Master Thesis

Visual Object Tracking applying Online Ensemble of multiple trackers

Author:
Jorge Martinez Gomez

Supervisor:
Juan Carlos Niebles

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

in the

Computer Vision Research Group
Electrical and Electronics Engineering Department

May 2015

"Thanks to my solid academic training, today I can write hundreds of words on virtually any topic without possessing a shred of information, which is how I got a good job in journalism."

Dave Barry

UNIVERSIDAD DEL NORTE

Abstract

ELECTRICAL AND ELETRONICS ENGINEERING DEPARTMENT
Computer Vision Research Group

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 paper, 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

In whatever you choose to do. Do it because it's hard, not because it's easy.

The acknowledgements and the people to thank go here, don't forget to include your project advisor...

Contents

List of Figures

List of Tables

Dedicated to my future brand new PS4

Chapter 1

Introduction

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. On the other hand, generative methods tend to suffer from model drifting and struggle to handle appearance variations.

In this paper, 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. Recently, the online tracking benchmark proposed in [**?**] 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.

The main contribution of this paper is an ensemble tracking framework that builds on top of the output of available online tracking algorithms running in parallel to produce 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.

The rest of the paper is organized as follows. We first briefly review the state-of-the-art of tracking algorithms in Section **??** and then present our online ensemble tracking algorithm in Section **??**. Section 4 illustrates quantitative and qualitative results of our tracker on a standard benchmarking dataset. Finally, we conclude the paper in Section **??**.

Chapter 2

Moving Object Detection Approaches, Challenges, Datasets and Object Tracking

An object can be considered simply as nothing but an entity of interest used for further analysis. These elements can be represented by their shape **Cite here** or appearance **cite color histograms, etc.**. In order to track objects, selecting the right features plays a critical role. In general, the most important property of a visual feature is its uniqueness so that could be easily distinguished from other objects. Mostly features are chosen manually by the user depending on the application domain. this problem of automatic feature selection has receibed significant attention in the pattern recognition community. The most common visual features selections are color, edges, displacement vectors and textures.

Among all features, color is the one of the most widely used feature for tracking. However, color features are sensity to illumination variation. To tackle this problem, in scenarios where this effect is inevitable, other features are incorporated to model object appearance.

2.1 Moving Object Detection

In a video, there are two sources of information that can be used for object detection and tracking: Visual features (color, texture and shape) and motion information. Robust approaches suggest that combining the statistical analysis of visual features and temporal

analysis of motion information. Moving object detection targets the extraction of moving objects that are of interest in sequences (e.g. people and vehicles).

A large number of methodologies have been proposed for object tracking, focusing on the task of object detection first. Most of them apply combinations and intersections 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.

2.1.1 Background Substraction

Background subtraction is a commonly used technique for object segmentation in static scenarios [3]. This task consist 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 (belongs to an object). The background image is created averaging images over time in an intiialization 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 (elimination of 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.

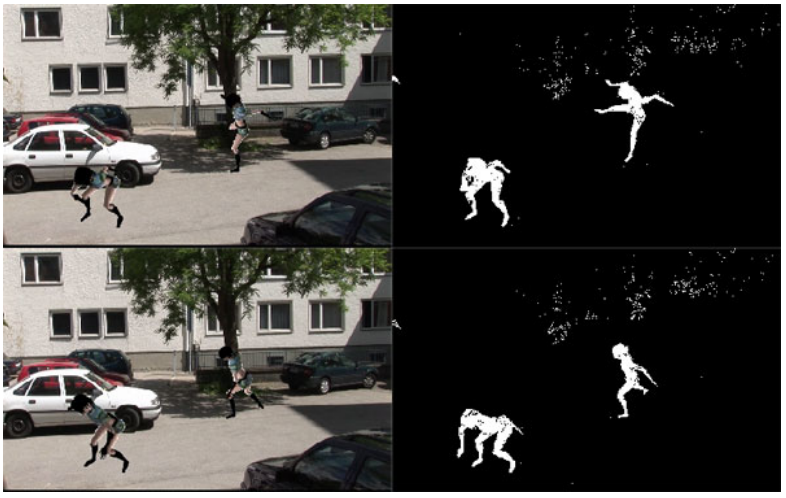


Figure 2.1: Object detection using Gaussian Mixture Models for background subtraction [1]. foreground pixels are drawn in white.

2.1.2 Temporal differencing

In temporal differencing, objects are detected by taking pixel-by-pixel difference of consecutive frames (generally two or three) in a video sequence. This method is most common for moving object detection in scenarios where camera is moving. Unlike static camera scenarios, the background is changing in time for moving camera (not appropiate to create a background model). Alternatively, 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 small regions as moving objects (ghost regions). Detection will not be correct also, for objects that preserve uniform regions (static objects).

