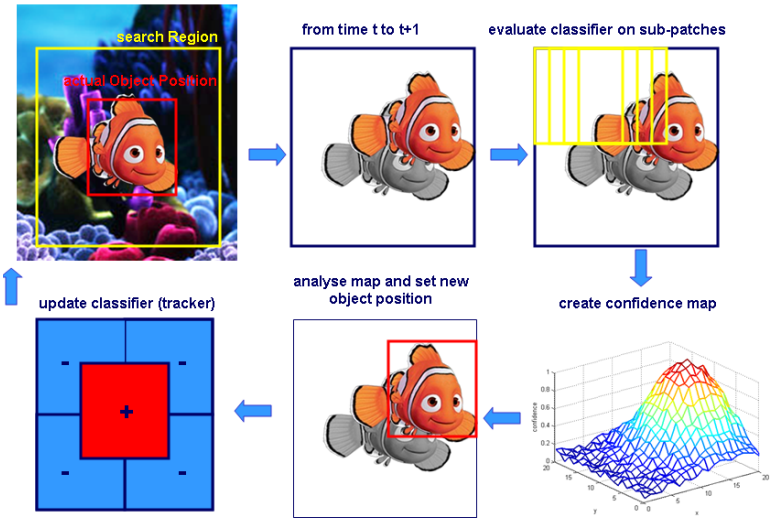


Figure 2.4: Overview for online boosting object tracker. For each classified patch, the system receives a confidence value, that is entered into a confidence map. Using the confidence map, the tracking windows is shifted to the best possible position. Then, the classifier is updated and the process is repeated. Figure reproduced from [33].

[48]. The goal of boosting is to combine many weak classifiers (usually decision stumps) into a linear strong classifier.

Sparse/non-sparse representation: In this type of tracking, a set of target samples is associated with serveral templates. The likelihood of a candidate sample belonging to the object class is often determined by the residual between the candidate sample and the reconstructed samples derived from a linear representation [49–53].

2.3.3 Silhouette Tracking

The object is tracked via estimation of the object region in each frame. Silhouette-based methods provide an accurate shape description for the objects that are tracked (Figure 2.5) . These approaches can be divided into two main categories, shape matching and contour tracking. Shape matching [54] approaches search object silhouette in the current frame. Contour based, evolve initial contour to its new position using state space models or direct minimization of an energy function [55–57].

2.4 Tracking Datasets and Challenges

Tracking Challenges. Object detection and tracking is still an open research problem in computer vision. The level of difficulty depends on how the object of interest is

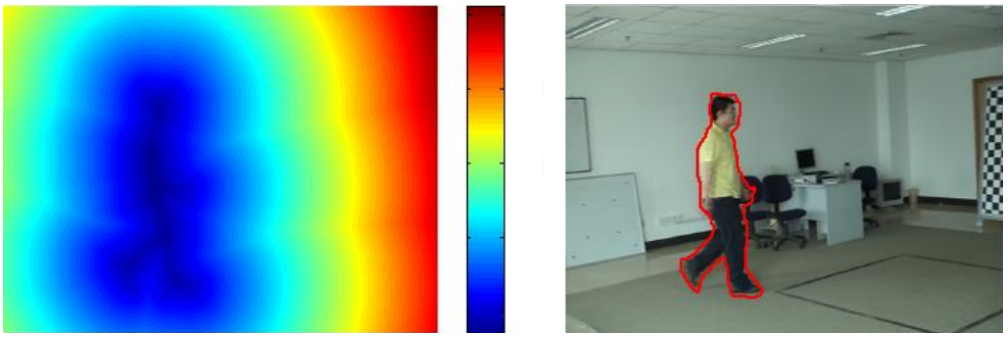


Figure 2.5: Illustration of an active contour representation. Left subfigure shows the signed distance map of a human contour; right image displays contour result. Figure reproduced from [25].

defined in terms of features, and particular challenges which appear in a video sequence. For instance, using color as object representation method, it is not difficult to identify all pixels with same color as the object. However, there are sequences where the values of pixels for the object and the background, are very similar. In addition, illumination changes in the scene do not guarantee that the pixel values of an object will be the same in all frames. These variabilities or challenges cause failures, listed below.

- **Illumination Variation (IV):** The illumination in the target region changes significantly.
- **Scale Variation (SV):** Ratio between initial object size and current object size differs.
- **Occlusion (OC):** Partially or full, occlusion affects the process of computing the background frame. In real life situations, occlusion can occur anytime the object of interest passes behind another object with respect to a camera.
- **Dynamic background:** Some scenery regions contain movement, but should be still remain as background, acoording to their relevance. Such movement can be periodical or irregular, causing blurring (motion blur - MB), e.g. traffic lights, waving trees).
- **Out of view (OV):** Some portion of the target leaves the field of view.
- **Background clutter (BC):** As stated before, this challenge makes the segmentation task difficult. It is hard to create a separate background model from moving foreground objects.
- **Fast Motion (FM):** The speed of a moving object plays an important role in its detection and track. If an object is moving too slow, the temporal differencing methods fails to detect object, because it preserves uniform region between frames.

On the other hand, fast moving object leaves ghost regions in a detected foreground model.

- **Object rotation and deformation (DEF):** Since natural objects move freely, they can appear slightly or completely transformed. Such rotations, in (IPR) or out (OPR) of plane, on the images affect object tracking considerably.
- **Low Resolution (LR):** Number of pixels inside the object bounding box is less than 400.

Tracking Dataset. In computer vision, a *dataset* could be defined as a collection of images or video sequences used for testing algorithms. The amount of data and characteristics presented, depend on the field that is studied. For instance, in scene recognition, a dataset contains images of landscapes or outdoor environments. Generally, this collection is shared between researchers and plays an important role in comparison and evaluation of state-of-the-art approaches. A list of datasets used in object tracking is summarized in table 2.1.

The Surveillance Performance Evaluation Initiative (SPEVI) [58] can be used for evaluating algorithms for surveillance-related applications. The first dataset contains 5 sequences applied to single person/face detection and tracking. The second dataset applies for multiple person/face detection and tracking. The sequences contain four targets occluding each other repeatedly. The ETISEO dataset [59] contains indoor and outdoor scenes, such as corridors, buildings entries, etc. This dataset can be used for surveillance applications.

The PETS dataset [60] became a surveillance project whose challenging scenarios are focused only on high level applications on this field. Some issues, like illumination or scale changes are not considered in these videos. Most of the sequences are used for person/vehicle tracking in outdoor environments(subway stations, building entrances). CAVIAR [61] is a dataset used generally for situation recognition systems. However, sequences can be applied for tracking evaluation methods. Includes videos of people walking alone, meeting other people, entering and exiting shops.

The VIdeo Surveillance Online Repository (VISOR) [62] database covers a wide range of scenarios and situations, including videos for human action recognition, outdoor videos for face detection, indoor videos for people tracking with occlusions, vehicles detection and surveillance. VISOR, includes several sequences for two separate tasks: First, an abandoned baggage scenario and second, a parked vehicle scenario.

Recently, the tracking community released evaluation suites containing a selection of videos and algorithms for testing trackers performance. These benchmarks test and

Dataset	# Sequences	GT-Available	Object
Bobot [64]	12	Yes	Generic
Cehovin [65]	5	Yes	Generic
Kalal [66]	10	Yes	Generic
Kwon [67]	4	Yes	Generic
Kwon VTD [68]	11	Yes	Generic
PROST [69]	4	Yes	Generic
Ross [40]	4	Yes	Generic
Thang [70]	4	Yes	Generic
Wang (NoRef)	4	Yes	Generic
Tracking Benchmark [1]	50	Yes	Generic
ALOV 300+ [63]	315	Yes	Generic
Godec [71]	7	Yes	Human
Babenko [48]	3	Yes	Human
SPEVI [58]	3,5	Yes	Human
ETISEO [59]	86	Yes	Human
PETS [60]	23	Yes	Human
CAVIAR [61]	25	Yes	Human
Clemson [72]	16	Yes	Human
VISOR [62]	6	No	Human

Table 2.1: Object tracking datasets.

compare many tracking approaches using fair evaluation criteria. Online Object Tracking Benchmark [1] contais 50 of the most commonly used sequences. Also, the authors classified tracking challenges (attributes) into subsets to report specific challenging conditions. The Amsterdam Library of Ordinary Videos for tracking, ALOV300+ [63], consists in 315 real-life video sequences from Youtube with 64 different targets. The collection is categorized for thirteen aspects of difficulty and evaluates a large variety of situations including low contrast, occlusion, ilumination variation, etc.

