A survey of datasets for visual tracking

Séverine Dubuisson, Christophe Gonzales

Noname manuscript No.
(will be inserted by the editor)

A survey of datasets for visual tracking

Séverine Dubuisson · Christophe Gonzales

Received: date / Accepted: date

Abstract For fifteen years now, visual tracking has been a very active research area of the computer vision community. But an increasing amount of works can be observed in the last five years. This has led to the development of numerous algorithms that can deal with more and more complex video sequences. Each of them has its own strengths and weaknesses. That is the reason why it becomes necessary to compare those algorithms. For this purpose, some datasets dedicated to visual tracking as well as, sometimes, their ground truth annotation files, are regularly made publicly available by researchers. However, each dataset has its own specificities and is sometimes dedicated to test the ability of some algorithms to tackle only one or a few specific visual tracking subproblems. This article provides an overview of some of the datasets that are most used by the visual tracking community, but also of others that address specific tasks. We also propose a cartography of these datasets from a novel perspective, namely that of the difficulties the datasets present for visual tracking.

Keywords visual tracking · datasets · survey

1 Introduction

"Visual tracking refers in general to the ability of the eyes to quickly and accurately look (fixate), visually follow a moving object (pursuit) and efficiently move our eyes so we can fixate on objects from point to point as in reading (saccades)" (Doctor Dan L. Fortenbacher, specialized in vision therapy). From the computer vision community point of view, visual tracking is the process of locating, identifying, and determining the dynamic configuration of one or many moving (possibly deformable) objects (or parts of objects) in each frame of one or several cameras [193,166]. As such, this involves solving many different challenging problems. What makes a good visual tracking algorithm is its capacity to handle all the variability in a video sequence caused by the tracked object, the scene and the camera acquiring the scene. Such variability can be caused, for instance, by pose and illumination variations, occlusions, varying and erratic motions. The algorithm should also be robust to cluttered backgrounds, blur, poor video acquisition quality and to the similarity between multiple objects to track. The combinations of these issues makes visual tracking very challenging, and more and more emerging applications, such as video surveillance, require the algorithms to work under such very difficult conditions. This has led researchers to develop many visual tracking algorithms that can tackle, at least partially, some of these issues. They all have their own strengths and weaknesses and, to compare them, some datasets are regularly made publicly available.

The goal of this paper is not to perform a review of all existing visual tracking algorithms (see [193,166] for two points of view on this matter, and [235,156, 157] for benchmarks). This would be impossible anyway: indeed, too many works have been proposed and are still proposed every year since the middle of the 90's. Rather, in this paper, we are interested in the way to evaluate and compare visual tracking algorithms. Our

S. Dubuisson
Sorbonne Universités, UPMC Univ Paris 06, CNRS, UMR 7222, 4 place Jussieu 75005 Paris, France
E-mail: severine.dubuisson@isir.upmc.fr

C. Gonzales
Sorbonne Universités, UPMC Univ Paris 06, CNRS, UMR 7606, 4 place Jussieu 75005 Paris, France.
E-mail: christophe.gonzales@lip6.fr

study is mainly articulated around an original cartography of visual tracking datasets depending on the issues involved in their video sequences. To our knowledge, this is the first study from this point of view. Of course, there exist websites providing some dataset repositories [152] but, more often than not, they do not describe the issues they address and, individually, each one only provides a small subset of all the datasets available. Therefore, they do not completely answer the above question. We hope that our cartography will be useful in that it gathers in a principled way many datasets scattered over the web. As such, we expect it should help researchers selecting easily the most appropriate datasets they need in order to evaluate their tracker's performances on specific visual tracking problems. As a side effect, this cartography should also prove useful for the evaluation of the components involved in more complex tasks than just visual tracking, e.g., human behavior or dynamic scene analysis.

This article is organized as follows. Section 2 first defines the notion of visual tracking as a specific task and a unique goal, by opposition to it being only a component of a more complex framework where visual tracking is not a goal but a mean (e.g., it is only a subproblem involved in scene analysis problems). This distinction enables to discriminate existing datasets that are only devoted to visual tracking, i.e., those which are designed explicitly as challenges to visual tracking algorithms, from datasets whose challenges are not only related to visual tracking. Section 3 describes the different issues that naturally arise in visual tracking and the way a visual tracking algorithm can be evaluated. Section 4 presents datasets that are most useful to test specific properties of visual tracking algorithms. In particular, some of the algorithms that were tested on these datasets are given, as well as the visual tracking issues involved in the datasets. Section 5 is dedicated to a classification of datasets used for research problems in which visual tracking is a mean and not a goal. Here, two main classes of such problems are considered: human motion understanding and scene dynamics understanding. Finally, concluding remarks and perspectives are given in Section 6.

2 Visual tracking: which dataset for which need?

In the rest of the paper, *visual tracking* will refer to the problem of estimating over time the position or the shape (rectangle, ellipse, *etc.*) of moving objects in the scene, without explicitly estimating their inter deformation, or managing their interactions with other objects and/or with the environmental context (see [193,166] for more details). *Per se*, visual tracking is a goal in itself. By opposition, we will refer to *dynamic scene analysis or understanding* as a more general task in which visual tracking is only one component. As such, the latter is thus more general than visual tracking; it can consider inter deformations, interactions with other objects, analyses of the output of visual tracking (human behavior analysis, for instance), *etc.*

Designing a visual tracking algorithm requires considering three main tasks, that necessitate to answer the following questions:

- Information extraction. Which information of the current frame is relevant to discriminate the tracked object from the rest of the scene?
- Object representation. How to represent the tracked object in a compact way so that most of its features (shape, appearance, *etc.*) are considered?
- Motion model. How to model the motion of the tracked object in a recursive way so that its instance at time t can be linked to the previous one?

Of course, by our definition, designing a visual tracking algorithm requires addressing the above three tasks but also two additional concerns related to the very goal of the dynamic scene analysis or understanding algorithm:

- Multi-object trajectories handling. How to associate current objects' positions and previous estimated trajectories? How to manage occlusions, appearance and disappearance?
- Dedicated application. Which information can be extracted from visual tracking to achieve specific goals?

Unlike the other questions, the last one implies to add information to correctly achieve the task (*i.e.*, the target application). Applications of visual tracking are numerous and include surveillance, biomedical applications, robotics, media production, *etc.*

As we have chosen to mainly focus on visual tracking datasets, we provide in Section 4 as exhaustive as possible a list of datasets dedicated to the evaluation of (and only of) a visual tracker's output, and we detail their features as well as the ground truth they provide. But as visual tracking in general is important for computer vision and has many applications, we also provide in Section 5 an overview of datasets dedicated to it.

3 Visual tracking issues

In this section, we address the issues raised by visual tracking and we detail the different metrics that can be

used to evaluate and compare the output of a visual tracking algorithm. In this article, we consider that the output of a visual tracker is a shape surrounding the tracked object.

3.1 Problems inherent to visual tracking

Visual tracking in video sequences raises many problems that can potentially cause a loosing track on the object. We provide below a list of some of the most challenging ones.

Illumination effects The ambient light of the scene can change due to outdoor conditions (weather, hour of the day) or indoor conditions (switching on or off lights). In the same spirit, shadows can change illumination conditions. Finally, non-lambertian, transparent or reflective objects can also occur in the scene whose appearance can vary depending on the incoming light or point of view. These kinds of variations make the color distribution of the object vary over time, which can confuse the visual tracking algorithm.

Scene clutter This can be due to a very textured background, or to other moving objects in the scene, sometimes similar to the tracked object. This feature can cause some drifts of the visual tracker, resulting in loosing track on the object.

Changes in object appearance Due to the projection of 3D movements onto a 2D plane (frames from sequences), the tracked object can have geometric deformations (caused by rotations about the depth axis for example). For non rigid objects, the shape can be deformed over time (for example a walking person). In addition, there can also be some changes in appearance caused by the object itself like facial expression changes or the addition or removal of some clothes. As for illumination changes, the appearance changes induce a variation of the object's color distribution over time.

Abrupt changes in motion The object velocity can vary with time. This can make the object very hard to track because its movement can become unpredictable and, therefore, the object can be lost.

Occlusions The object to track can be occluded by other objects of the scene, or can even be self-occluded (for instance, a walking person can sometimes self-occlude one of her arms or legs). Occlusions are difficult to handle because some parts of the object (sometimes nearly the whole object) can disappear from the scene.

Similar appearances When different objects have similar appearances in the video sequence (e.g., when tracking some people in a crowd or cars on a road, *etc.*), it may be difficult for the visual tracker to discriminate between these appearances to correctly track its target(s).

Camera motions With new applications of visual tracking (for video surveillance, for example), it is sometimes necessary to embed a camera into a moving vehicle (car, motorbike, drone, *etc.*), or on other people. In such cases, it is difficult to discriminate the motion due to the object movements from that due to the camera. In addition, the motion induced by an embedded camera can be very irregular and is often not modeled. It is characterized in frames by motion blurs that strongly degrade their quality and, therefore, make visual tracking harder.

Appearance/disappearance The tracked object can enter or leave the view field on one edge of the frame but also inside the frame due to the presence of another object (for instance, a wall in the middle of the scene can hide the object temporarily, or a person can leave the room by a door, *etc.*). In such cases, it is necessary for the visual tracker to keep in memory the object and find it again when it reappears in the field of view (if it ever does). Appearance/disappearance differ from occlusion in the sense that, in the latter, at least a small part of the tracked object is still visible.

Quality of frames This property directly follows from the sensor or from the acquisition conditions of the video sequence. For example, blur can appear if the object or the camera is moving too quickly. Block artifacts can be visible if the video sequence was compressed. All these can of course confuse the visual tracker.

A complete visual tracking algorithm should be able to deal with all the above problems occurring in the video sequence, which, in itself, is still a challenge. Of course, such a visual tracker should stay robust over time, *i.e.*, losing its target(s) should be an exception and, in this case, it should be able to quickly reacquire it/them. In addition, this robustness and the quality of the visual tracking should not be at the expense of computational efficiency. To our knowledge, today, there exists no algorithm that satisfies all these features, especially when dealing with long-term visual tracking, *i.e.*, when tracking some objects during a very long period of time (*e.g.*, several hours). This is still a challenge for the future visual tracking algorithms, with many applications such as video surveillance.