A two-frame differencing method is presented in [4], 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 in some cases. For instance, a three-frame differencing method [5] and a hybrid method that combines three-frame differencing with an adaptive background subtraction model [6].

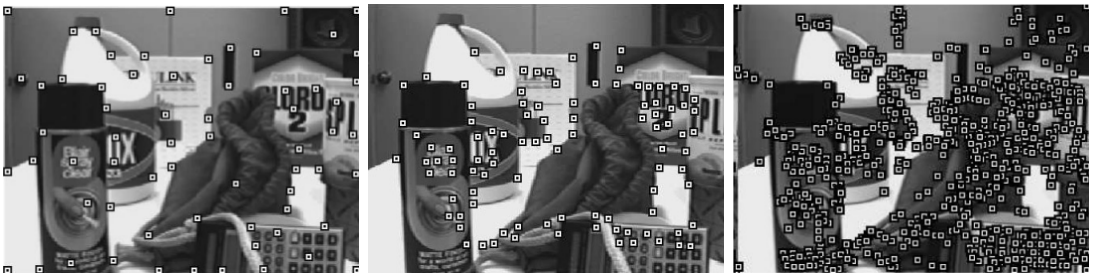
2.1.3 Statistical Approaches

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 pixels statistics 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. For instance, some approaches apply Hidden Markov Models (HMM). These methods [2, 7] represent the intensity variation of a pixel in an image sequence as discrete states

The statistical method proposed in [1] describes and 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 is evaluated in order to detect whether a pixel belongs to foreground and background.

2.1.4 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 approach [8]. These methods make 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 [9] and Harris [10] corners detectors.



2.2 Challenges

Object detection and tracking is still an open research problem in computer vision. A robust, accurate and high performance approach is still a great challenge. The level of difficulty depends on how the object of interest is defined in terms of features. For instance, Using color as object representation method, it is not difficult to identify all pixels with same color as the object. However, there is always a probability of existence a background region with same color information (background clutter). In addition, illumination changes in the scene does not guarantee that the pixel values of an object will be the same in all frames. These variabilities or challenges which are random in object tracking causes wrong object tracking, and are listed below.

- **Illumination Variation (IV):** It is desirable that background model adapts to gradual changes of the appearance of the environment.
- **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 o the target leaves the view.
- **Background clutter (BC):** As stated before, this challenge makes the segmentation task difficult. It is hard to create ans 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. In the other case, 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.

2.3 Tracking Datasets

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 in the list, 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.

The Surveillance Performance Evaluation Initiative (SPEVI) 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. ETISEO dataset contains indoor and outdoor scenes, such as corridors, buildings entries, etc. This dataset can be used for surveillance applications.

PETS dataset became a surveillance project whose challenging scenarios are focused only on high level applications of 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 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) 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. The VIdeo Surveillance Online Repository, includes several sequences for two separate tasks: First, an abandoned baggage scenario and second, a parked vehicle scenario.

In generic visual tracking, a dataset is a collection of videos that contains and object moving in some scenario. The sequences vary in length from hundreds of frames to thousands. Diverse object types are used. Different scene settings (indoor or outdoor, static or moving camera). Also different challenges, such as object occlusions or illumination conditions are presented. Most commonly used tracking benchmarks are summarized in table 2.1. Recently, the authors in [11] released a benchmark containing 50 most commonly used sequences from some datasets mentioned 2.1, to facilitate fair performance evaluation. Also for better evaluation and analysis of strengths and weakness of tracking approaches. They classified sequences, considering a object tracking challenge, as a category, constructing several subsets to report specific challenging conditions. Some attributes occur more frequently, and some sequences are annotated with several attributes.

Name/Author/Paper	Sequences
Babenko	3
Bobot	12
Cehovin	5
Ellis IJCV2011	3
Godec	7
Kalal	10
Kwon	4
Kwon VTD	11
PROST	4
Ross	4
Thang	4
Wang	4

Table 2.1: Popular object tracking datasets

2.4 Object Tracking

The goal of an object tracker is to generate an object path over time. This trajectory consists of the object position over time in every frame of the video. The tracker may provide complete region in the image that is occupied by the object at every time instant. Certainly, this list is not meticulous and covers popular approaches on each category.

2.4.1 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 [12] or greedy search methods.