2.5 Object tracking applying ensemble

Several methods have proposed the use of ensemble classifiers within the tracking-by-detection framework. In this perspective, instead of ensembling the outputs of other trackers, these methods focus on building ensemble classifiers, such as Adaboost [44] or Bayesian probabilistic fusion [73], from weak low-level image features. The classifier is then used to detect the target object in new frames. Instead of working directly with low-level image features, our ensemble tracker builds on top of tracker outputs which are used as a more powerful mid-level representation of the target.

Another alternative is to manually select a small set of trackers and use prior knowledge on the behaviour of each tracker to build an ensemble. In [69], the authors combine a

template-based tracker, an optical flow tracker, and an online-random forest tracking-by-detection method into a cascade. In that case, the authors manually predefine a set of rules that decide how to ensemble the tracker outputs. In contrast, our method can handle a larger pool of trackers in the ensemble, since their outputs are combined in a data-drive fashion that does not require manual rules or prior knowledge on the behaviour of each tracker.

Sampling based approaches have also been explored for ensemble tracking. Examples of this are the Visual Tracking Detector (VTD) [67] and the Visual Tracker Sampler (VTS) [68] trackers. In this perspective, there is a sampling process that generates multiple samples of target and tracker states. These trackers run in parallel and their outputs are fused by probabilistic weighting. Therefore, the ensemble uses a set of trackers of similiar architecture with varying parameters. Our ensemble has the advantage of fusing outputs of multiple trackers with no assumptions on their architectural similarity and can leverage strenghts that different tracker architectures can provide.

Finally, one can consider the case of offline fusion of trackers, such as in [74], where all trackers are applied to the entire sequence and the ensemble is performed after the entire sequence is processed. However, we are interested in the case of online tracking, where the full sequence is not available beforehand. Furthermore, our online approach is capable of steering and reinitializing failed trackers, increasing the chances of better long-term tracking performance.

Chapter 3

Proposed approach

As it can be seen in Chapter 2, several tracking approaches have been proposed before. These methods attain relative success under particular circumstances. Based on the above, we present in this chapter our ensemble tracking approach, which combines many tracking results into an online ensemble. We consider that combining each tracker virtues, while evading their weaknesses, could outperform each single algorithm used individually.

First, in this chapter we give an introduction to our online ensemble tracking approach, and then describe the details of each stage of our processing pipeline. Given tracking inputs, the algorithm estimates the trajectory of an object as it moves around a scene. Then, the tracker outputs labels of the tracked object in every frame of the video (Figure 3.1). These labels provide position and scale information for evaluation and analysis.

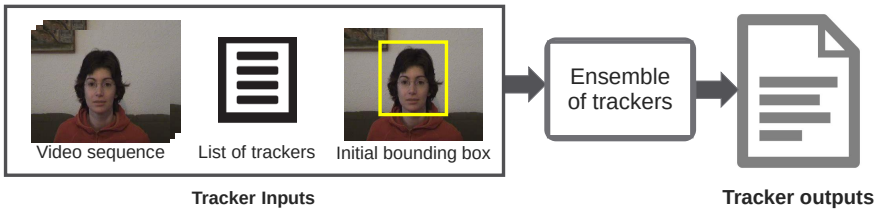


Figure 3.1: Tracking diagram. The proposed system takes an initial bounding box as region of interest, image sequences and the list of trackers as inputs. After tracking ensemble, the system outputs tracking results for each frame.

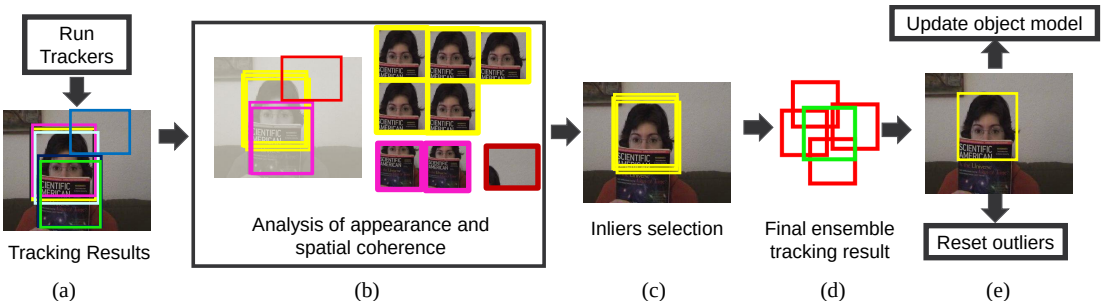


Figure 3.2: Basic diagram of our approach. On each frame, we analyze coherence between tracking results. We apply spatial and appearance models with the goal of finding inliers. Then, we select best tracker that follows the target. Finally, we reset outliers and update object model on necessary cases.

3.1 Overview

We illustrate the main components of our method in Figure 3.2. At each frame, our method proceeds as follows. First, we independently run all trackers $T = \{t_1, t_2, ..., t_n\}$ in a pool of size n, which produce a set of predicted target states $X = \{x_1, x_2, ..., x_n\}$ (Fig. 3.2a). These predictions are the raw input of our ensemble algorithm, and may be in the form of bounding boxes. A bounding box is defined as the set of pixels covered by a rectangular area. Our ensemble methodology then looks for spatial coherence among the predicted target states, and verifies the appearance of the predicted image regions with respect to an object model (Fig. 3.2b). The idea is that for a given frame, we would like to discover the subset of trackers that are correctly estimating the target state and to ignore all erroneous predictions. The result is a selection of inlier tracking predictions that contain the object with high confidence (Fig. 3.2c). A final ensemble prediciton of the target state is then derived from these inlier estimations (Fig. 3.2d). Finally, we use the prediction of the ensemble to update a model of the target appearance, and periodically reinitialize outlier trackers (Fig. 3.2e). This entire process is then repeated for each new frame in the video sequence.

In spite of the simplicity of our ensemble procedure, our experiments evidence that the ensemble is able to produce more accurate tracking results than any of the trackers in the pool. Furthermore, the ensemble can achieve state-of-the-art performance on benchmarking videos.

In the following, we provide the details of each processing stage and their role in the overall ensemble procedure.

3.2 Analysis of appearance and spatial coherence

After running all trackers in the pool, our method uses their state estimates X as input for the appearance and spatial coherence stage (Fig. 3.2b). The goal is to determine the set of trackers that correctly estimate the true target state. We denote these trackers as inliners in Fig. 3.2c.

In order to achieve this, we note a simple but key observation: in most cases, there is only a small subset of trackers that fail to correctly track the object in any given frame. In such scenario, the majority of the trackers focus correctly in the object, and we could use clustering techniques to automatically determine this set of inlier trackers. A more difficult scenario is when most of the trackers fail, but only a few focus in the correct image region. But even in this case, tracker failures tend to be distributed in varied image locations, while the few correct trackers focus on a spatially coherent region. Once again, spatial clustering of the locations X can help discover the underlying true object location. In an extreme scenario, when only one tracker correctly follows the object, we can no longer rely on spatial clustering alone. A complementary cue can be introduced to overcome this issue: a target appearance model. We therefore explore the use of these two cues to select a subset of inlier trackers that follow the object region with high confidence.

3.2.1 Spatial clustering

Cluster analysis is the formal study of algorithms and methods for grouping, or classifying entities. These entities are described as a set of measurements or by relationships between the entity and other entities. A *cluster* is comprised of a number of similar entities collected or grouped together. Other authors define a cluster as a set of instances which are alike, and instances from different clusters are not alike. Also, a cluster could be defined as an aggregation of points in the test space such that the *distance* between any two points in the cluster is less than the distance between any point in the cluster and any point not in it [75].