A qualitative evaluation of the output of the visual tracker can be made, but this may be insufficient sometimes, especially when visually, two (or more) visual tracker's outputs seem to be very close. This is the reason why a quantitative evaluation is necessary, and for that, different metrics have been proposed, that are listed in the next subsection.

3.2 Quantitative evaluation of visual tracking

The quality of the output of a visual tracker can only be evaluated if a precise and efficient ground truth (GT) is available with the tested video sequences. In all existing evaluation protocols, the quality of visual tracking is measured by a distance or similarity between the visual tracker's output and the GT. In the rest of the paper, let us denote by $\mathcal{A}_t^{GT}$ the annotated area corresponding to the object in the GT, and by $\mathcal{A}_t^T$ the output of the visual tracker at time step t. Here, by area, we mean a set of pixels.

Error score. The easiest way to evaluate the quality of visual tracking is to measure at each time step t the distance between the centers of the GT's area and of the visual tracker output's area. This is given by the center location error:

$$S_t^{Err} = d\left(Center(\mathcal{A}_t^{GT}), Center(\mathcal{A}_t^T)\right). \quad (1)$$

Here, d is most of the time the Euclidean distance. Note that this is possible to weight the error according to its value: in such cases, d should be defined as a Lp-norm, i.e.:

$$S_t^{Err} = \sqrt[p]{\|Center(\mathcal{A}_t^{GT}) - Center(\mathcal{A}_t^T)\|^p} \quad (2)$$

A curve of the evolution of the center location error throughout the sequence can then be drawn. S_t^{Err} can also be averaged over all the sequence (T frames) to get the overall trajectory error $S^{traj} = \frac{1}{T}\sum_{t=1}^{T} S_t^{Err}$.

Accuracy score. The accuracy can be measured by the overlapping rate between the GT's area and that resulting from the visual tracker. It is defined as follows:

$$S_t^{Acc} = \frac{|\mathcal{A}_t^{GT} \cap \mathcal{A}_t^T|}{|\mathcal{A}_t^{GT} \cup \mathcal{A}_t^T|}. \quad (3)$$

Success score. This score allows to give an overall evaluation of the visual tracker throughout the sequence (of T frames). It is defined by:

$$S^{Succ} = \frac{\#Success}{T} \quad (4)$$

Here, a "Success" is a frame for which the output of the visual tracker (evaluated for example by one of the aforementioned scores) is considered to be correct. A score corresponds to a "Success" if it is larger than a given threshold.

For the error score S_t^{Err}, a threshold is hard to fix since it depends on the size of the objects, the point of view, *etc.* As an example, it is fixed to 20 pixels in [160]. This error threshold is called here T^{Err}, and we have

$$S_{T^{Err}}^{Succ} = \frac{|\{t : S_t^{Err} < T^{Err}\}|}{T}.$$

Concerning the accuracy score, the majority of the works consider that it is successful if it is larger than 0.5 (i.e., at least half of the object's area is overlapped by the visual tracker's output area). If we call T^{Acc} this threshold, we get:

$$S_{T^{Acc}}^{Succ} = \frac{|\{t : S_t^{Acc} > T^{Acc}\}|}{T}.$$

Detection scores. The visual tracker's output can be evaluated as for a classification problem: is the tracked object well detected of not? And how well is it detected? For this purpose, it is customary to define a pixel to be a true positive (TP) (resp. true negative (TN)) whenever this pixel belongs to both $\mathcal{A}_t^T$ and $\mathcal{A}_t^{GT}$ (resp. to none of these sets). Similarly, a pixel is considered to be a false negative (FN) (resp. a false positive (FP)) if it belongs to $\mathcal{A}_t^{GT}$ but not to $\mathcal{A}_t^T$ (resp. to $\mathcal{A}_t^T$ but not to $\mathcal{A}_t^{GT}$). To assess how well the tracked object is detected, three scores are defined at each time step t, $Precision_t$ (the proportion of estimated object pixels that are correct), $Recall_t$ (the proportion of object pixels that are correctly estimated) and the $F_{1,t}$ measure (the harmonic mean of $Precision_t$ and $Recall_t$) as:

$$Precision_t = \frac{|TP|}{|TP| + |FP|} = \frac{|\mathcal{A}_t^{GT} \cap \mathcal{A}_t^T|}{|\mathcal{A}_t^T|} \quad (5)$$

$$Recall_t = \frac{|TP|}{|TP| + |FN|} = \frac{|\mathcal{A}_t^{GT} \cap \mathcal{A}_t^T|}{|\mathcal{A}_t^{GT}|} \quad (6)$$

$$F_{1,t} = 2\frac{Precision_t \times Recall_t}{Precision_t + Recall_t} \quad (7)$$

A weighted version of the $F_{1,t}$ measure is also given by:

$$F_{\beta,t} = (1+\beta^2)\frac{Precision_t \times Recall_t}{\beta^2 Precision_t + Recall_t}$$

where $\beta \in \mathbb{R}^+$ allows to adjust the $Precision_t$'s importance. Note that different definitions can be used for $Precision_t$ [220]. $Precision_t$ and $Recall_t$ should have high values, as well as the $F_{1,t}$ measure. As for the S_t^{Err} measure, a curve of the evolution of these 3 measures throughout the sequence can be drawn and their average over all the sequence (T frames) can be computed to get the overall trajectory measures.

Curves for comparing tracking algorithms. Different curves can be drawn to compare visual tracking algorithms. One can for example cite:

- The visual tracking error curve: is a plot of S_t^{Err} depending on t. This curve should be as horizontal as possible, with low values.
- The visual tracking accuracy curve: is a plot of S_t^{Acc} depending on t. This curve should be as horizontal as possible, with high values.
- The success curves: are plots of $S_{T^{Err}}^{Succ}$ or $S_{T^{Acc}}^{Succ}$ depending on T^{Err} or T^{Acc} respectively. The first curve increases with T^{Err}, whereas the second decreases with T^{Acc}. The success of a visual tracker can be measured by the AUC (Area Under Curve) that should be as high as possible.

Robustness of evaluation A good visual tracker should have small average errors/failures, but also small variations in this error/failure rate. This is the reason why it is also necessary to evaluate the robustness of the visual tracker to the initialization conditions. The initialization can vary both in time and in space. We can then get a mean and a standard deviation of the previous scores by launching the visual tracker at different time steps, or, at a same time step, in different positions around the GT. The visual tracker is stable if the visual tracking scores are only marginally influenced by the initialization conditions.

There exist of course other metrics. We just mentioned here the most common ones for mono-object visual tracking. One can refer to [199] for a more complete overview on the evaluation of object visual tracking algorithms in general. In particular, the case of multi-object visual tracking requires to take into account the correct data association.

Many visual tracking algorithms have been and are still proposed, that can solve, if not all, at least some of the difficulties listed in Section 3.1, and it has become more and more necessary to provide some means to evaluate and compare them. This is the reason why many datasets have also been made publicly available by researchers, sometimes with ground truth to make easier the comparison between algorithms by enabling both qualitative and quantitative comparative results (using the aforementioned metrics). In the next section, we describe some of the most popular datasets, as well as some datasets only dedicated to test some specificities of video sequences. We also propose an original all-in-one view of these datasets in Tables 1 and 2, in which we list most of the video sequences belonging to these datasets, sorted by alphabetical order, and we mention some of their key properties: sizes, number of frames, video resolutions, year and the challenging visual tracking problems they induce.

4 Datasets for visual tracking

This section is dedicated to a selection of the datasets publicly available on the web, that cover most of the problems in visual tracking (see Section 3 for details). We have chosen to describe these datasets by alphabetical order (note that we refer to some of them by their author's name – a personal choice). All the video sequences they contain are listed in Tables 1 and 2 (in chronological order of their appearance on the web), as well as some information about the number and the size of their frames, their color properties and where to find their ground truth, when available. These tables also mention the main issues for visual tracking raised by these sequences. We have classified these difficulties into nine categories: 1: illumination effects, 2: scene clutter, 3: appearance changes, 4: abrupt motion, 5: occlusions, 6: appearance/disappearance, 7: quality of frames, 8: similar appearance (this also means multiple objects can be tracked), 9: camera motion. For some of these categories, some additional information are provided. Concerning the appearance changes, for instance, we distinguish five cases: "scale" refers to zoom in or out cases, "view" to the changes in the point of view (corresponding to depth rotations), "shape" refers to deformable objects, "look" refers to variations such as glass removing, facial expressions, *etc.*, and "rotation" to image plane rotations. The quality of the frames can be altered by blur, fuzzy or block artifacts due to video compression. Note that video sequences from BEHAVE, CAVIAR and PETS do not appear in these tables because these datasets contain many video sequences: short descriptions of them are given respectively in Sections 4.1, 4.6 and 4.17. We have also reported in these tables minimum tracking errors (in pixels) for sequences that were part of the Tracker Benchmark [235], as well as the algorithms that provided these scores. This indicates which algorithms are the current best ones for these sequences.

4.1 The BEHAVE dataset

The BEHAVE Interactions Test Case Scenarios [164] dataset, proposed in 2009, contains various color video sequences with different scenarios of people having different interactions. Ten specific types of behaviors or interactions were defined, such as *WalkTogether*, *Meet* or *Split*. A complete description of this dataset can be

Table 1 Video sequences dedicated to test visual tracking algorithms (from 2006 to 2013). Name of the sequence and of the dataset it belongs to, year, number of frames, resolution, color or not, ground truth availability, properties and difficulties for visual tracking (1: illumination effects, 2: scene clutter, 3: appearance changes, 4: abrupt motion, 5: occlusions, 6: appearance/disappearance, 7: quality of frames, 8: similar appearance, 9: camera motion) and minimum tracking errors in the Tracker Benchmark [235].