Statistical methods for Point Tracking: Statistical correspondence methods solve tracking problems whose measurements obtained from video sensors contain nose, or object motion can undergo random perturbations. These approaches take measurements and model uncertainties into account during object state estimation. Applying state space approach to model the object properties such as position, velocity and acceleration. In single object state estimation, tYhe optimal state of an object is giben by the Kalman Filter [13, 14], assuming measurement noise have a Gaussian distribution. In the general case, that is, object state is not assumed as Gaussian, estimation can be performed using particle filters [2, 15].

In the case of multiobject data association, state estimation using Kalman or particle filters, it is necessary to solve first correspondence problem before these filters can be applied. However, in cases when two objects are close each other, the correspondence could be incorrect. Then, an incorrectly associated measurement can cause the filter to fail to converge. In order to tackle this problem, Joint Probability Data Association Filtering (JPDAF) [16] and Multiple Hypothesis Tracking (MHT) [17] are two used techniques for data association.

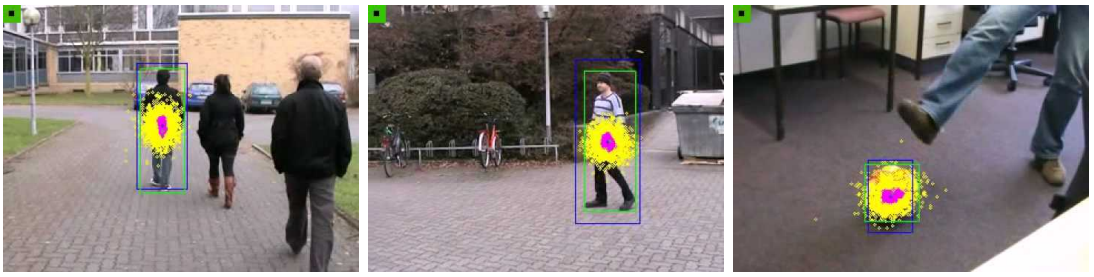


Figure 2.2: [2]

2.4.2 Kernel Tracking

In this type of tracking, 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 (features extraction) used, the number of objects tracked, and the method used for object motion estimation.

Density-based tracking: According to [18], the object is modelled with one or more probability density functions, such as Gaussian, mixture of Gaussian, Parzen windows or histogras, that describe the probability of object appearance. Mean-shift is an approach to feature space analysis. This method shifts a data point to the average of data points in its neighborhood. Mean shift uses fixed color distribution. A similar approach is called CAMSHIFT [19] that handles dynamically changing color distribution by adapting the search window size and computing color distribution in the search 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* [20] that searchs accross the image, a region similar to the object template, defined in previous frames. The 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 near previous position.

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 the rectangular or ellipsoidal regions. To reduce computational complexitu, the similarity between object model and the hypothesized position, is computed evaluating the ratio betweem color means between model and position. The position with highest ratio is selected as current object location.

"Tracking by detection" or "Tracking by repeated recognition" [21] systems generally perform target object appearance learning. These methods are closely related to object detection (an area with great progress in computer vision) and has encouraged some

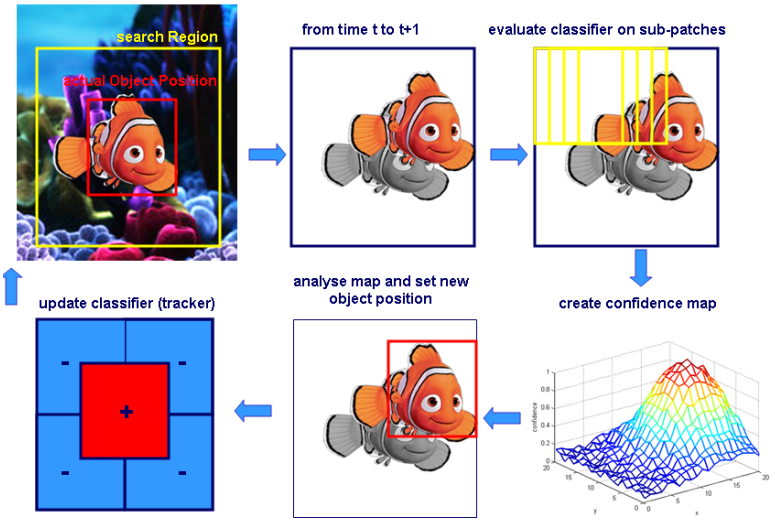


Figure 2.3: Overview for online boosting object tracker

successful real-time tracking algorithms [22, 23]. However, many tracking algorithms employ static appearance models that are defined manually or trained at the first frame only [24–28], these methods are often unable to deal with significant appearance changes. This situations are difficult when there is limited knowledge of the object of interest. In order to cope this problem, an adaptive appearance model that changes during the tracking process as the appearance of the object changes, gets better results [29–31].