Cluster analysis is the process of classifying entities into subsets that have meaning in the context of a particular problem. The entities are thereby organized into an efficient representation that characterizes the data. Clustering methods require that an index of proximity, or alikeness, or affinity, or association be established between pairs or patterns. A *proximity matrix* $|d(i,j)|$ accumulates the pairwise indices of proximity in a matrix in which each row and column represents an entity. Diagonal entries of a proximity matrix are ignored since all paterns are assumed to have the same degree of

proximity with themselves. Also it is assumed that all proximity matrices are symmetric, so all pairs of entities have the same proximity index, independent of the order in which they are written. A proximity index is either a *similarity* or a *dissimilarity*. The more the ith and jth entities are similar one another, the larger a similarity index and the dissimilarity index are.

At this stage, we perform spatial clustering of all tracker predictions X. Bounding boxes with large overlap, similar location and scale should be grouped into the same cluster. Since we do not know the number of natural groups beforehand, we use an agglomerative clustering technique to achieve this. In practice, we apply complete-link hierarchical agglomerative clustering (CL) [75]. We define a dissimilarity measure between pairs of bounding boxes equal to:

$$d(x_i, x_j) = 1 - \frac{|x_i \bigcap x_j|}{|x_i \bigcup x_j|} \tag{3.1}$$

Using all dissimilarity values, we construct a symmetric $n \times n$ proximity matrix D. In CL, a pair of bounding boxes (x_i, x_j) will be grouped in the same cluster if their dissimilarity is below some threshold v. For all our experiments we set this value to 0.8, which generally corresponds to a small number of clusters. The result of CL is a grouping of the input bounding boxes X into m clusters $C = \{c_1, c_2, ..., c_m\}$, which are illustrated with colors in Fig. 3.2b. These clusters satisfy the following:

- $c_i \cap c_j = \emptyset$ for i and j from 1 to m, $i \neq j$
- $c_1 \cup c_2 \cup ... \cup c_m = T$

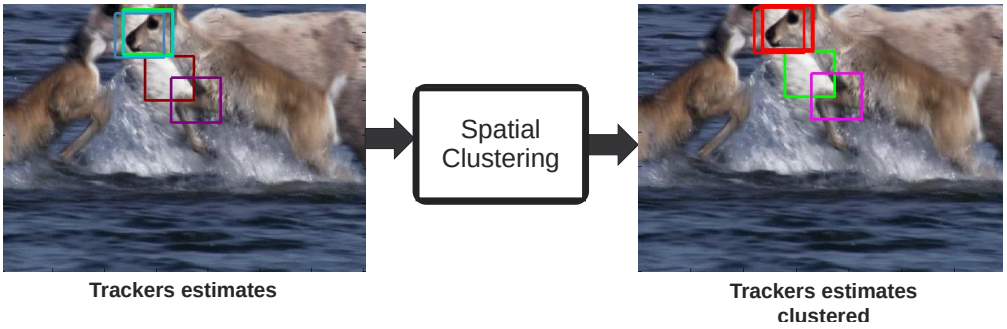


Figure 3.3: Spatial clustering stage. Using predicted positions from each tracker, we cluster similar bounding boxes using equation 3.1. In the image on the right, each color represents a cluster.

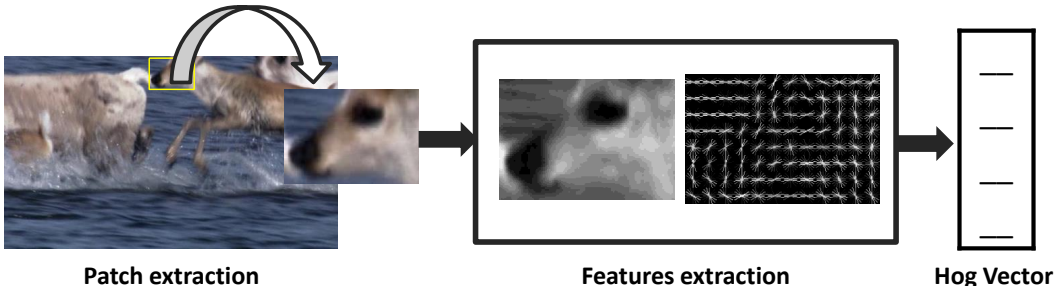


Figure 3.4: Patches and features extraction process. For a given patch, Histograms of Oriented Gradients are extracted. The final result, corresponds to a hog feature descriptor.

3.2.2 Object modeling

Visual object tracking has been formulated as a tracking-by-detection problem recently. Object modeling is dynamically performed to support object detection over all frames. Mostly all the approaches can be classified into two categories: *Generative appearance models*, that mainly focus on how to fit the data into their correspondent object class; and *discriminative appearance models*, which assume object tracking as a binary classification issue. Main goal is to maximize the separability between object and non-object regions discriminately.

Generally, discriminative methods train a classifier using data acquired from previous frames, and subsequently use the trained classifier to evaluate possible object regions at the current frame. After localization, a set of *positive* and *negative* samples are heuristically selected to update the classifier. Some approaches apply online boosting [33, 45, 48], that make a discriminative evaluation of features taken from a candidate feature pool, and then select the top ranked features to conduct the tracking process. Other methods apply Support Vector Machines (SVM) method, which learns a margin-based discriminative appearance model, in order to maximize inter-class separability. These classifiers are trained using visual representations of the object.

At this stage, we would like to verify the appearance of all tracker predictions X in comparison to an object appearance model. In order to do so, we train an appearance classifier that aims at separating positive target bounding boxes from background bounding boxes. We extract positive samples (Figure 3.4) from the target location given at the first frame and also from uniformly sampling background bounding boxes around the target bounding box. In practice, we adopt a feature representation based on Histogram of Oriented Gradients: (HOG) and a SVM classifier with probability outputs. Using the classifier, we compute appearance scores $S = \{s_1, s_2, ..., s_n\}$ for each tracking result x_i in X (Figure 3.5). We considered this model, because tracking methods that applied SVM classifier [73, 76], obtained better results than other classifiers *e.g.*

Boosting [33, 45]. Also, HOG features are more robust to deformable objects than other image features used in tracking-by-detection methods *e.g.* Haar [44].

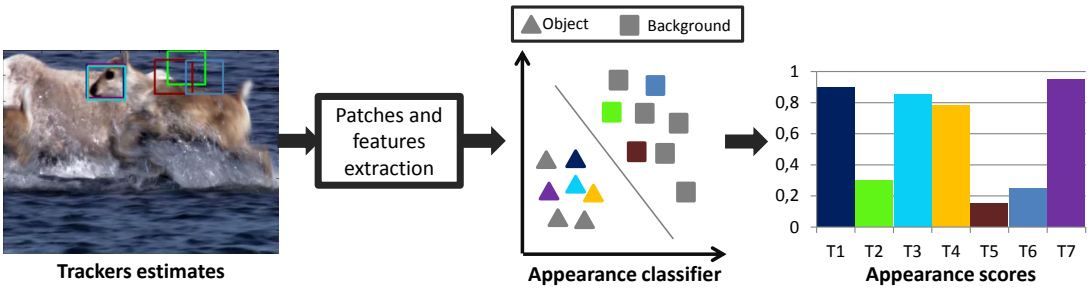


Figure 3.5: Appearance estimation for trackers estimates. On each frame, We compute appearance score for each tracker result using a previously trained classifier. In this figure, patches classified as object, have high appearance scores.

3.3 Selection of inliers

The clustering stage and appearance scoring provide cues about which trackers can be considered as correctly tracking the target. We evaluate two simple criteria that optimally select a group of trackers as inliers (Fig. 3.2c). In our experiments, we consider cluster size and appearance scores as potential cues to select inliers. The user defines initially which criteria will use in the ensemble.