Sequence	Dataset	Year	#	Res.	Color	GT	Difficulties for visual tracking	Min error [235]
Seq_bb	Birchfield [10]	1998	51	128 × 96	Yes	[10]	3 (view, scale), 4	-
Seq_cubicle	Birchfield [10]	1998	51	128 × 96	Yes	[10]	3 (view), 5	-
Seq_dhb	Birchfield [10]	1998	51	128 × 96	Yes	[10]	3 (view)	-
Seq_djb	Birchfield [10]	1998	51	128 × 96	Yes	[10]	3 (view)	-
Seq_dk	Birchfield [10]	1998	51	128 × 96	Yes	[10]	3 (view)	-
Seq_dp	Birchfield [10]	1998	101	128 × 96	Yes	[10]	3 (view, scale), 4, 5	-
Seq_dt	Birchfield [10]	1998	151	128 × 96	Yes	[10]	1, 3 (scale, rotation)	-
Seq_fast	Birchfield [10]	1998	31	128 × 96	Yes	[10]	4	-
Seq_jd	Birchfield [10]	1998	101	128 × 96	Yes	[10]	3 (view, scale), 5	-
Seq_mg	Birchfield [10]	1998	31	128 × 96	Yes	[10]	3 (view, scale, rotation), 4, 5	-
Seq_ms	Birchfield [10]	1998	51	128 × 96	Yes	[10]	3 (view, scale)	-
Seq_nb	Birchfield [10]	1998	501	128 × 96	Yes	[10]	5	-
Seq_sb	Birchfield [10]	1998	501	128 × 96	Yes	[10]	3 (view, scale, rotation), 4, 5	-
Seq_simultaneous	Birchfield [10]	1998	41	128 × 96	Yes	[10]	3 (view), 5	-
Seq_vilains1	Birchfield [10]	1998	201	128 × 96	Yes	[10]	3 (view, scale), 4, 8	-
Seq_vilains2	Birchfield [10]	1998	201	128 × 96	Yes	[10]	3 (view, scale), 8	-
Body	BLUT [217]	2011	233	640 × 480	Yes	-	3 (scale, view, shape), 7 (blur), 9	-
Car 1	BLUT [217]	2011	751	640 × 480	Yes	-	2, 3 (look), 7 (blur), 8, 9	-
Car 2	BLUT [217]	2011	584	640 × 480	Yes	-	2, 3 (scale, view, look), 7 (blur), 8, 9	-
Car 3	BLUT [217]	2011	356	640 × 480	Yes	-	2, 3 (look), 7 (blur), 8, 9	-
Car 4	BLUT [217]	2011	233	640 × 480	Yes	-	2, 3 (look), 7 (blur), 8, 9	1.5 / ASLA [196]
Face	BLUT [217]	2011	493	640 × 480	Yes	-	2, 3 (look, view), 4, 7 (blur), 9	-
Owl	BLUT [217]	2011	631	640 × 480	Yes	-	2, 3 (rotation), 4, 7 (blur), 9	-
Seq A	BoBoT [201]	2010	602	320 × 240	Yes	[201]	4 (motion direction), 9	-
Seq B	BoBoT [201]	2010	629	320 × 240	Yes	[201]	2, 3 (scale) 9	-
Seq C	BoBoT [201]	2010	404	320 × 240	Yes	[201]	3 (scale), 4 (motion direction), 9	-
Seq D	BoBoT [201]	2010	947	320 × 240	Yes	[201]	3 (shape, view), 9	-
Seq E	BoBoT [201]	2010	305	320 × 240	Yes	[201]	5, 9	-
Seq F	BoBoT [201]	2010	453	320 × 240	Yes	[201]	3 (shape), 5, 6, 8, 9	-
Seq G	BoBoT [201]	2010	716	320 × 240	Yes	[201]	3 (view), 9	-
Seq H	BoBoT [201]	2010	412	320 × 240	Yes	[201]	1	-
Seq I	BoBoT [201]	2010	1017	320 × 240	Yes	[201]	3 (shape, view), 5, 6, 8, 9	-
Seq J	BoBoT [201]	2010	388	320 × 240	Yes	[201]	3 (shape), 5, 6, 9	-
Seq K	BoBoT [201]	2010	1020	320 × 240	Yes	[201]	3 (view, scale), 8, 9	-
Seq L	BoBoT [201]	2010	1308	320 × 240	Yes	[201]	3 (view, scale), 9	-
Book 1	Cannons [106]	2010	370	340 × 256	No		2, 3 (rotation)	-
Book 2	Cannons [106]	2010	191	256 × 331	No	-	2, 3 (view)	-
Book 3	Cannons [106]	2010	311	256 × 268	No	-	2, 3 (scale)	-
Illumination	Cannons [106]	2010	60	360 × 240	No	[106]	1, 2	-
Pop Machines	Cannons [106]	2010	37	320 × 240	No	[106]	1, 3 (shape), 5	-
Dinosaur	Cehovin [59]	2011	356	320 × 240	Yes	[59]	2, 3 (view, scale)	-
Gymnastics	Cehovin [59]	2011	206	320 × 180	Yes	[59]	3 (scale, shape, view), 4, 5, 9	-
Hand	Cehovin [59]	2011	244	320 × 240	Yes	[59]	2, 3 (shape, view, scale), 4	-
Hand 2	Cehovin [59]	2011	267	320 × 240	Yes	[59]	2, 3 (shape, view, scale), 4, 7 (blur)	-
Torus	Cehovin [59]	2011	264	320 × 240	Yes	[59]	2, 3 (view, scale)	-
Head Motion	Ellis [47]	2008	2351	320 × 240	No	[47]	3 (view), 5	-
Shaking Camera	Ellis [47]	2008	990	320 × 240	No	[47]	9	-
Track Running	Ellis [47]	2008	503	768 × 576	No	[47]	3 (shape, scale), 5, 8, 9	-
Face	Fragtrack [29]	2006	899	352 × 288	Yes	[129, 152]	5	
Woman	Fragtrack [29]	2006	597	352 × 288	Yes	[129, 152]	3 (view, scale, shape), 5, 9	4.1 / Struck [189]
Cliffdive 1	Godec [96]	2011	76	400 × 226	Yes	[96]	3 (rotation, shape, view), 9	-
Cliffdive 2	Godec [96]	2011	69	400 × 226	Yes	[96]	1, 3 (shape), 4, 6, 9	-
Motorcross 1	Godec [96]	2011	164	640 × 360	Yes	[96]	1, 2, 3 (view, rotation, scale), 4, 5, 9	80.9 / TLD [198]
Motorcross 2	Godec [96]	2011	23	640 × 360	Yes	[96]	1, 2, 3 (rotation), 4, 9	-
Mountain Bike	Godec [96]	2011	228	640 × 360	Yes	[96]	3 (shape, view), 9	6.5 / CSK [191]
Skiing	Godec [96]	2011	81	640 × 360	Yes	[96]	1, 3 (shape, view, rotation, scale), 4, 9	75.2 / TLD [198]
Volley-Ball	Godec [96]	2011	501	720 × 576	No	[96]	2, 3 (shape, view), 5, 7 (compressed), 8	-
Car	Kalal [107]	2011	945	320 × 240	No		3 (view), 5	-
Car Chase	Kalal [107]	2011	9928	290 × 217	Yes	[107]	1, 3 (view), 6, 8, 9	-
Jumping	Kalal [107]	2011	313	352 × 248	No	[107]	2, 3 (view, look), 4, 7 (blur), 9	-
Motocross	Kalal [107]	2011	2665	470 × 310	Yes	[107]	1, 3 (view, shape, scale), 4, 5, 6, 9	-
Panda	Kalal [107]	2011	184	320 × 240	Yes	[107]	2, 3 (scale, view, shape), 5, 6, 7 (compressed	-
Pedestrian 3	Kalal [107]	2011	140	320 × 240	Yes	[107]	2, 3 (view, shape), 5, 8, 9	-

Table 2 Video sequences dedicated to test visual tracking algorithms (from 2006 to 2013). Name of the sequence and of the dataset it belongs to, year, number of frames, resolution, color or not, ground-truth availability, properties and difficulties for visual tracking (1: illumination effects, 2: scene clutter, 3: appearance changes, 4: abrupt motion, 5: occlusions, 6: appearance/disappearance, 7: quality of frames, 8: similar appearance, 9: camera motion) and minimum tracking errors in the Tracker Benchmark [235].