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 [32] and is related to the work on Online Adaboost [33–35], multi-class boost [36] and MILBoost [37]. The goal of boosting is to combine many weak classifiers (usually decision stumps) into a linear strong classifier.

2.4.3 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. These approaches can be divided into two main categories, shape matching and contour tracking. Shape matching [38] approaches search object silhouette in the current frame. Contour based, evolve initial contour to its new position in the current frame using state space models or direct minimization of some energy function [39].

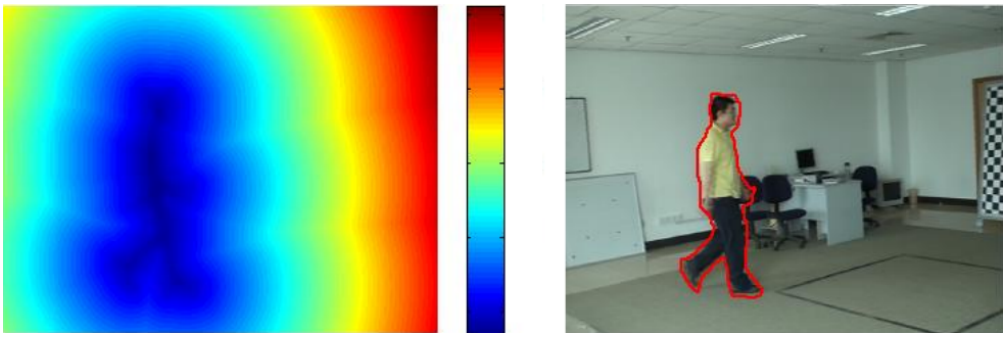


Figure 2.4: Illustration of an active contour representation. Left subfigure shows the signed distance map of a human contour; right image displays contour result.

2.4.4 Tracking applying ensemble of trackers

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 [33] or Bayesian probabilistic fusion [40], 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 [41], 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 VTD [42] and VTS [43] 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 [44], 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

In this section, we first present an overview of our online ensemble tracking approach, and then describe the details of each stage of our processing pipeline.

3.1 Overview

We illustrate the main components of our method in Figure 3.1. 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. 3.1a). These predictions are the raw input of our ensemble algorithm, and may be in the form of bounding boxes. 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. 3.1b). 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. 3.1c). A final ensemble prediciton of the target state is then derived from these inlier estimations (Fig. 3.1d). Finally, we use the prediction of the ensemble to update a model of the target appearance, and periodically reinitialize outlier trackers (Fig. 3.1e). 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.

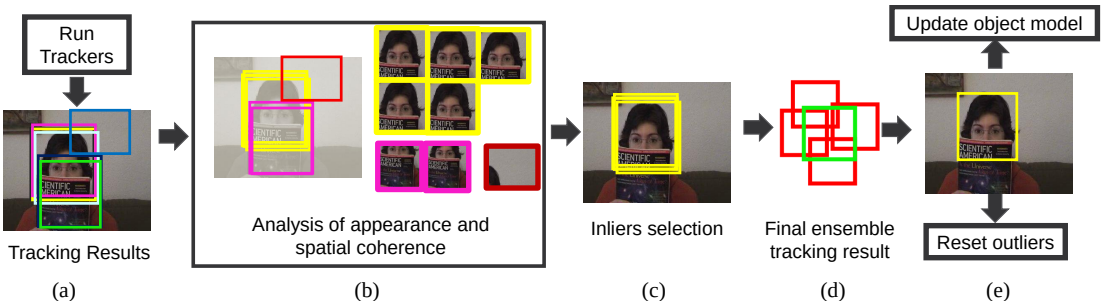


Figure 3.1: Basic diagram of our approach. On each frame, we analyze coherence between tracking results. We apply spatial and appearance models with the goal of finding inliers. Then, we select best tracker that follows the target. Finally, we reset outliers and update object model on necessary cases.

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 of the appearance and spatial coherence stage (Fig. 3.1b). The goal is to determine the set of trackers that correctly estimate the true target state. We denote these trackers as inliners in Fig. 3.1c.

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 objects. These objects are described as a set of measurements or by relationships between the object and other objects. A *cluster* is comprised of a number of similar objects collected or group together. Other authors define a cluster as a set of entities which are alike, and entities from different clusters are not alike, or "A cluster is 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". This theory is taken from **Jane and Dubes.**

Cluster analysis is the process of classifying objects into subsets that have meaning in the context of a particular problem. The objects 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 a pattern. 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 objects 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 objects 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) [**?**]. 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. 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. 3.1b. These clusters satisfy the following:

- $c_i \cap c_j = \emptyset$ for i and j from 1 to m, $i \neq j$

- $c_1 \cup c_2 \cup ... \cup c_m = T$

3.3 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 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 [23, 34, 37], 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 Nachines (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 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 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. We considered this model, because other methods found in [40**?**], obtained better results. Also, HOG features are more robust to deformable objects than other image features used in tracking-by-detection methods *e.g.* Haar.