Cluster size: This critera follows the idea that the largest spatially coherent cluster is associated to the true target location. Therefore, we flag all trackers in the largest cluster as inliers, and all other trackers as outliers.

$$c^* = \arg\max_{c_i \in C} |c_i| \tag{3.2}$$

Appearance scores: In this criteria, we trust our object appereance model to validate if a cluster contains bounding boxes that focus on the true target location. To do so, we max-pool the appereance scores s of the bounding boxes x within each cluster:

$$c^* = \arg\max_{c_i \in C} \max_{x_j \in c_i} s_j \tag{3.3}$$

3.4 Final ensemble tracking result

In order to provide an estimation of the state of the target using our ensemble, we fuse the outputs of all inlier trackers from the previous stage. There are multiple choices

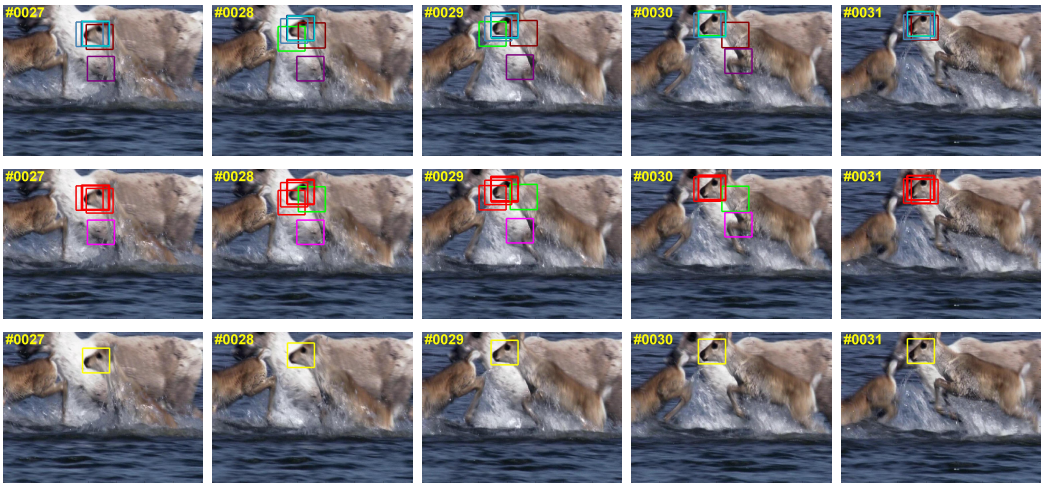


Figure 3.6: System behavior in five frames. Given image input, trackers will give results on where the object might be (first row). Then, all results are clustered using hierarchical agglomerative clustering (second row). We focus on selecting the group with highest number of members, or cluster with best appearance score (third row). Finally, we reinitialize outliers.

on how to implement this fusion. For example, it can be a linear combination of the inlier bounding boxes. In practice, we use a simpler approach that selects the medoid bounding box as the final ensemble output. That is, we output a bounding box x^* whose sum of distances with the rest of inlier bounding boxes is minimum:

$$x^* = \arg\min_{x_i \in c^*} \sum_{x_j \in c^*} d(x_i, x_j) \tag{3.4}$$

3.5 Model update and outlier reset

Once our ensemble estimates the target state, we can proceed to update the object model and steer failed trackers in the pool.

The first goal of this final stage is to keep our target appearance model updated. In order to avoid model drifting, we only update the appearance model periodically when our tracking ensemble has high confidence in the selection of inliers. Such confidence can be measured using the appearance scores S or the percentage of trackers in the inliner cluster. When confidence is high, a image patch is extracted at the predicted location and provided as a new positive sample to retrain or update our object appearance model.

The second goal is to steer failed trackers in the pool back to the target. When a tracker is tagged as an outlier, our algorithm automatically resets its state to the latest target state prediction from the ensemble (Figure 3.7). This reinitialization procedure is only performed periodically every 15 frames so that we can also take advantage of the capability of some trackers to recover from short-term tracking failures.

The proposed object tracking using ensemble of multiple tracking algorithms, can be visualized in Figure 3.6. In it, after running each tracker individually, the results are clustered using CL. Then, the system finds inliers applying the potential cues explained in Section 3.3. Finally, the medoid bounding box is selected from inliers. Each 15 frames, outliers are reinitialized.

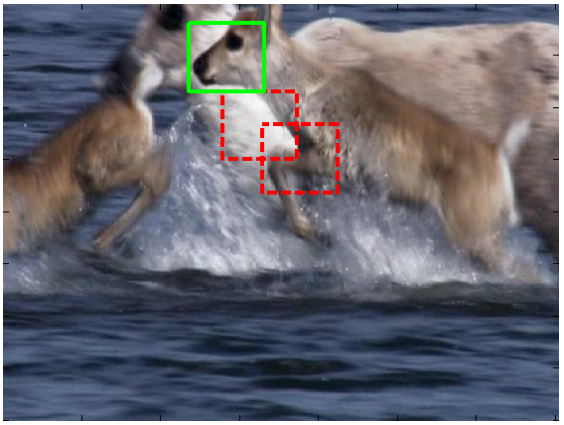


Figure 3.7: Outliers reinitialization. In this Figure, dashed red bounding boxes are tagged as outliers. State from each outlier are reset to the latest target state prediction from the ensemble (green bounding box.)

Chapter 4

Experiments

In this section we report evaluations for our proposed approach. We compare our tracking ensemble method and other state-of-the-art trackers using the complete 50 sequences tracking benchmark [1]. We also present qualitative and quantitative analysis of our ensemble tracker, such as individual performance of trackers and effect of the tracker pool.

4.1 Implementation details

We implemented our online ensemble approach in MATLAB. CL clustering method is implemented in C++ and ported as a *mex* function. Object model corresponds to a standard RBF SVM, implemented in LIBSVM library [77]. To extract HOG features, we made use of Piotr Toolbox for Matlab [78] using default parameters. All experiments were performed on a PC with an Intel Xeon CPU and 16 GB of RAM.

4.2 Evaluation Metrics

The trackers were put into test using two evaluation metrics commonly employed in the literature. The first is *distance precision* (DP), which is the number of frames where the estimated center location is within a certain distance threshold d from the ground truth center location. The second metric is *overlap precision* (OP), defined as the number of frames where the overlap between the estimated and ground truth bounding box exceeds a threshold b. On each sequence, these two measures are calculated using equations 4.1a and 4.1b. Estimated center location and ground truth are denoted as $\mathbf{p}_f$ and $\hat{\mathbf{p}}_f$ respectively, where f is the frame number. Futhermore, B_f and $\hat{B}_f$ denote the

estimated and ground truth bounding boxes of the object. N is the number of frames in the sequence.

$$\mathrm{DP}(d) = \frac{1}{N}|\{f : ||\hat{\mathbf{p}}_f - \mathbf{p}_f|| \leq d\}|,\ d \geq 0 \tag{4.1a}$$

$$\mathrm{OP}(b) = \frac{1}{N}\left|\left\{f : \frac{|\hat{B}_f \bigcap B_f|}{|\hat{B}_f \bigcup B_f|} \geq b\right\}\right|,\ 0 \leq b \leq 1 \tag{4.1b}$$

In the recent literature, there has not been much agreement on which performance measure to use for comparing visual trackers. DP focuses only on estimated center location, which is beneficial for trackers that do not estimate scale of the object. Instead, OP also takes estimated scale into account and penalizes if the scale of the target is estimated incorrectly.

Recently, the authors of the Online Object Tracking Benchmark (OOTB) [1] suggest the usage of precision and success plots. These curves show distance and overlap precision metrics over a range of thresholds. For instance, in precision plot, a higher precision at low thresholds means the tracker is more accurate, while a lost target will not achieve perfect precision on a large threshold range. In both types of plots, a *ranking score* is computed to line up trackers overall performance. In precision plot, the DP value of 20 pixels is used as ranking score. In contrast to precision, in success plots, trackers are ranked using area under the curve (AUC), which relates to the average overlap across all frames.

To validate the performance of our proposed approach, we follow the one-pass evaluation methodology (OPE) proposed in [1]. OPE evaluates trackers performance running them throughout a test sequence with initialization from the ground truth position in the first frame. We summarize results using the precision and success plots in the complete dataset.

4.3 Dataset

In order to evaluate the performance of our proposed ensemble tracking, we adopt the OOTB from [1]. This is an extensive benchmarking dataset that includes 50 sequences annotated with 11 attributes. The dataset includes challenging scenarios such as motion blur, illumination changes, scale variation, occlusions, in-plane and out-of-plane rotations, object deformation, background clutter and low resolution. The selection criteria of this dataset are listed below.

Table 4.1: Selected tracking algorithms for ensemble method. **Code Column**: M: Matlab, MC: Mixture of Matlab and C/C++, other: DLL files.