Sequence	Dataset	Year	#	Res.	Color	GT	Difficulties for visual tracking	Min error [235]
Pedestrian 4	Kalal [107]	2011	3000	312 × 233	Yes	[107]	3 (view, shape), 6, 8, 9	-
Pedestrian 5	Kalal [107]	2011	338	320 × 240	Yes	[107]	3 (view, shape), 4, 6, 7 (fuzzy), 9	-
Volkswagen	Kalal [107]	2011	8576	640 × 480	Yes	[107]	2, 3 (view, scale), 4, 5, 6, 7, 8, 9	-
Boxing	Kwon [11]	2008	31	512 × 336	Yes	[107]	3 (shape, view, scale), 4, 5, 8, 9	-
Tennis	Kwon [11]	2008	813	360 × 240	Yes	[11]	3 (shape), 4, 7, 9	-
Diving	Kwon [66]	2009	231	400 × 224	Yes	[11]	2, 3 (rotation, shape), 5, 9	-
Gymnastics	Kwon [66]	2009	767	426 × 234	Yes	[11]	3 (rotation, shape), 9	-
High Jump	Kwon [66]	2009	122	416 × 234	Yes	[11]	3 (rotation, shape), 4, 5, 7 (blur), 9	-
Transformer	Kwon [66]	2009	124	640 × 332	Yes	[11]	3 (shape), 4, 5	-
Animal	Kwon [88]	2010	71	704 × 400	Yes	[11]	2, 3 (shape), 5, 8, 9	-
Basket	Kwon [88]	2010	725	576 × 432	Yes	[11]	2, 3 (shape, view), 5, 6, 8, 9	5.6 / VTD [209]
Football	Kwon [88]	2010	362	624 × 352	Yes	[11]	2, 3 (shape, view), 4, 5, 8, 9	1.9 / DFT [229]
Iron	Kwon [88]	2010	166	720 × 304	Yes	[11]	1, 2, 3 (shape, view), 4, 5, 9	63.2 / VTD [209]
Matrix	Kwon [88]	2010	100	800 × 336	Yes	[11]	1, 2, 3 (shape, scale), 4, 5, 8	48.2 / SCM [242]
Shaking	Kwon [88]	2010	365	624 × 352	Yes	[11]	1, 3 (shape, scale), 5, 8	9.0 / VTD [209]
Singer 1	Kwon [88]	2010	351	624 × 352	Yes	[11]	1, 3 (shape, view, scale), 8, 9	2.7 / SCM [242]
Singer 2	Kwon [88]	2010	366	624 × 352	Yes	[11]	1, 3 (shape, view, scale), 8, 9	21.8 / DFT [229]
Skating 1	Kwon [88]	2010	400	624 × 360	Yes	[11]	1, 3 (shape, view, scale), 4, 8, 9	7.7 / CSK [191]
Skating 2	Kwon [88]	2010	707	640 × 352	Yes	[11]	3 (shape, view, scale), 4, 8, 9	-
Soccer	Kwon [88]	2010	392	624 × 360	Yes	[11]	1, 2, 3 (shape, view, scale), 5, 8, 9	23.2 / VTS [210]
jp1	Litiv [155]	2014	608	320 × 240	Yes	[155]	1, 2 (view), 5, 8	-
jp2	Litiv [155]	2014	229	320 × 240	Yes	[155]	1, 2 (view, scale), 5, 8	-
wbook	Litiv [155]	2014	581	320 × 240	Yes	[155]	1, 2 (view), 5	-
wguest	Litiv [155]	2014	709	320 × 240	Yes	[155]	1, 5	-
Cliffbar	MILtrack [64]	2009	472	320 × 240	No	[64]	2, 3 (rotation, view, scale), 4	-
Coke Can	MILtrack [64]	2009	292	320 × 240	No	[64,235]	1, 3 (view), 4, 5, 6	12.0 / Struck [189]
Coupon Book	MILtrack [64]	2009	326	320 × 240	No	[64]	3, 5, 8	-
Occluded Face 2	MILtrack [64]	2009	820	320 × 240	No	[64,235]	3 (rotation, shape), 5	10.9 Fragtrack[158]
Surfer	MILtrack [64]	2009	842	320 × 240	No	[64]	3 (scale, view, shape), 5, 9	-
Tiger 1	MILtrack [64]	2009	354	320 × 240	No	[64,235]	2, 3 (shape, view, scale), 4, 5	23.1/ VR-V [173]
Tiger 2	MILtrack [64]	2009	365	320 × 240	No	[64,235]	2, 3 (shape, view, scale), 4, 5	12.2 / DFT [229]
Twinnings	MILtrack [64]	2009	472	320 × 240	No	[64]	3 (scale, view)	-
Cartoon	Nejhum [57]	2008	200	320 × 240	No	-	3 (shape), 4	-
Dancer	Nejhum [57]	2008	225	320 × 246	No	-	3 (view, shape), 4	-
Female Skater	Nejhum [57]	2008	160	320 × 246	No	[235]	3 (view, shape), 4, 9	7.7 / CSK [191]
Indian Dancer	Nejhum [57]	2008	150	320 × 262	No	[235]	3 (shape, view)	9.0 / VTD [209]
Male Skater	Nejhum [57]	2008	436	320 × 262	No	-	3 (view, shape), 4, 9	-
Boxing	Oron [120]	2012	368	450 × 360	Yes	[120]	2, 3 (shape, view, scale), 4, 5, 8, 9	-
Dh	Oron [120]	2012	481	800 × 600	Yes	[120]	1, 2, 3 (shape, view, scale), 4, 5, 8, 9	-
Shirt	Oron [120]	2012	1484	320 × 240	Yes	[120]	2, 3 (shape, scale)	-
Board	PROST [83]	2010	698	640 × 480	Yes	[83]	2, 3 (view, scale)	-
Box	PROST [83]	2010	1161	640 × 480	Yes	[83]	2, 3 (view, scale), 4, 5	-
Lemming	PROST [83]	2010	1336	640 × 480	Yes	[83]	2, 3 (view, scale), 4, 5	11.7 / CPF [226]
Liquor	PROST [83]	2010	1741	640 × 480	Yes	[83]	2, 3 (view, scale), 4, 5, 8	8.5 / LOT [223]
Car 11	Ross [44]	2007	393	320 × 240	No	[152,235]	1, 3 (scale)	0.9 / Struck [189]
Car 4	Ross [44]	2007	659	360 × 240	No	[152,235]	1, 3 (view, scale), 8, 9	1.5 ASLA [196]
David Indoor	Ross [44]	2007	761	320 × 240	No	[130,235]	1, 3 (view, scale, look), 9	4.3 / SCM [242]
Dog	Ross [44]	2007	1381	320 × 240	No	[152,235]	3 (scale, view), 6	3.2 / ORIA [237]
Dudek	Ross [44]	2007	573	720 × 480	No	[152,235]	1, 3 (rotation, view, scale), 9	9.6 / IVT [227]
Fish	Ross [44]	2007	476	320 × 240	No	[152,235]	1, 2, 7 (blur)	3.4 / Struck [189]
Sylvester	Ross [44]	2007	1345	320 × 240	No	[152,235]	1, 3 (shape, view, rotation)	5.6 / ORIA [237]
Treillis	Ross [44]	2007	501	320 × 240	Yes	[152,235]	1, 3 (view, shape), 9	4.6 / LSK [218]
Airshow	SPOT [151]	2013	928	424 × 240	Yes	[151]	1, 3 (view), 8	-
Car Chase 2	SPOT [151]	2013	350	640 × 456	Yes	[151]	1, 3 (scale, view), 6, 8, 9	-
Hunting	SPOT [151]	2013	1805	480 × 266	Yes	[151]	3 (shape, view, scale), 4, 5, 9	4.9 / CSK [191]
Parade	SPOT [151]	2013	493	480 × 272	Yes	[151]	2, 3 (shape), 5, 8, 9	-
Red Flowers	SPOT [151]	2013	2249	360 × 240	Yes	[151]	3 (shape), 8	-
Skydiving	SPOT [151]	2013	1237	656 × 480	Yes	[151]	3 (rotation, shape), 8, 9	-
Bird 1	Wang [100]	2011	408	720 × 420	Yes	-	3 (shape, view), 4, 6, 8, 9	-
Bird 2	Wang [100]	2011	99	720 × 420	Yes	-	3 (shape, scale), 5	-
Girl	Wang [100]	2011	1500	640 × 480	Yes	-	2, 3 (shape, view, scale), 5, 6, 8	2.5 / Struck [189]
Bike	Zhang [132]	2012	180	640 × 360	Yes	[132,235]	3 (scale, view, shape)	80.9 / TLD [198]
Bolt	Zhang [132]	2012	293	480 × 270	Yes	[132]	2, 3 (shape, view, scale), 5, 8, 9	6.9 / LSK [218]
Kitesurf	Zhang [132]	2012	84	480 × 270	Yes	[132]	3 (shape, scale, view), 4, 9	-
Panda 2	Zhong [124]	2012	313	640 × 272	Yes	-	2, 3 (rotation), 5	-
Stone	Zhong [124]	2012	593	320 × 240	Yes	-	5	

found in [164]. A very complete ground truth visual tracking information is provided for all the sequences but one. It is composed of the bounding box for each interactive person, as well as labels in case of interactions, in `XML` format. Although this dataset was mainly proposed for behavior analysis of interacting groups, some of its sequences are used to validate visual tracking algorithms that consider occlusions or similar appearance objects [180,179,200], or fast and varying motions of objects [224] for example. This dataset is used by a very large computer vision community working on video tracking, motion analysis and scene understanding.

Fig. 1 The BEHAVE dataset [164]. Two snapshots of sequences, with the ground truth bounding boxes of the objects to track.

Fig. 2 The Birchfield dataset [10]. Four of the seventeen proposed sequences, addressing complex visual tracking problems such as appearance changes, occlusions, similar objects (heads), *etc.*

4.2 The Birchfield dataset

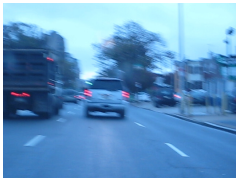
The Birchfield dataset [10] is probably the first dataset that was made publicly available, in 1998, for visual tracking. The goal was to test an elliptical head visual tracker [163] under varying conditions (appearance changes, occlusions) or motions (rotations of objects, zoom of camera, *etc.*). Seventeen color sequences are provided. They show different people (from one to four) moving, occluding, showing their face, but also their profile, or the back of their head. Moreover, due to file compression, the quality of these video sequences is very low, which makes them very challenging. A ground truth text file is also provided, containing bounding ellipses of tracked objects. Lots of works were validated on these sequences. This is for instance the case of the algorithms proposed in [161,160,159,223,216,238,234]. Especially, sequence `Seq_mb` (see Figure 2) is one of the most used.

4.3 The BLUT dataset

The BLUT dataset [217] was proposed by Haibin Ling in 2011 to test a visual tracking algorithm dedicated to blurred properties of the video sequences, called the L_1 tracker [236]. This dataset is composed of seven video sequences containing various objects (cars, faces, human body or poster). They were recorded using a fast moving camera, which causes blur in the frames of the sequences. If the movement of the objects to track is quite simple, the blur considerably decreases the quality of the sequences, which makes the objects harder to track. To our knowledge, only a few works have been tested on this dataset [236,232], even though some previous works have already addressed the blur problem [188] (but, unfortunately, without providing any dataset).

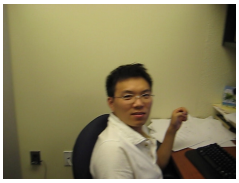
4.4 The BoBoT dataset

The Bonn Benchmark on Tracking (BoBoT) [201] is a 2010 dataset providing a benchmark for testing and comparing different properties of visual tracking algorithms. A total of twelve video sequences are provided in `avi` format. All frames have a size of 320×240 pixels and their numbers vary from 305 to 1308. Ground truth annotations are also given for each sequence. They correspond to the coordinates of the target object's bounding box and its size (in text format files). A java tool evaluator enables to compare different algorithms in terms of tracking errors, with respect to a specific scheme. As shown in Figure 4, each sequence (named from A to L) is dedicated to highlight specific performances of the algorithms (for example, tracking specific motions such as translations, rotations, oscillations, or tracking under occlusion, appearance and disappearance of the object,



`Car 1`
clutter, look, blur
camera motion, similar appearance

`Body`
scale, view, shape
camera motion



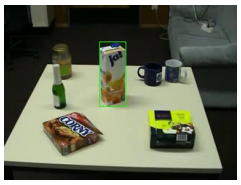
`Face`
clutter, look, view, abrupt motion
clutter, blur, camera motion

`Owl`
clutter, rotation, abrupt
motion blur, camera motion

Fig. 3 The BLUT dataset [217]. Four of the seven video sequences with strong blur properties.

illumination changes, *etc.*). This dataset was used in several recent works, such as [202, 204, 203, 231, 230].