3.3.1 Selection of inliers

The clustering stage and appearance scoring provide cues about which trackers can be considered as correctly tracking the target. We evalute two simple criteria that optimally

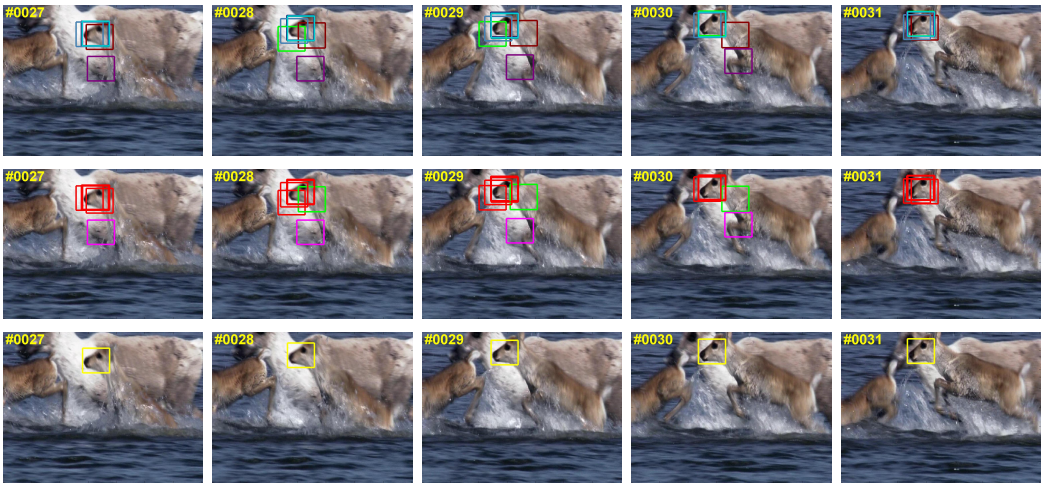


Figure 3.2: System behavior in five frames. Given image input, trackers will give results on where the object might be (first row). Then, all results are clustered using hierarchical agglomerative clustering (second row). We focus on selecting the group with highest number of members, or cluster with best appearance score (third row). Finally, we reiniialize outliers (frame 30 in this case).

select a group of trackers as inliers (Fig. 3.1c). In our experiments, we consider cluster size and appearance scores as potential cues to select inliers.

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 inliners, 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:

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

3.3.2 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 on how to implément 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:

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

3.3.3 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 consistently tagged as an outlier, our algorithm automatically resets its state to the latest target state prediction from the ensemble. 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.

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 complete 50 sequences tracking benchmark [**?**]. We also present qualitative and quantitative analysis of our ensemble tracker, such as individual performance of trackers and effect of the tracker pool.

4.0.4 Experimental setup

In order to evaluate the performance of our proposed ensemble tracking, we adopt the Online Object Tracking Benchmark from [**?**]. 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.

We implement our online ensemble tracking algorithm in MATLAB. All experiments are performed on desktop with an Intel Xeon CPU and 16 GB of RAM.

4.0.5 Evaluation Methodology

To validate the performance of our proposed approach, we follow the one-pass evaluation methodology (OPE) proposed in [**?**]. We summarize the performance using the precision and success plots.

In *precision*, a frame is considered correctly tracked if the predicted target center is within a distance threshold of the ground truth. A higher precision at low thresholds