Method	**Type**	**Code**
Template matching - TM [79]	Template-based	other
Mean Shift - MS [79]	Density-based	other
Variance Ratio - VR [79]	Density-based	other
Peak Difference - PD [79]	Point-based	other
Ratio Shift - RS [79]	Point-based	other
Adaptive Structural Local Sparse Appearance Model - ASLA [80]	Sparse representation	MC
Compressive Tracker - CT [81]	Sparse representation	MC
Minimum Experts Entropy Minimization - MEEM [76]	tracking by detection	MC
Self-paced learning for long-term tracking - SPLTT [82]	tracking by detection	MC
Kernelized Correlation Filters - KCF [83]	Template-based	M
Dual Correlation Filters - DCF [83]	Template-based	M
Spatio-temporal Context Tracker - SCT [84]	Template-based	M
Circulant Structure Kernel - CSK, sKCF [83]	Template-based	M
Robust CSK tracker - RCSK [85]	Template-based	M
Minimum Output Sum of Squared Error - MOSSE [86]	Template-based	M

- **Standarized sequences:** For better comparison, OOTB videos are widely used as benchmark sequences in recent literature, allowing us to make a fair comparison to state-of-the-art algorithms. Also, all of the sequences were selected for generic single-object tracking purposes.
- **Manual annotations available:** OOTB provides ground truth annotations for each sequence. This avoids a costly labeling process.
- **Image frames:** Each sequence corresponds to a set of image frames. This format allows to evaluate global processing time of the ensemble approach.
- **Attributes classification:** Each sequence provides an annotation with 11 attributes, which explain the challenging scenarios encountered. This allows us to make attribute-based comparisons, which can show strengths and weaknesses of different trackers.

4.4 Tracker pool

Table 4.1 shows the list of the selected tracking algorithms that will be included in the pool for our experiments. The selection criteria for the trackers are listed below.

- **Source code:** We integrate into our pool tracking algorithms whose original source code is publicly available. Executable files were not selected.

- **Store current state:** On each frame, a tracker must be able to store current state and give current result. This allow the ensemble to perform spatial clustering and compute appearance score for each tracker.
- **Trackers bounding box result:** Each tracker estimate must give current target location and scale overall frames. Also, the current result must be able to be transformed into a generic format, in order to perform ensemble.
- **Reinitialization available:** A tracker is able to reset its state using final ensemble tracking result.

4.5 Benchmarking results

We report the performance of our ensemble tracking algorithm on the 50 benchmarking sequences in Figure 4.1. The plots compare our approach with each individual tracker in Table 4.1 and state-of-the-art trackers: Struck tracker [87], Sparse Collaborative Model (SCM) [88], TLD tracker [66]. Our online ensemble tracking algorithm is denoted *OET*.

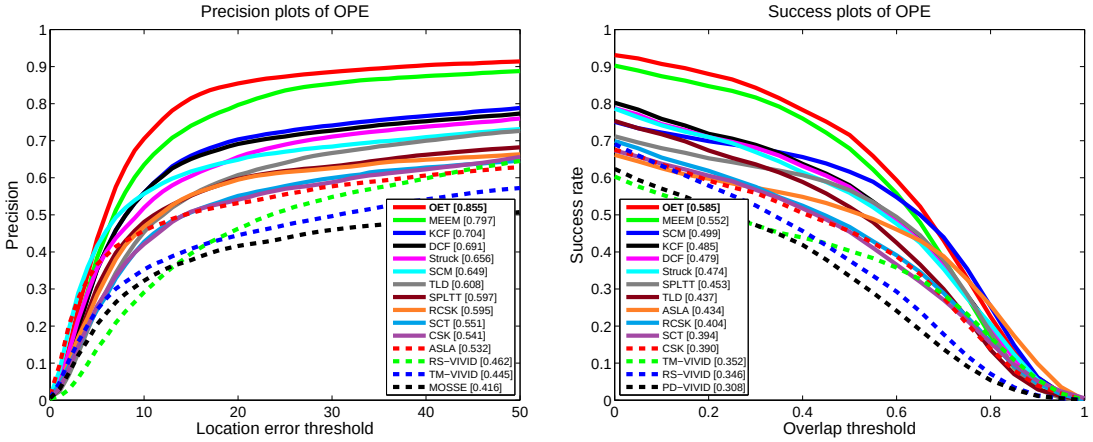


Figure 4.1: Precision and success plots for all 50 sequences. Precision and success ratios are measured by center location error and overlap ratio, respectively. Trackers are ranked using scores of 20 pixels for precision and AUC for success.

We note that our tracking ensemble improves performance over current state-of-the-art tracking algorithms in the benchmark. From Figure 4.1, our method outperforms each individual tracking score, in both precision and success. When inspecting individial sequences, we note that our method handles camera rotation sequences better, such as *singer1, singer2, fish, car4*, unlike the competing MEEM tracker. Also, our ensemble method can overcome complete object occlusion. In sequences where this situation happens like *jogging, david3*, KCF trackers fails tracking the object.

4.6 Analysis of the tracking ensemble

In this subsection, we focus on the relevance of the number of trackers in the pool. We vary the size of the tracker pool in order to evaluate its effect on the performance of the ensemble. When selecting small tracker pools, we first chose the best 5 trackers from Fig. 4.1b. We then augment the pool with the next 5 trackers to form a pool of size 10. Finally we add the worst 5 trackers for a pool of size 15. We also study the effect of the inlier selection criteria of Section 3.3 (appearance score and cluster size).

Table 4.2: Average AUC and precision for inliers selection methods.

Method	# trackers	Success(AUC)	Precision(20px)
Appearance score	5	0.585	0.855
	10	0.569	0.811
	15	0.560	0.804
Cluster size	5	0.528	0.765
	10	0.493	0.715
	15	0.480	0.678

The results are summarized in Table 4.2 for all 50 sequences in the benchmark. We report AUC ranking score for success, and 20-pixel threshold for precision. Our algorithm generally outperforms other methods using appearance score selection criteria. In case of cluster size selection, this method usually fails handling occlusions. In some sequences, many trackers lose object target and keep steady when occlusion happens, creating a big cluster which has the biggest number of members (Figure 4.2). However, this method is smoother and handles better fast motion and background clutter sequences.

We also note that adding trackers with lower performance hurts the ensemble. However, the drop in performance when adding weaker trackers, is less than 5% (∼1500 frames out of 29519) in success and 10 % in precision (∼3000 frames). For instance, when

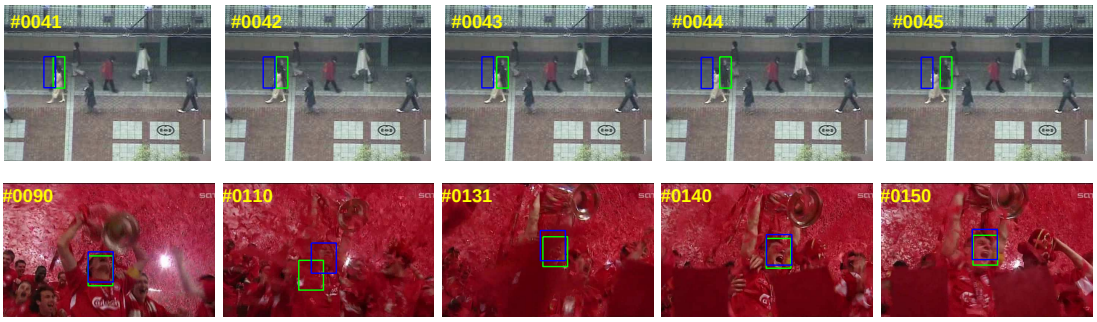


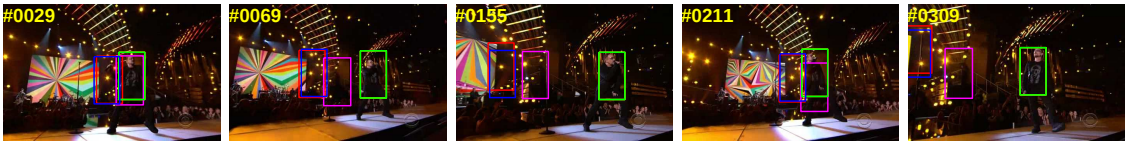
Figure 4.2: Qualitative results for object tracking applying both inliers selection methods in two sequences. **Green** bounding box corresponds to appearance score selection. **Blue** box to cluster size. In *subway* when occlusion happens, many trackers are lost, creating a big cluster. In cases of background clutter *soccer*, appearance selection loses target in some frames.

(a) bolt sequence.