`SeqA`
motion direction,
camera motion



`SeqC`
scale, motion direction,
camera motion

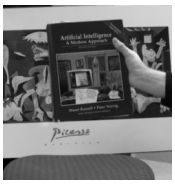
`SeqE`
occlusions, camera motion

`SeqJ`
shape, occlusions, camera motion
appearance/disappearance

Fig. 4 The Bonn Benchmark on Tracking (BoBoT) dataset [201]. Four of the twelve proposed video sequences for testing different cases: types of motions, occlusions, *etc.*

4.5 The Cannons dataset

This dataset [106] was proposed in 2010 by the Vision Laboratory at York University. It contains well-known sequences (`Car 11`, `Woman`, `Sylvester`, *etc.*) from other datasets with complete ground truth annotations, but also five new sequences (see Figure 5), each one having different specificities. In the first three sequences of Figure 5, that are not provided with ground truth, a rigid object (a book) is subject to some specific transformations (scales, rotations, or perspective) in front of a cluttered background. The `Illumination` sequence shows a man walking from left to right with illuminations changes. Finally, in the `Pop Machines` sequence, two people are exchanging an object in a hall: there are occlusions and shape deformations. For these last two sequences, a ground truth annotation is proposed, consisting of the coordinates of the tracked object bounding box. This dataset was essentially used by its providers [168, 167].



`Book 2`
clutter, view

`Book 3`
clutter, scale

`Illumination`
illumination, clutter

`Pop Machines`
illumination, shape, occlusions

Fig. 5 The Cannons dataset [106]. Four sequences, containing different features, such as specific motions, illumination changes or occlusions.

4.6 The CAVIAR dataset

The CAVIAR project (Context Aware Vision using Image-based Active Recognition) [14], from INRIA Labs, started in 2002 and ended in 2005. It was dedicated to the development of algorithms to richly describe and understand video scenes. In CAVIAR Test Case Scenario [15], two sets of data were provided (see Figure 6). A first set was filmed in the entrance lobby of INRIA Labs (indoor), and the second one in a hallway in a shopping center in Lisbon (also indoor). For the first set, six different scenarios were considered (*Walking*, *Browsing*, *Resting, slumping or fainting*, *Leaving bags behind*, *People/groups walking together and splitting up* and *Two people fighting*), and for each scenario, a various number of sequences were recorded (from three to six, for

a total of 28 video sequences). For the second set, two views (corridor view and front view) of twenty-two different scenes were given (hence providing forty-four sequences). Here again, different scenarios were considered (for example *Three persons walking in the corridor* or *Couple leaves a store while browsing*). The ground truth is given for each sequence of these two sets, in `XML` format. It contains a lot of information (depending on the set of data), such as rectangular bounding boxes' locations and sizes, head and feet positions, body direction, *etc.* The CAVIAR dataset is very popular and used by a lot of computer vision research teams. Concerning the visual tracking task, it is challenging since it addresses the problems of occlusions, appearance/disappearance, similar object tracking, appearance changing, and so on. Many works test their algorithms on these datasets, among which some recent visual tracking works [161,223,216,195,243,242,244,175,219].

Fig. 6 The CAVIAR project dataset [14]. From left to right, two frames (ground truth superposed) from sequences of datasets 1 (entrance lobby of INRIA Labs) and 2 (hallway of a shopping center).

4.7 The Cehovin dataset

The Cehovin dataset [59] was provided by the Visual Cognitive Systems Laboratory of the University of Ljubljana, Faculty of Computer and Information Science in 2011. The dataset contains five video sequences whose visual properties such as color, shape and apparent local motion, are changing with time. This dataset mainly addresses the appearance changes of the tracked object (rigid or not). Ground truth is also provided (coordinates of the bounding box). Because of its recent availability, only a few works have used this dataset yet [230,169,170,214,171].

4.8 The Ellis dataset

Liam Ellis proposed this dataset [47] in 2008. It contains three grayscale video sequences in which the object to track changes its appearance with time, due to strong motions of the object and/or the camera. Sequences `Head Motion` and `Camera Shake` (see Figure 8) are not of a high complexity for visual tracking (except for the blur that can appear in case of fast movements). On the opposite, Sequence `Track Running` is very challenging because the point of view of the camera evolves with time, sometimes showing the runner full-face, sometimes 3/4 face, and sometimes it even zooms on him. Moreover, the background is cluttered and some people are also running near the tracked runner. All sequences are provided with ground truth corresponding to the center and size of the bounding boxes in each frame. To our knowledge, this dataset was only used by its providers [177,178].

Fig. 7 The Cehovin dataset [59]. Four of the five video sequences showing objects deforming and changing their appearance.

Fig. 8 The Ellis dataset [47]. Two video sequences in which the appearance of the objects to track vary due to the movements of both the camera and the objects.

4.9 The FRAGtrack dataset

The FRAGtrack dataset [29] is very popular, even if it was provided very early (2006). This dataset contains four sequences with occlusions and appearance changes. One sequence is indoor (`Living room`) while the other three are outdoor (see Figure 9). This dataset is still used by the visual tracking community because the objects it contains are sometimes highly occluded (more that 90% sometimes, see for example sequences `Woman` and `Face`). This makes visual tracking in these sequences still very challenging. Only two sequences (`Woman` and `Face`) are provided with ground truth (a `Matlab` file giving the center of the bounding box every five frames). Note that more complete ground truths have been provided by other researchers: in [129,152] the ground truth contains the coordinates (four corners, or one corner plus width and height) of the bounding box for each frame for the `Woman` sequence. Among the recent works that use this dataset, we can cite [160,179, 159,195,214,233].

`Face`
occlusions

`Woman`
view, scale, shape, occlusions
camera motion

Fig. 9 The FRAGtrack dataset [29]. Two video sequences showing objects that can sometimes be highly occluded and whose shape can change with time.

4.10 The Godec dataset

Godec's dataset [96] was provided in 2011 in the context of the Austrian FFG project MobiTrick, to test the robustness of a visual tracker to high deformations of the object's shape and occlusions. The dataset contains seven video sequences (see Figure 10) with various types of objects (human body, motocross, *etc.*) under heavy non-rigid transformations, occlusions, scale changes and rotations. The ground truth is available for each sequence, as the coordinates of the bounding box (text file). For the moment, only few works are using this dataset [230,185,240].

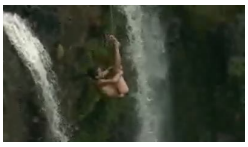
4.11 The Kalal dataset

Zdenek Kalal's dataset [107], proposed in 2011, contains nine video sequences (see some examples in Figure 11). Eight of the nine sequences were taken by cameras with strong motions (most of the time, cameras were embedded in vehicles): for example, a car chase is filmed from a helicopter in Sequence `Car Chase`, and some people are filmed in a parking lot from a flying engine in `Pedestrian 3`, 4 and 5 sequences. The stability of the camera is usually very poor, with jerky movements involved. This makes the global quality of the sequences low. Moreover, sometimes, the object to track can disappear from the scene. These sequences are very hard to work with due to the uncontrolled environment in which the objects are moving, their unpredictable movements, the cluttered backgrounds, and the illumination variations resulting from the outdoor conditions. This is the reason why only few algorithms have been validated on them [232,230,242,198,183,239]. Note that ground truth annotations are provided for each sequence, as a text file containing the corners' coordinates of the bounding boxes of the object to track.

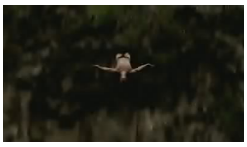
4.12 The Kwon datasets

Junseok Kwon made publicly available three datasets. The first one was proposed in 2008 [11] and contains two video sequences with very abrupt motions, showing tennis players and boxers, with annotated ground truth (center and size of bounding box). The second one was provided in 2009 [66]. It contains four sequences in which the geometric appearance of the object to track drastically changes as well as the background while the camera is moving. The ground truth annotations are given as the center and size of the box surrounding the object. The third dataset, proposed in 2010, can be found on two websites [88,86]. It contains a total of six sequences with severe occlusions, pose variations, illumination changes, and abrupt motions. The ground truth is also given (manual detection of the ellipse surrounding the object to track). Frames from these different datasets are given in Figure 12.

All these video sequences are probably, today, the hardest to work with. For example, sequence `Soccer` shows victorious soccer players filmed by a moving camera. They are sometimes occluded by a red smoke, making them really hard to track. However, a lot of recent works have validated their algorithms on these datasets [216,238,232,195,242,214,240,209,211,241,208, 197,213,212].



`Cliff Dive 1`
rotation, shape, view
camera motion



`Cliff Dive 2`
illumination, shape
abrupt motion, camera motion
appearance/disappearance,

`Motorcross 1`
illumination, clutter, view
rotation, scale, abrupt motion
occlusions, camera motion

`Motorcross 2`
illumination, clutter, rotation
abrupt motion, camera motion

Fig. 10 The Godec dataset [96]. Four of the seven sequences, with objects having non-rigid transformations, scale changes, rotations, as well as occlusions, and illumination changes properties.

`Car Chase`
illumination, camera motion
appearance/disappearance
similar appearance, view

`Motocross`
illumination, scale, shape, occlusions
abrupt motion, camera motion
view, appearance/disappearance

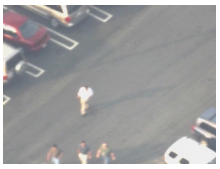
`Panda`
clutter, scale, view
occlusions, shape, quality
appearance/disappearance

`Car`
view, occlusions

`Jumping`
clutter, view, look, blur
abrupt motion
camera motion

`Pedestrian 3`
clutter, view, occlusions
similar appearance
camera motion, shape

`Pedestrian 4`
view, similar appearance
shape, camera motion
appearance/disappearance



`Pedestrian 5`
view, shape, abrupt motion
fuzzy, camera motion
appearance/disappearance

Fig. 11 The Kalal dataset [107]. Eight of the nine proposed video sequences filmed in uncontrolled environments, with moving objects, illumination variations, and appearance changes.