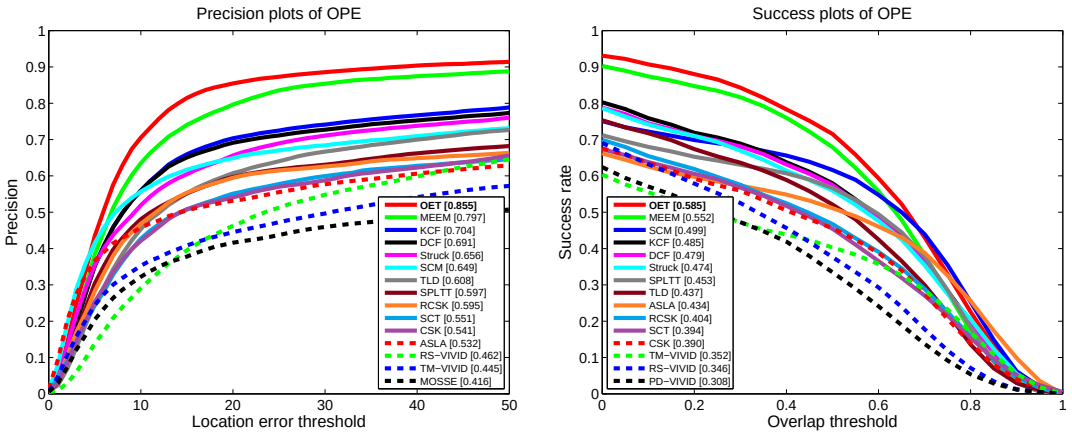


Figure 4.1: Precision and success plots for all 50 sequences. Precision and success ratios are measured by center location error and overlap ratio, respectively. Trackers are ranked using scores of 20 pixels for precision and AUC for success.

means the tracker is more accurate, while a lost target will not achieve perfect precision on a large threshold range. We rank tracking algorithms using their performance at 20 pixels as in [37**? ?**].

Success measures intersection over union overlap between a tracker output and a ground truth box and evaluates whether it exceeds a threshold. This overlap measure penalizes if the scale of the target is estimated incorrectly. In contrast to precision, trackers are ranked using area under the curve (AUC) metric, which relates to the average overlap across all frames.

4.0.6 Tracker pool

We integrate into our pool tracking algorithms whose original source code is publicly available. Table 4.1 shows the list of the selected tracking algorithms.

4.0.7 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 [**?**], Sparse Collaborative Model (SCM) [**?**], TLD tracker [**?**]. Our online ensemble tracking algorithm is denoted *OET*.

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

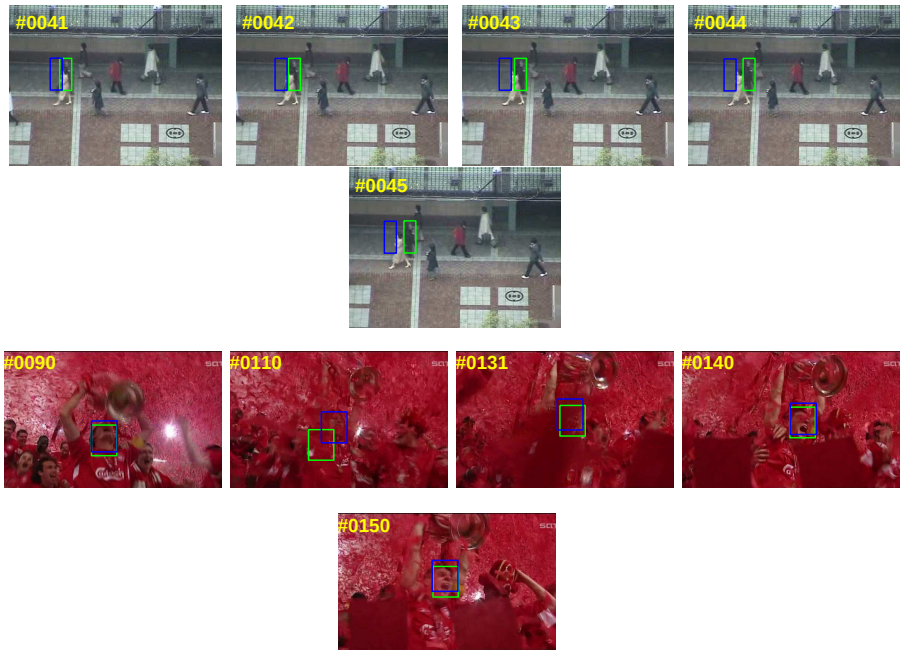


Figure 4.2: Qualitative results for object tracking applying both selection criterias in two sequences. **Green** bounding box corresponds to appearance score selection. **Blue** box to cluster size. In *subway* when occlusion happens, many trackers are lost, creating a big cluster. In cases of background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

sequences, we note that out method handles better camera rotation sequences, 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 *jogging, david3*, KCF trackers fails tracking the object.

4.0.8 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 3.3.1 (appearance score and cluster size).

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

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

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

Table 4.2: Average AUC and precision for live fusion methods tested in 50 videos dataset.

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

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% (∼300 frames) in success and 10% in precision (∼500 frames). For instance, when 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.