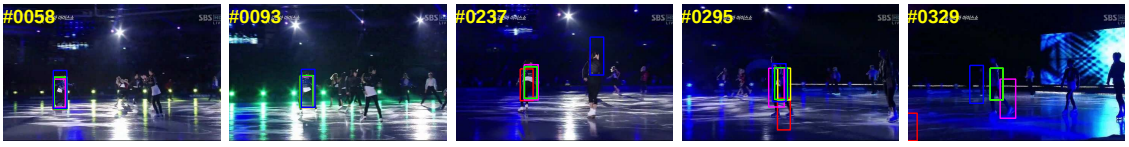
(b) jogging-2 sequence.

(c) singer-1 sequence.

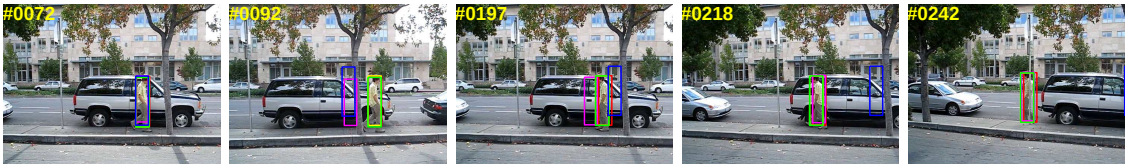
(d) football sequence.



(e) singer-2 sequence.



(f) skating1 sequence.



(g) david3 sequence.

Figure 4.3: Screenshots of tracking results.

performing inlier selection using the appearance score criteria, a spurious tracker may focus on a region with very similar apperance to the object, *e.g.* background clutter, which can make tracking fail. This is very common in the *soccer* and *shaking* sequences. It is important to note that even a small percentage difference could be significant in terms of sequences.

In terms of running time, the average time cost for our ensemble method is 0.062 s/frame for 5 trackers, 0.122 s/frame for 10 trackers, and 0.198 s/frame for 15 trackers. These timings do not include the processing time of each tracker.

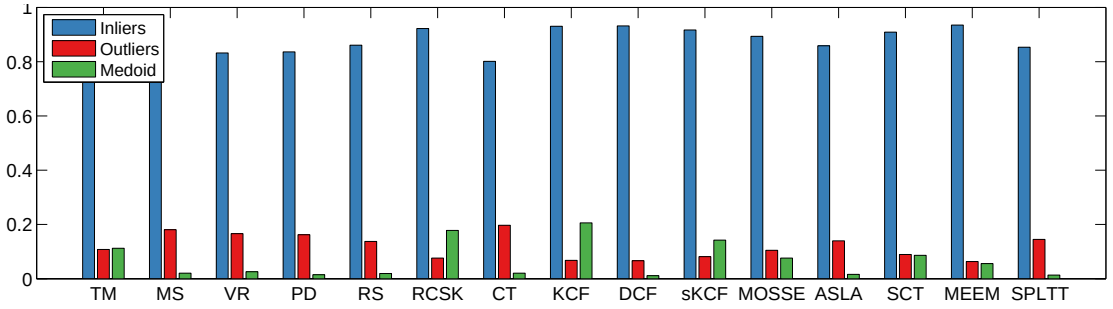


Figure 4.4: Statistics for each tracker in the ensemble over all frames.

On the other hand, we present statistics about how often trackers in the pool are selected as outlier, inliner, and medoid. In figure 4.4, for each tracker, there are 3 bars: percentage of frames that a tracker was in best cluster(inlier - blue bar); percentage of frames where a tracker was considered outlier (red bar); finally, percentage of frames that a tracker was selected as medoid bounding box (green bar). Based on this result, trackers whose individual performance is very high, have low percentage of being considered outliers (KCF, MEEM, RCSK). MEEM individual performance is very high. However, it does not have the highest frame percentage of being selected as central bounding box, in comparison with KCF or RCSK. Also, trackers that were considered spurious in previous experiments, have high rate in outliers bar (MS, VR, PD, RS).

4.7 Experiments with sequence attributes

The videos in the benchmark dataset are organized and selected with attributes, which describe challenges present in the sequence - *e.g.* occlusion, object deformations. These properties are useful for diagnosing tracking behavior, without the need of analyzing each video separately. Figure 4.5 shows AUC ranking scores of recent trackers on different sequences, grouped by attributes. For instance, background clutter (BC) contains all sequences whose target pixels might be confused with background.

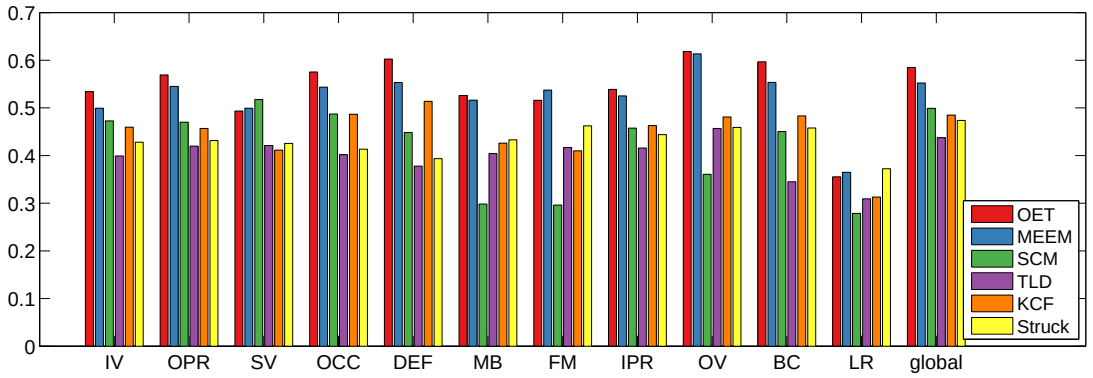


Figure 4.5: Average AUC ranking scores of top trackers on different subsets of test sequences in OPE. Each subset of sequences corresponds to an attribute: IV - illumination variation, OPR - out of plane rotation, SV- scale variation, OCC - occlusion, DEF - deformation, MB - motion blur, FM - fast motion, IPR - in plane rotation, OV - out of view, BC - background clutter, and LR - low resolution. Average AUC for all 50 videos is presented as global.

From figure 4.5, our approach using appearance selection outperforms other trackers in 8 of 11 attributes. Specifically, in attributes such as IV (illumination variation), OPR (out of plane rotation), OCC (occlusion), DEF (deformation), MB (motion blur), IPR (in plane rotation), OV (out of view), and BC (background clutter). It is important to note that SCM tracker is better than most recent trackers in terms of scale variation. Results show that its affine motion models handle scale variation better than other trackers, which are designed to account translational motion [88]. In our system, scale is not determined. We are dependent of each separated tracker scale, and some trackers do not consider scale correction. Some of them apply initialization scale over all sequence.

Chapter 5

Conclusion and Future Work

In this thesis, we demonstrated that an online ensemble of trackers can result in improved performance with respect to each individual tracker. Our framework considers the spatial coherence and appearance of the predicted target locations to ensemble a final estimate of the target state. We leverage high confidence estimation to reinitialize trackers that fail and steer them towards the true region. Our experiments show that our simple but effective technique can achieve state-of-the-art tracking performance in benchmarking sequences.

Using a quite simple framework, further improvements are expected to be possible. For instance, the ensemble method does not estimate scale. This limitation needs to be addressed in future research, since it might give significant performance gain. One possible solution could be the application of a sliding window in the scale dimension, preserving low computational cost.

We plan to enlarge evaluation of the ensemble of trackers using large scale evaluation benchmark. Moreover, we expect to formulate a novel appearance model update, that will perform this task more efficiently, avoiding spurious samples aggregation into the model. Finally, we propose to apply better object modeling using more robust features.

Glossary

Clustering: Task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar to each other than to those in other groups (clusters).

Feature descriptor: Or feature vectors are algorithms that take an image and outputs numerical vectors. These vectors include location of the feature as well as other information.

Feature: Part of an image that contains interesting details or a property of the image which we are interested in.

Ground truth: Set of measurements that is known to be much more accurate than measurements from the system that is tested. These data usually comes from manual annotations.

Histogram of Oriented Gradients: Or HOG, is a feature descriptor used to detect objects. The HOG descriptor technique counts occurrences of gradient orientation in localized portions of an image - detection window, or region of interest.