4.13 The Litiv dataset

The Litiv dataset [155] is very recent and is very similar to Birchfield [10], MILtrack [64] or FRAGtrack [29] datasets. Four sequences with people are proposed whose challenge consists in tracking their head in various conditions (occlusions, appearance changes, similar appearance, cluttered backgrounds, *etc.*). Video sequences are provided with a ground truth containing centers, widths and heights of tracked objects bounding boxes. Due to its recent public availability, only its providers have used it [165].

4.14 The MILtrack dataset

The MILtrack [64] dataset is provided in Boris Babenko's website. It is probably the earliest dataset proposed to test visual tracking algorithms that is currently one of the most used by the community: most of the proposed video sequences are still used for comparing recent algorithms. It contains twelve video sequences (series of `png` files) as well as ground truth object location (center of the bounding boxes every five frames) in text format. Complete ground truths for some of the MILtrack dataset sequences have been made available on other websites (`Surfer` [152], `Occluded Face 2` [152,124,121,130]). This dataset is dedicated to test visual tracking under changes of appearance, illumination conditions, occlusions and cluttered background conditions. Most of the works on visual tracking, even the most recent ones, have been validated on some sequences of this dataset [160,161,159,223,195,242,239,189].

4.15 The Nejhum dataset

The Nejhum dataset [57] was proposed in 2008 by Shahed Nejhum to validate his visual tracking algorithm based on block histograms (called BHT tracker [221, 222]). This dataset contains five grayscale sequences

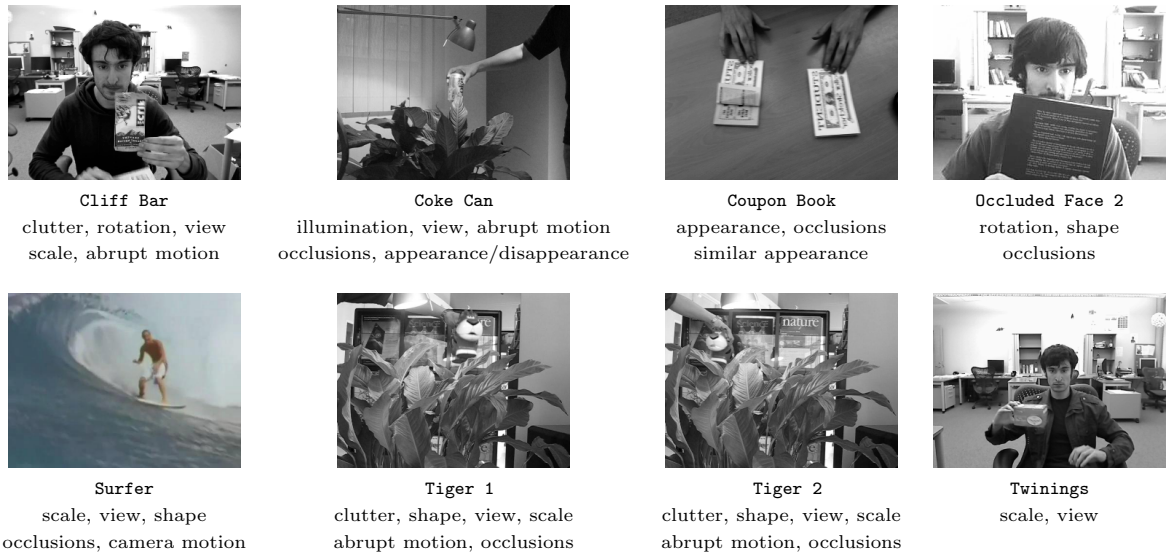


Fig. 14 The MILtrack dataset [64]. Five of the twelve proposed video sequences with occlusions, similar objects, and appearance changes.

in which objects are strongly deformed due to complex movements (people dancing, or skating, plus a cartoon). Motion is very high and is due to both camera and tracked object. These sequences are especially dedicated to algorithms that can adapt online their appearance model (this is compulsory due to the significant shape variations). Unfortunately, no ground truth is provided. These sequences were used to test the algorithms proposed in [179,212,221,222].

4.16 The Oron dataset

The Oron dataset [120] was proposed in 2012 to test the robustness of algorithms w.r.t. rigid and deformable object visual tracking, whose appearance can change with time. It contains three sequences (`Dh`, `Shirt` and `Boxing`, see two of them in Figure 16) with deforming objects whose appearances drastically change over time (see in particular the `Shirt` sequence). All sequences are provided with a ground truth corresponding to the center and size of the object's bounding box in each frame. This dataset was used by different works [223, 213,207].

4.17 The PETS datasets

Since 2000, the International Workshop on Performance Evaluation of Tracking and Surveillance proposes a visual tracking competition whose objectives vary from competition to competition. For example, in 2013 [143] two of the objectives were to track and count people in crowds to estimate their density, and to detect events by crowd analysis. In 2006 [32] a left luggage detection in public area competition was also proposed. For these two competitions, two benchmark data were proposed, corresponding to multi sensor sequences containing different activities, and different scenarios. They were filmed by several synchronized cameras. Some frames of the sequences proposed in 2006 and 2013 are given in Figure 17. Calibration data are also given. Note that, due to the competition's purpose, almost no ground truth is available. However, for some of the sequences, people have provided some, see for instance [121,30]. PETS datasets are very popular among the computer vision community. Some proposed scenarios address hard challenges for the visual tracking community, such as similarity of appearance of the tracked objects, changes in illumination (for the outdoor sequences) or occlusions. Some recent algorithms were tested on some sequences of these datasets [161,223,244,219,174].

4.18 The PROST dataset

The Parallel Robust Online Simple Tracking (PROST) dataset [83] was provided in 2010 by the Institute for computer graphics and vision of Graz University of Technology. The dataset contains four sequences (see Figure 18) either in `wmv` format or provided as a series of `jpg` files. It also contains their ground truth annotation, corresponding to the coordinates of the bounding box of the tracked object (text file). Two `Matlab` scripts enable to read the visual tracking results and to create videos of these results. Moreover, results given by

`Boxing (2008)`
shape, view, scale, abrupt motion
occlusions, camera motion
similar appearance

`Tennis (2008)`
shape, abrupt motion
quality, camera motion

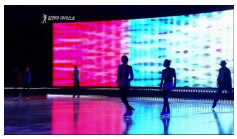
`High Jump (2009)`
rotation, shape
camera motion

`Transformers (2009)`
shape, occlusions
abrupt motion

`Shaking (2010)`
illumination, shape, scale
occlusions, similar appearance

`Animal (2010)`
clutter, shape, occlusions
similar appearance, camera motion

`Soccer (2010)`
illumination, clutter, shape
view, scale, occlusions
similar appearance, camera model



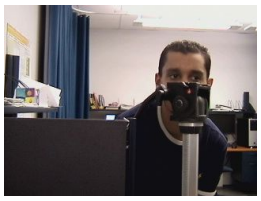
`Skating 1 (2010)`
illumination, shape, view, scale
abrupt motion, camera model
similar appearance

Fig. 12 The Kwon datasets were provided in three steps: in 2008 [11] (abrupt motions – first row), in 2009 [66] (geometrical changes of object's shapes – second row) and in 2010 [88, 86] (abrupt motions, illumination changes, occlusions – last two rows).

`jp1`
clutter, view, occlusions
similar appearance

`jp2`
clutter, view, scale
occlusions, similar appearance

`wbook`
clutter, view, occlusions



`wguest`
clutter, occlusions

Fig. 13 The Litiv dataset [155]. Four sequences with strong occlusions, similar appearances and cluttered backgrounds.

`Female Skater`
view, camera motion
abrupt motion, shape

`Male Skater`
view, camera motion
abrupt motion, shape

`Dancer`
view, shape
abrupt motion

`Cartoon`
shape
abrupt motion

Fig. 15 The Nejhum dataset [57]. Four of the five grayscale sequences, each one containing an object to track whose shape strongly varies with time.

different visual trackers are also provided as text files (PROST [228], Fragtrack [158], MILtrack [160, 159] and GRAD [202]). In these sequences, similar appearances of objects as well as partial occlusions and cluttered background problems are addressed. This dataset was used in different works [230, 214].

4.19 The Ross dataset

The Ross dataset [44] was proposed in 2007 to test the robustness of visual tracking algorithms in situations of strong appearance changes. This dataset contains seven sequences, without any ground truth. They are comparable to those of FRAGtrack and MILtrack, but they contain some specificities, such as highly cluttered background (see Sequence `Car 11`, Figure 19, for example), or illumination changes (Sequence `Fish`, Figure 19, is only dedicated to test illumination changes: neither the object nor the camera are moving). Some other sites also propose complete ground truths for some of these sequences: `David Indoor` (one of the most famous sequence) [130], `Car 5`, `Car 11`, `Dudek`, `Sylvester`, `Dog` and `Treillis` [152]. This dataset is also frequently used for testing visual tracking algorithms [160, 179, 161, 159, 223, 238, 232, 195, 242, 219, 189, 215].

Dh
illumination, clutter, shape, view
abrupt motion, scale, occlusions
camera motion, similar appearance

Shirt
clutter, scale
shape

Fig. 16 The Oron dataset [120]. Two video sequences containing highly deformable objects to track as well as cluttered backgrounds.

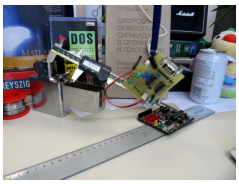
PETS'06 Dataset S2

PETS'06 Dataset S4

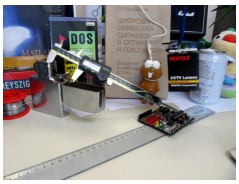
PETS'13 view 001

PETS'13 view 002

Fig. 17 PETS datasets. Two frames from sequences belonging to the benchmarks of PETS 2006 [32] and 2013 [143].



Board
clutter, view, scale



Box
clutter, view, scale, occlusions
abrupt motion

Lemming
clutter, view, scale, occlusions
abrupt motion

Liquor
clutter, view, scale, occlusions
abrupt motion, similar appearance

Fig. 18 The PROST dataset [83]. Four video sequences dedicated to visual tracking in cluttered background conditions.