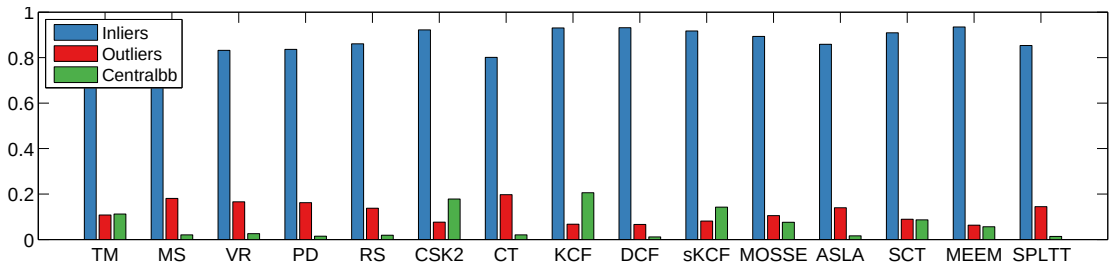


Figure 4.3: Statistics for each tracker in the ensemble over all frames.

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.

On the other hand, we present statistics about how often trackers in the pool are selected as outlier, inliner, and medoid. In figure 4.3, 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).

4.0.9 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.

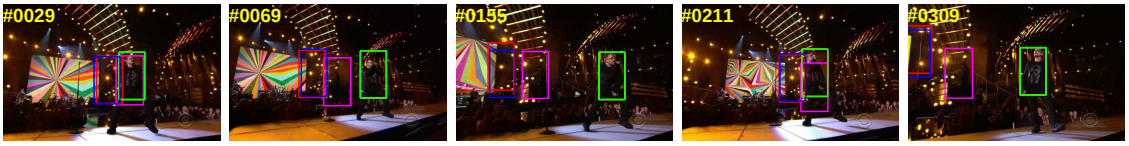
From figure 4.5, our approach using appearance selection outperforms other trackers in 8 of 11 attributes. Specifically, in attributes such as IV (illumination variation), OPR (out of plane rotation), OCC (occlusion), DEF (deformation), MB (motion blur), IPR (in plane rotation), OV (out of view), and 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 [**?**]. In our system, scale is not

(a) bolt sequence.

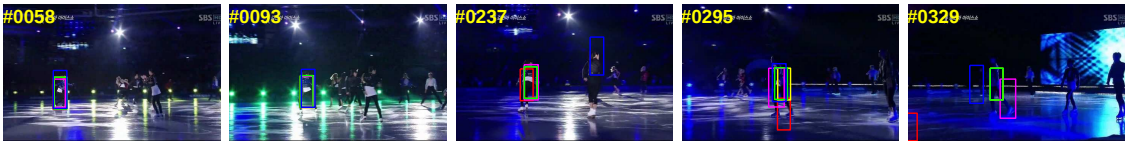
(b) jogging-2 sequence.

(c) singer-1 sequence.

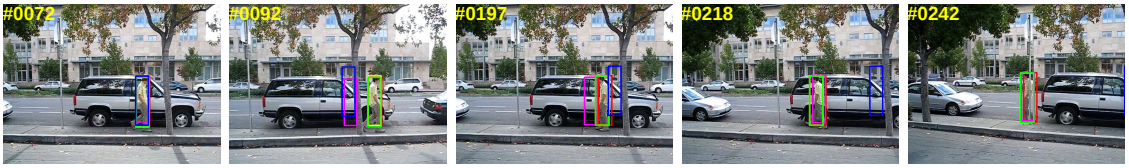
(d) football sequence.



(e) singer-2 sequence.



(f) skating1 sequence.



(g) david3 sequence.

Figure 4.4: Screenshots of tracking results.

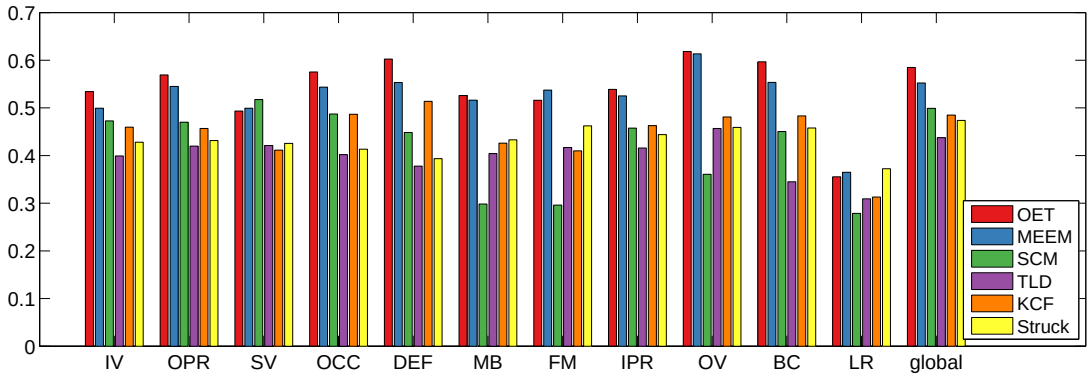


Figure 4.5: Average AUC ranking scores of top trackers on different subsets of test sequences in OPE. Each subset of sequences corresponds to an attribute: IV - illumination variation, OPR - out of plane rotation, SV- scale variation, OCC - occlusion, DEF - deformation, MB - motion blur, FM - fast motion, 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.