Inliers: Group or cluster whose estimates represent correctly the object..

Matching: Is the process of establishing a correspondence between two or more objects by comparing features of these objects to one another.

Outliers: Cluster or clusters that are not correctly following the object..

Overlap: Intersection over union of two bounding boxes.

Region of interest(ROI): A particular portion of the image that seems important.

Bibliography

[1] Yi Wu, Jongwoo Lim, and Ming-Hsuan Yang. Online Object Tracking: A Benchmark. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2411–2418, 2013.

[2] Cor J. Veenman, Marcel J T Reinders, and Eric Backer. Resolving motion correspondence for densely moving points. *IEEE transactions on Pattern Analysis and Machine Intelligence*, 23:54–72, 2001.

[3] Khurram Shafique and Mubarak Shah. A noniterative greedy algorithm for multiframe point correspondence. *IEEE transactions on Pattern Analysis and Machine Intelligence*, 27:51–65, 2005.

[4] Dorin Comaniciu, Visvanathan Ramesh, and Peter Meer. Kernel-based object tracking. *IEEE transactions on Pattern Analysis and Machine Intelligence*, 25:564–577, 2003.

[5] Ismail Haritaoglu, David Harwood, and Larry S. Davis. W4: Real-time surveillance of people and their activities. *IEEE transactions on Pattern Analysis and Machine Intelligence*, 22: 809–830, 2000.

[6] G. Paschos. Perceptually uniform color spaces for color texture analysis: An empirical evaluation. *IEEE Transactions on Image Processing*, 10:932–937, 2001.

[7] K. Y. Song, J. Kittler, and M. Petrou. Defect detection in random colour textures. *Image and Vision Computing*, 14:667–683, 1996.

[8] J Canny. A computational approach to edge detection. *IEEE transactions on Pattern Analysis and Machine Intelligence*, 8:679–698, 1986.

[9] Kevin Bowyer, Christine Kranenburg, and Sean Dougherty. Edge Detector Evaluation Using Empirical ROC Curves. *Computer Vision and Image Understanding*, 84:77–103, 2001.

[10] Michael J Black and Allan D Jepson. EigenTracking: Robust Matching and Tracking of Articulated Objects Using a View-Based Representation. *International Journal of Computer Vision*, 26:63–84, 1996.

[11] Bruce D Lucas and Takeo Kanade. An Iterative Image Registration Technique with an Application to Stereo Vision. *International joint conference on Artificial intelligence*, 130: 674–679, 1981.

[12] C. Harris and M. Stephens. A Combined Corner and Edge Detector. *Proceedings of the Alvey Vision Conference 1988*, pages 147–151, 1988.

[13] Robert M. Haralick, K. Shanmugam, and Its'Hak Dinstein. Textural Features for Image Classification. *IEEE Transactions on Systems, Man, and Cybernetics*, 3, 1973.

[14] Kevin M. Nickels and Seth Hutchinson. Textured image segmentation, 1997.

[15] Stephane G. Mallat. Theory for multiresolution signal decomposition: the wavelet representation. *IEEE transactions on Pattern Analysis and Machine Intelligence*, 11:674–693, 1989.

[16] Am McIvor. Background subtraction techniques. *Proceedings of Image and Vision Computing*, 2000.

[17] Vu Pham, Phong Vo, Vu Thanh Hung, and Le Hoai Bac. GPU Implementation of Extended Gaussian Mixture Model for Background Subtraction. *2010 IEEE RIVF International Conference on Computing & Communication Technologies, Research, Innovation, and Vision for the Future (RIVF)*, pages 1–4, 2010.

[18] A J Lipton, H Fujiyoshi, and R S Patil. Moving target classification and tracking from real-time video. *Proceedings Fourth IEEE Workshop on Applications of Computer Vision*, 98:8–14, 1998.

[19] Liang Wang, Weiming Hu, and Tieniu Tan. Recent developments in human motion analysis. *Pattern Recognition*, 36:585–601, 2003.

[20] Robert T Collins, Alan J Lipton, Takeo Kanade, Hironobu Fujiyoshi, David Duggins, Yanghai Tsin, David Tolliver, Nobuyoshi Enomoto, Osamu Hasegawa, Peter Burt, and Lambert Wixson. A System for Video Surveillance and Monitoring, 2000.

[21] B. Stenger, V. Ramesh, N. Paragios, F. Coetzee, and J.M. Buhmann. Topology free hidden Markov models: application to background modeling. In *IEEE International Conference on Computer Vision*, volume 1, 2001.

[22] J Rittscher, J Kato, S Joga, and A Blake. A probabilistic background model for tracking. In *European Conference on Computer Vision*, pages 336–350, 2000.

[23] Jianbo Shi and C. Tomasi. Good features to track. In *IEEE Conference on Computer Vision and Pattern Recognition*, 1994.

[24] David G Lowe. Distinctive Image Features from Scale-Invariant Keypoints. pages 1–28, 2004.

[25] Alper Yilmaz, Omar Javed, and Mubarak Shah. Object tracking: A survey. *ACM Computing Surveys*, 38:13, 2006.

[26] Zhen Qin and Christian R. Shelton. Improving multi-target tracking via social grouping. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1972–1978, 2012.

[27] Xiaofeng Ren. Finding people in archive films through tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[28] Janne Heikkilä and Olli Silvén. A real-time system for monitoring of cyclists and pedestrians. *Image and Vision Computing*, 22:563–570, 2004.

[29] Kenji Okuma, Ali Taleghani, Nando De Freitas, James J Little, and David G Lowe. A Boosted Particle Filter : Multitarget Detection and Tracking. In *European Conference on Computer Vision*, pages 28–39, 2004.

[30] Yizong Cheng. Mean shift, mode seeking, and clustering. *IEEE transactions on Pattern Analysis and Machine Intelligence*, 17:790–799, 1995.

[31] David Exner, Erich Bruns, Daniel Kurz, Anselm Grundhöfer, and Oliver Bimber. Fast and robust CAMShift tracking. In *IEEE Conference on Computer Vision and Pattern Recognition- Workshops*, pages 9–16, 2010.

[32] Simon Korman, Daniel Reichman, Gilad Tsur, and Shai Avidan. FasT-match: Fast affine template matching. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2331–2338, 2013.

[33] Helmut Grabner, Michael Grabner, and Horst Bischof. Real-Time Tracking via On-line Boosting. *British Machine Vision Conference*, 1:1–10, 2006.

[34] Greg Mori and Jitendra Malik. Recovering 3D human body configurations using shape contexts. *IEEE transactions on Pattern Analysis and Machine Intelligence*, 28:1052–1062, 2006.

[35] Xiaoming Liu and Ting Yu. Gradient feature selection for online boosting. In *IEEE International Conference on Computer Vision*, 2007.

[36] M. Isard and J. MacCormick. BraMBLe: a Bayesian multiple-blob tracker. In *IEEE International Conference on Computer Vision*, volume 2, 2001.

[37] Vincent Lepetit and Pascal Fua. Keypoint recognition using randomized trees. *IEEE transactions on Pattern Analysis and Machine Intelligence*, 28:1465–1479, 2006.

[38] D Comaniciu, V Ramesh, and P Meer. Real-time tracking of non-rigid objects using mean shift. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 142–149, 2000.

[39] Amit Adam, Ehud Rivlin, and Ilan Shimshoni. Robust fragments-based tracking using the integral histogram. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 798–805, 2006.

[40] David a. Ross, Jongwoo Lim, Ruei-Sung Lin, and Ming-Hsuan Yang. Incremental Learning for Robust Visual Tracking. *International Journal of Computer Vision*, 77:125–141, 2007.

[41] Iain Matthews, Takahiro Ishikawa, and Simon Baker. The template update problem. *IEEE transactions on Pattern Analysis and Machine Intelligence*, 26:810–815, 2004.

[42] A D Jepson, D J Fleet, and T F El-Maraghi. Robust online appearance models for visual tracking. *IEEE transactions on Pattern Analysis and Machine Intelligence*, 25:1296–1311, 2003.

[43] Y Freund and R E Schapire. A Decision-theoretic Generalization of On-line Learning and an Application to Boosting. *Journal of Computing Systems and Science*, 55:119–139, 1997.