David Indoor
illumination, view, scale
look, camera motion

Sylvester
illumination, shape, view, rotation

Dudek
illumination, rotation, view, scale
camera motion

Car 11
illumination, scale

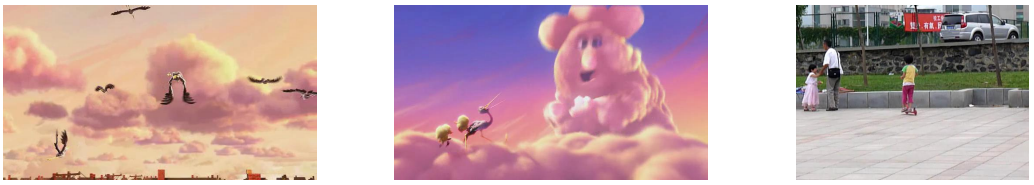
Fig. 19 The Ross dataset [44]. Four of the seven proposed video sequences showing objects to track with strong appearance changes, illumination changes and cluttered backgrounds.

4.20 The SPOT dataset

The SPOT dataset [151] is very recent (2013). It proposes six new and very challenging video sequences that were collected from *Youtube*. The originality of this dataset lies in the fact that it is dedicated to track simultaneously multiple objects, sometimes with similar appearances (see Figure 20). The movements of the objects in a same sequence are related to each other. For example, in the `Red Flowers` sequence, a bunch of tulips are moving due to the wind, and in the `Skydiving` sequence, six people are doing synchronized acrobatics in the sky. The appearance of the tracked objects can also strongly vary with time (see the `Hunting` sequence). For all the sequences, a ground truth annotation file (`Matlab` matrices format) is provided, containing the center and size of the bounding box of each tracked object. For example, for the `Airshow` sequence, four planes have been manually annotated, and three people for the `Parade` sequence. Because of its novelty, only one work uses this dataset and proposes to exploit structural relations between the objects during visual tracking [238, 241].

4.21 The Wang dataset

The Wang dataset [100] was provided in 2011 to test the robustness of algorithms to large changes in scale, motion and shape deformation with occlusions. Four sequences are proposed (see Figure 21): two of them

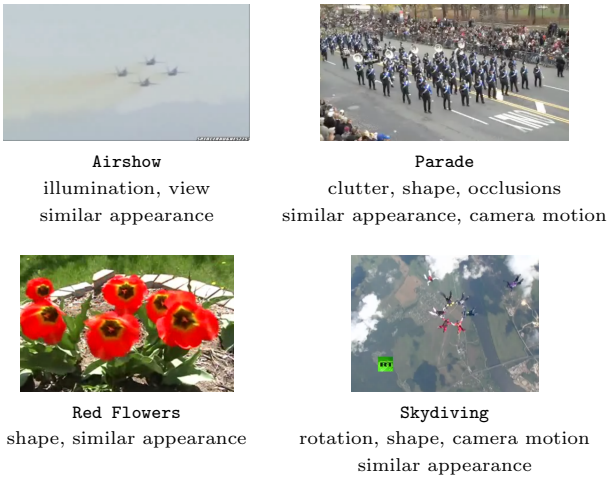


`Bird 1`
shape, view, abrupt motion, camera motion
similar appearance
appearance/disappearance

`Bird 2`
shape, scale, occlusions

`Girl`
clutter, view, scale, camera motion
shape, similar appearance, occlusions
appearance/disappearance

Fig. 21 The Wang dataset [100]. Four sequences with high deformations and occlusions of tracked objects.



`Airshow`
illumination, view
similar appearance

`Parade`
clutter, shape, occlusions
similar appearance, camera motion

`Red Flowers`
shape, similar appearance

`Skydiving`
rotation, shape, camera motion
similar appearance

Fig. 20 The SPOT dataset [151]. Four of the six video sequences with multiple and often similar objects to track.

are animation movies, whereas the other two are real sequences. In all these sequences, the object to track is highly deformed, the camera is moving and, moreover, it is sometimes zooming very fast on the object. The object to track is also occluded, sometimes it can disappear (for example, in the `Bird 1` sequence, the bird is hidden by clouds during some time interval). This dataset was not used in a lot of works [214].

4.22 The Zhang dataset

The Zhang dataset [132] was provided in 2012 with a real-time visual tracker (Real-time Compressive Tracking [240,176]) that can deal with pose variations, blur, variation in illumination and that can tackle the well-known drift problem encountered during visual tracking. In the three outdoor color video sequences, there are out of plane rotations and very abrupt motions (see for example the `Kitesurf` and `Biker` sequences, Figure 22). Sometimes the object to track can disappear during a few frames. In the `Bolt` sequence, we face a lot of occlusions and pose variations, due to both the runner and the camera zooming at the end of the line for example. The centers of the bounding boxes of the objects to track are provided as ground truth (`Matlab` format). Those sequences are very challenging and are used by a new generation of algorithms dedicated to visual tracking under highly realistic and hard conditions [240,239,176,190].

4.23 Zhong dataset

This dataset [124] was provided in 2012 and only contains two sequences, `Panda 2` and `Stone` (see Figure 23) with heavy occlusions, as well as appearances/disappearances of the tracked object. The four corners of the bounding boxes of the tracked objects are given as ground truth annotations. The `Stone` sequence in particular is very interesting to test visual tracking in very cluttered background with similar objects. This dataset was used in several works [195,242,197].

`Panda 2`
clutter, rotation, occlusions

`Stone`
similar appearance

Fig. 23 The Zhong dataset [124]. Two video sequences with heavy occlusions, appearance/disappearance of object and cluttered background.

4.24 Other datasets for visual tracking

We could not cite all the visual tracking datasets. One can however cite the Gauglitz's dataset [77] that is more dedicated to the evaluation of detector-descriptor-based visual camera tracking (see a survey in [184]). This dataset contains 96 video sequences with ground

Bike
scale, view, shape

Bolt
clutter, shape, view, scale, occlusions
similar appearance, camera motion

Kitesurf
shape, scale, view
abrupt motion, camera motion

Fig. 22 The Zhang dataset [132]. Three video sequences with high deformations and motions of the objects to track.

truth annotations. The INRIA-Video dataset [97] was provided for spatio-temporal people segmentation: each person in the scene is separately segmented and annotated. PaFiSS dataset [123] was proposed for video segmentation purposes. It contains 13 video sequences (some of them belong to other datasets) with rigid or deformable objects (cars, people, motorbike, ...) and was used to test a conjoint segmentation-and-tracking algorithm [162]. A ground truth is provided for each sequence containing, for each tracked object, the center and size of its bounding box. We can also cite video sequences that have been involved in the TRECVID [7] workshops since 2001. TRECVID are evaluation meetings focusing on content-based video retrieval and analysis. They provide large test collections. Some of the datasets can be used for visual tracking, but they are not primarily dedicated to this task. Note that we have not considered into this article the datasets from the cell imaging community, but these are also challenging for visual tracking, because they are often very noisy due to the acquisition sensors and, moreover, the tracked objects are similar, deforming and multiple.

4.25 Discussion

We have listed in the preceding subsections numerous visual tracking datasets and given, for each one, its features (which visual tracking issues were raised). We have also described how these datasets are annotated. Among them, only a subset raised just a few issues: illumination effects [106,44], appearance changes [10, 59,29,64,83,44], occlusions [10,29,155,64,83], specific motions [201,106,124], blur [217] or similar appearance of tracked objects [83,151,124]. The other datasets involved a high level of difficulties for visual tracking, ranging from highly deforming objects [96,66,88,57,100, 132], abrupt motions [96,107,11], severe occlusions [88] objects entering or living the field of view [107], to camera motions [47,96,107,120,100]. The higher the number of issues a visual tracker will be able to deal with, the more it will be fit for the new perspectives offered by the current challenges in visual tracking.

Table 3 summarizes the results obtained in the Tracker Benchmark [235], that compared 29 visual tracking algorithms. Here we can find, for each of the visual tracking difficulties, the average, highest and lowest tracking errors (in pixels) obtained in the benchmark. We also report the algorithm that achieved the best score in this benchmark. This table shows that the best current visual tracking algorithms are SCM [242], VTD [209], TLD [198] and STRUCK [189]. Note that the tested video sequences are a subset of those presented in this article. The lowest tracking errors for this subset of sequences are reported in Tables 1 and 2.

With the development of RGB-D sensors, lots of video datasets have been proposed to test visual tracking including depth information. Some of them are cited in Section 5 depending on their objective. Even if this does not exactly refer to visual tracking, including depth information into a visual tracker enables tracking in 3 dimensions as well as analyzing movements with more precision. Another difficulty for visual tracking is when motion comes from both the tracked object and the camera that is acquiring the scene. Some datasets with this feature have already been proposed, for example for traffic surveillance when the camera is positioned in a car windscreen driving on a road [125,118,110]. In the same spirit, developing first-person or egocentric view datasets [145] is required because of a growing field of research due to the miniaturization of cameras and to the development of visual sensors on body (Google glasses, GoPro camera, *etc.*). Aerial images are also challenging for visual tracking tasks because all the objects to track have similar appearances [76,103]. Finally, a visual tracker can also be tested on datasets that were acquired in specific modalities such as infrared [25] or live-cell imaging [144].

Tracking during long time periods is still a challenging problem. It could be very useful for the community to provide new data that include a unique issue (illumination changes, single appearance changes – shape, or scale, or view, or look –, motion directions, clutter, *etc.*)

Table 3 Average, highest and lowest tracking errors (in pixels) and best tracking algorithm obtained in the Tracker Benchmark [235], over the 29 tested trackers, depending on the difficulty in the video sequences.

Visual tracking difficulty	Highest	Average	Lowest	Best algorithm
Illumination effects	114.3	81.7	51.6	SCM [242]
Scene clutter	116.9	79.0	46.8	VTD [209]
Appearance changes	167.4	76.1	42.1	TLD [198]
Abrupt motion	130.6	86.8	43.4	TLD [198]
Occlusions	158.3	77.4	47.7	SCM [242]
Appearance/disappearance	748.13	135.9	56.1	TLD [198]
Quality of frames	368.6	126.8	62.3	TLD [198]
Similar appearance	214.5	90.1	35.6	VTD [209]
Camera motion	334.2	116.2	56.4	Struck [189]

observable during a long time. This would allow testing the capacity of visual trackers not to diverge with time by considering this specific issue.