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.

Bibliography

[1] 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.

[2] J Rittscher, J Kato, S Joga, and A Blake. A probabilistic background model for tracking. *Computer Vision—ECCV 2000*, pages 336–350, 2000.

[3] Am McIvor. Background subtraction techniques. *Proc. of Image and Vision Computing, ...*, 2:13, 2000.

[4] 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 WACV98 Cat No98EX201*, 98:8–14, 1998. ISSN 09031936.

[5] Liang Wang, Weiming Hu, and Tieniu Tan. Recent developments in human motion analysis. *Pattern Recognition*, 36:585–601, 2003. ISSN 00313203.

[6] 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. ISSN 19406029.

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

[8] Jianbo Shi Jianbo Shi and C. Tomasi. Good features to track. *Computer Vision and Pattern Recognition, 1994. Proceedings CVPR '94., 1994 IEEE Computer Society Conference on*, 1994. ISSN 1063-6919.

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

[10] C. Harris and M. Stephens. A Combined Corner and Edge Detector. *Procedings of the Alvey Vision Conference 1988*, pages 147–151, 1988. ISSN 09639292.

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

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

[13] Xiaofeng Ren. Finding people in archive films through tracking. In *26th IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2008.

[14] Janne Heikkilä and Olli Silvén. A real-time system for monitoring of cyclists and pedestrians. *Image and Vision Computing*, 22:563–570, 2004. ISSN 02628856.

[15] Kenji Okuma, Ali Taleghani, Nando De Freitas, James J Little, and David G Lowe. A Boosted Particle Filter : Multitarget Detection and Tracking. *Proceedings of the 8th European Conference on Computer Vision - ECCV 2004*, pages 28–39, 2004. ISSN 03029743.

[16] Dirk Schulz, Wolfram Burgard, Dieter Fox, and Armin B. Cremers. People Tracking with Mobile Robots Using Sample-Based Joint Probabilistic Data Association Filters, 2003. ISSN 02783649.

[17] Mohd Asyraf Zulkifley, Bill Moran, and David Rawlinson. Robust hierarchical multiple hypothesis tracker for multiple object tracking. In *Proceedings - International Conference on Image Processing, ICIP*, pages 405–408, 2012.

[18] Yizong Cheng. Mean shift, mode seeking, and clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17:790–799, 1995. ISSN 01628828.

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

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

[21] 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. ISSN 01628828.

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

[23] Helmut Grabner, Michael Grabner, and Horst Bischof. Real-Time Tracking via On-line Boosting. *Technology*, 1:1–10, 2006. ISSN 0162-8828.

[24] M. Isard and J. MacCormick. BraMBLe: a Bayesian multiple-blob tracker. *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, 2, 2001.

[25] Vincent Lepetit and Pascal Fua. Keypoint recognition using randomized trees. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28:1465–1479, 2006. ISSN 01628828.

[26] 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. ISSN 0920-5691.

[27] D Comaniciu, V Ramesh, and P Meer. Real-time tracking of non-rigid objects using mean shift. *IEEE Conference on Computer Vision and Pattern Recognition*, 2:142–149, 2000. ISSN 01628828.

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

[29] 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. ISSN 0920-5691.

[30] Iain Matthews, Takahiro Ishikawa, and Simon Baker. The template update problem. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26:810–815, 2004. ISSN 01628828.

[31] 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. ISSN 0162-8828.

[32] 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. ISSN 00220000.

[33] Shai Avidan. Ensemble tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29:261–271, 2007. ISSN 01628828.

[34] 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.

[35] NC Oza and S Russell. Online ensemble learning. *AAAI/IAAI*, 6837:1109–1109, 2000.

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

[37] 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. ISSN 1939-3539.

[38] 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. ISSN 10577149.

[39] Daniel Cremers and Christoph Schnörr. Statistical shape knowledge in variational motion segmentation. *Image and Vision Computing*, 21:77–86, 2003. ISSN 02628856.

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

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

[42] 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 *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2009*, pages 1208–1215, 2009.

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

[44] 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.