[44] Shai Avidan. Ensemble tracking. *IEEE transactions on Pattern Analysis and Machine Intelligence*, 29:261–271, 2007.

[45] Helmut Grabner, Christian Leistner, and Horst Bischof. Semi-supervised on-line boosting for robust tracking. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 5302 LNCS, pages 234–247, 2008.

[46] NC Oza and S Russell. Online ensemble learning. *Association for the Advancement of Artificial Inteligence*, 6837:1109–1109, 2000.

[47] Amir Saffari, Martin Godec, Thomas Pock, Christian Leistner, and Horst Bischof. Online multi-class LPBoost. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3570–3577, 2010.

[48] Boris Babenko, Ming-Hsuan Yang, and Serge Belongie. Visual Tracking with Online Multiple Instance Learning. *IEEE transactions on Pattern Analysis and Machine Intelligence*, pages 983–990, 2010.

[49] Tianzhu Zhang, Bernard Ghanem, Si Liu, and Narendra Ahuja. Robust visual tracking via multi-task sparse learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2042–2049, 2012.

[50] Kaihua Zhang, Lei Zhang, and Ming Hsuan Yang. Real-time compressive tracking. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 7574 LNCS, pages 864–877, 2012.

[51] Chenglong Bao, Yi Wu, Haibin Ling, and Hui Ji. Real time robust L1 tracker using accelerated proximal gradient approach. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1830–1837, 2012.

[52] Xue Mei, Haibin Ling, Yi Wu, Erik Blasch, and Li Bai. Minimum error bounded efficient L1 tracker with occlusion detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1257–1264, 2011.

[53] Xue Mei and Haibin Ling. Robust visual tracking and vehicle classification via sparse representation. *IEEE transactions on Pattern Analysis and Machine Intelligence*, 33:2259–2272, 2011.

[54] Baoxin Li, Rama Chellappa, Qinfen Zheng, and Sandor Z. Der. Model-based temporal object verification using video. *IEEE Transactions on Image Processing*, 10:897–908, 2001.

[55] Daniel Cremers and Christoph Schnörr. Statistical shape knowledge in variational motion segmentation. *Image and Vision Computing*, 21:77–86, 2003.

[56] Maria Alejandra Dávila. Seguimiento Visual de Objetos Articulados Utilizando Modelos Gráficos Basados en Energía, 2014.

[57] Juan Carlos Niebles, Bohyung Han, and Li Fei-Fei. Efficient extraction of human motion volumes by tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 655–662, 2010.

[58] Emilio Maggio and Andrea Cavallaro. Hybrid particle filter and mean shift tracker with adaptive transition model. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, volume II, 2005.

[59] S. Munder and D. M. Gavrila. An experimental study on pedestrian classification. *IEEE transactions on Pattern Analysis and Machine Intelligence*, 28:1863–1868, 2006.

[60] Alexandre Alahi, Laurent Jacques, Yannick Boursier, and Pierre Vandergheynst. Sparsity-driven People Localization Algorithm: Evaluation in Crowded Scenes Environments. In *IEEE International Workshop on Performance Evaluation of Tracking and Surveillance*, Snowbird, Utah, 2009.

[61] A. Torralba, K.P. Murphy, W.T. Freeman, and M.A. Rubin. Context-based vision system for place and object recognition. In *IEEE International Conference on Computer Vision*, 2003.

[62] Roberto Vezzani and Rita Cucchiara. Video surveillance online repository (ViSOR): An integrated framework. *Multimedia Tools and Applications*, 50:359–380, 2010.

[63] Arnold W M Smeulders, Dung M. Chu, Rita Cucchiara, Simone Calderara, Afshin Dehghan, and Mubarak Shah. Visual tracking: An experimental survey. *IEEE transactions on Pattern Analysis and Machine Intelligence*, 36:1442–1468, 2014.

[64] Dominik A. Klein, Dirk Schulz, Simone Frintrop, and Armin B. Cremers. Adaptive real-time video-tracking for arbitrary objects. In *IEEE Int. Conf. on Intelligent Robots and Systems (IROS)*, pages 772–777, Oct 2010.

[65] Luka Cehovin, Matej Kristan, and Ales Leonardis. Robust visual tracking using an adaptive coupled-layer visual model. *IEEE transactions on Pattern Analysis and Machine Intelligence*, 35:941–53, 2013.

[66] Zdenek Kalal, Krystian Mikolajczyk, and Jiri Matas. Tracking-Learning-Detection. *IEEE transactions on Pattern Analysis and Machine Intelligence*, 34:1409–1422, 2011.

[67] Junseok Kwon and Kyoung M. Lee. Tracking of a non-rigid object via patch-based dynamic appearance modeling and adaptive basin hopping monte carlo sampling. In *IEEE Conference on Computer Vision and Pattern Recognition- Workshops*, pages 1208–1215, 2009.

[68] Junseok Kwon and Kyoung Mu Lee. Tracking by sampling trackers. In *IEEE International Conference on Computer Vision*, pages 1195–1202, 2011.

[69] Jakob Santner, Christian Leistner, Amir Saffari, Thomas Pock, and Horst Bischof. PROST: Parallel robust online simple tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 723–730, 2010.

[70] Thang Ba Dinh, Nam Vo, and Gérard Medioni. Context tracker: Exploring supporters and distracters in unconstrained environments. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1177–1184, 2011.

[71] Martin Godec, Peter M. Roth, and Horst Bischof. Hough-based tracking of non-rigid objects. In *IEEE International Conference on Computer Vision*, 2011.

[72] Stan Birchfield. Elliptical head tracking using intensity gradients and color histograms. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 232–237, 1998.

[73] Qinxun Bai, Zheng Wu, S Sclaroff, M Betke, and C Monnier. Randomized Ensemble Tracking. In *IEEE International Conference on Computer Vision*, pages 2040–2047, 2013.

[74] Christian Bailer, Alain Pagani, and Didier Stricker. A Superior Tracking Approach: Building a strong Tracker through Fusion. In *European Conference on Computer Vision*, pages 170–185, 2014.

[75] Anil K. Jain and Richard C. Dubes. *Algorithms for clustering data*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1988.

[76] Jianming Zhang, Shugao Ma, and Stan Sclaroff. MEEM: robust tracking via multiple experts using entropy minimization. In *European Conference on Computer Vision*, 2014.

[77] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[78] Piotr Dollár. Piotr's Computer Vision Matlab Toolbox (PMT). `http://vision.ucsd.edu/~pdollar/toolbox/doc/index.html`.

[79] Robert Collins, Xuhui Zhou, and Seng Keat Teh. An Open Source Tracking Testbed and Evaluation Web Site. In *IEEE International Workshop on Performance Evaluation of Tracking and Surveillance*, 2005.

[80] Xu Jia, Huchuan Lu, and Ming Hsuan Yang. Visual tracking via adaptive structural local sparse appearance model. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1822–1829, 2012.

[81] Kaihua Zhang, Lei Zhang, and Ming Hsuan Yang. Real-time compressive tracking. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 7574 LNCS, pages 864–877, 2012.

[82] James Steven Supancic and Deva Ramanan. Self-paced learning for long-term tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2379–2386, 2013.

[83] JF Henriques, Rui Caseiro, Pedro Martins, and Jorge Batista. High-Speed Tracking with Kernelized Correlation Filters. *IEEE transactions on Pattern Analysis and Machine Intelligence*, pages 1–14, 2015.

[84] Kaihua Zhang, Lei Zhang, Ming-Hsuan Yang, and David Zhang. Fast Tracking via Spatio-Temporal Context Learning. In *European Conference on Computer Vision*, pages 1–16, 2013.

[85] Martin Danelljan, Fahad Shahbaz Khan, Michael Felsberg, and Joost Van De Weijer. Adaptive Color Attributes for Real-Time Visual Tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[86] D S Bolme, J R Beveridge, B a Draper, and Yui Man Lui Yui Man Lui. Visual object tracking using adaptive correlation filters. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2544–2550, 2010.

[87] Sam Hare, Amir Saffari, and Philip H S Torr. Struck: Structured output tracking with kernels. In *IEEE International Conference on Computer Vision*, pages 263–270, 2011.

[88] Wei Zhong, Huchuan Lu, and Ming-Hsuan Yang. Robust object tracking via sparsity-based collaborative model. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1838–1845, 2012.