A major challenge for visual tracking datasets is to provide a correct GT, *i.e.*, one that is constructed with high precision: a manual annotation seems to be, today, the best way to achieve such a precision but this is very tedious. To our opinion, the construction of a GT for visual tracking is still an open problem. Nevertheless, the community has developed its own solutions, that are more often than not sufficient, even if they do not always consider all the visual tracking difficulties (which GT could be well adapted in case of blur, scale changes, appearance/disappearance problems, for example?). Of course, the comparison between the visual tracker's output and the GT is more relevant if both represent the same thing. For example, it can be hard to measure the quality of a visual tracker when its output is a rectangular region surrounding the tracked object whereas the GT's annotation corresponds to a group of pixels (a silhouette, a blob, *etc.*).

As mentioned in the beginning of this section, we consider that a visual tracker's output is a shape surrounding the tracked object and that correctly estimating this shape is the purpose of the visual tracker. However, today, for many applications, this is not the ultimate goal but rather just a mean to achieve a more complex goal, e.g., analyzing the behaviors of individuals in a scene. Therefore, in addition to the aforementioned visual tracking difficulties, these applications raise new challenges and, thus, require other datasets to enable researchers to assess the effectiveness of their algorithms. The next section describes some of these datasets.

5 Datasets for scene analysis and understanding based on visual tracking

The progress in sensor technology enabled the details of perception of the movements in a scene to become increasingly finer. Moreover, the ever decreasing costs of cameras contributed to their proliferation within companies and in individual's homes. There is then a need for automatically processing videos and analyzing and understanding their content. These applications, in which visual tracking is only a component and not an end in itself, raise new difficulties, e.g., multiple-object considerations, multiple crossing trajectories, multi-views, articulated shapes, multiple modalities (RGB-D, IR, panoramic, first eye) and many others. We propose in this section a panorama of some well-known datasets dedicated to them, or more precisely to *dynamic scene analysis and understanding* as we defined it in Section 2.

There exist many other datasets that were made publicly available for dynamic scene analysis and understanding applications (3D multi-view human tracking, articulated object tracking, tracking in the invisible spectrum, crowd analysis, *etc.*) and that can also be used to validate visual tracking algorithms. Thus, this article does not claim to be exhaustive. Fortunately, general dataset repositories can be found on the web [3,61,94,50,49]. Note that some of them do not contain ground truth annotations, which makes quantitative comparisons with other algorithms somewhat difficult.

We divided the datasets described in this section into two categories: the first one (Section 5.1) only concerns human motion analysis and understanding, whereas the second one (Section 5.2) deals with global scene motion understanding (involving humans or not). Tables 4 and 5 list all the datasets.

5.1 Human understanding

Tracking and understanding the behavior of human beings is a very important issue for the computer vision community. It has a lot of applications in the areas of human-computer interaction and video surveillance for example. The main goal of visual tracking in this context is to analyze human gestures, faces, gaits or poses. This enables to extract from video sequences information that are then used to understand humans:

Table 4 Datasets for scene analysis and understanding based on visual tracking.

	Dataset	Reference
Hand gesture	Thomas Moeslund's gesture recognition database	[16]
	Sign language recognition dataset	[22]
	3D iconic gesture dataset	[108]
	Cambridge hand gesture data set	[36]
	Multi-modal gesture recognition dataset	[147]
	Pointing gestures: video sequence database	[21]
	Two-handed datasets	[18]
	Kinect gesture data set	[116]
	MSR action recognition datasets	[79]
	Multi-modal Gesture Recognition Challenge 2013	[181]
	ChAirGest'13 dataset	[138]
Face	The Bircheld dataset	[10]
	Incremental learning for robust visual tracking project website	[44]
	The Honda/UCSD video database	[17]
	RS-DMV dataset	[84]
	CLEAR 2007 evaluation datasets	[39]
	Head pose image database	[19]
	CAVA database	[38]
	YouTube celebrities face tracking and recognition dataset	[131]
	Automatic naming of characters in TV video	[72]
	Hannah dataset	[141]
	The big bang theory dataset	[137]
	SARC3D dataset	[99]
	3D PeS dataset	[91]
	ChokePoint dataset	[92]
Facial expression and emotions	Talking face video	[13]
	Cohn-Kanade AU-coded expression database	[12]
	MMI facial expression database	[78]
	Facial expressions in the wild (SFEW / AFEW)	[95]
	Biwi 3D audiovisual corpus of affective communication - B3D(AC)2	[73]
	Facial expressions and emotion database	[28]
	RS-DMV dataset	[84]
	The UNBC-McMaster shoulder pain expression archive database	[35]
Body motion	CLEAR 2007 evaluation datasets	[39]
	Buffy stickmen dataset	[74]
	ETHZ pascal stickmen dataset	[52]
	ICPR'12 contest	[114]
	HumanEva dataset	[31]
	Ultrech multi-person motion benchmark	[104]
	The CMU motion of body (MoBo) database	[186]
	CMU Graphics Lab motion capture database	[2]
	CASIA gait database	[23]
	DRAG: a database for recognition and analysis of gait	[205]
	The OU-ISIR gait database	[192]
	WVU outdoor SWIR gait dataset	[153]
	DGait database	[111]
Action/activity	Laptev's dataset	[20]
	i3DPost multi-view human action datasets	[43]
	Weisman's dataset	[26]
	CASIA action database	[37]
	KTH multi view football dataset	[119]
	IXMAS actions dataset	[33]
	Hollywood datasets	[62]
	HMDB: A large video database for human motion recognition	[206]
	TREC video retrieval evaluation: TRECVID	[6]
	The LIRIS human activities dataset	[63]
	G3D: A gaming action dataset	[113]
	Hollywood 3D dataset	[187]
	CMU Graphics Lab motion capture database	[2]
	INRIA Xmas motion acquisition sequence	[45]
	MuHAVi: Multicamera human action video data	[80]
	WVU multi-view action recognition dataset	[9]
	Berkeley multimodal human action database (MHAD)	[136]
	UCF-ARG dataset	[103]
	VIRAT dataset	[105]
	University of Rochester activities of daily living dataset	[69]
	YouCook dataset	[154]
	Airport dataset	[133]
	3DLife dataset	[90]
	UCF sports action data set	[55]
	Olympic sports dataset	[82]
	UIUC datasets	[56]
Behavior	Violent scenes dataset	[128]
	Multimodal dyadic behavior dataset (MMDB)	[4]
	BINED dataset: Belfast natural induced emotion datasets	[109]
	SSPNet conflict corpus	[150]
	Canal 9 political debates dataset	[75]

Table 5 Datasets for scene analysis and understanding based on visual tracking.

	Dataset	Reference
Interaction, Social activities	CAVIAR test case scenarios	[15]
	PlacLab dataset	[34]
	The BEHAVE video dataset	[164]
	SDHA dataset	[85]
	AMI meeting corpus	[70]
	TV human interactions dataset	[87]
	TalkingHeads dataset	[102]
	CoffeeBreak dataset	[93]
	IGCLab 6 dataset	[142]
	USAA dataset	[126]
	JPL first-person interaction dataset	[145]
	CMU Graphics Lab motion capture database	[2]
	Collective activity dataset	[60]
	TA2 database	[101]
	WOLF dataset	[89]
Video surveillance	ETISEO dataset	[41]
	ETH Zurik datasets	[112]
	MIT Traffic data set	[65]
	QMUL junction dataset	[149]
	Public dataset of traffic video	[148]
	The Terrascope dataset	[194]
	Airport dataset	[134]
	KIT AIS dataset	[117]
Event	International workshop on performance evaluation of tracking and surveillance website	[32]
	i-Lids bag and vehicle detection challenge	[42]
	CANDELA dataset	[1]
	UCSD anomaly detection dataset	[8]
	BOSS dataset	[48]
	Multiple cameras fall dataset	[81]
	Anomalous behavior data sett	[71]
Crowd	TUD campus and crossing datasets	[54]
	Mall dataset	[146]
	ETH-Person datasets	[51]
	Edinburgh informatics forum pedestrian database	[76]
	VABENE dataset	[127]
	IPSU HUB dataset	[115]
	Crowd segmentation data set	[40]
	Collective motion database	[139]
	OTCBVS benchmark dataset collection	[25]
	Bedre brug af hallen dataset	[135]
	Crowd counting dataset	[140]
	3D PeS dataset	[91]
	UCSD pedestrian database	[67]
	International workshop on performance evaluation of tracking and surveillance website 2007	[46]
	Optical flow dataset	[5]
	UMN dataset	[68]
	Honey bee dance data	[53]
	Multiple ant tracking dataset	[98]
Sport games	APIDIS basket-ball dataset	[58]
	CVBASE'06 dataset	[27]
	Multi-camera and virtual PTZ dataset	[122]
	Hocker players dataset	[24]

what are the actions and the behaviors of humans and how do they interact? We briefly define these topics below, detail the visual tracking difficulties and cite some datasets that are publicly available to test algorithms.

Hand gesture tracking. Hand gestures are one of the main features of body language and are used in many aspects of human communication. For instance, by emphasizing the messages conveyed in speeches, they are an essential companion of the latter. This is also necessary to study the interaction between the hand and some objects of the scene. This explains why hand tracking is a necessary step not only for gesture dynamics analysis, but also for human action, behavior or interaction inferences.

A good hand tracker should be robust to occlusions, abrupt motions and strong shape changes. It should also be able to track two objects similar in appearance (the hands). A taxonomy of hand gestures [225] divides them into two groups according to their purposes: either a communicative goal or a manipulative goal. For the first group, datasets were proposed for applications such as sign language understanding [16, 22] or iconic gesture recognition [108]. For the second group, datasets mainly focus on human-computer interactions or human-object interactions [36,147,21,18]. The most recent datasets have been acquired by RGB-D sensors [147,116,79]. Each year, different challenges are proposed to compare algorithms, and those provide new gesture datasets [181,138].

Face tracking. As for gestures, faces and facial expressions are a natural way for humans to communicate. They can express lots of information, such as inertia or current emotions. Therefore, automatic extraction of face tracks is an important task component for video processing. As for visual tracking, the main challenges concern the tracking of faces in hard conditions (illumination changes, changes in appearance because of the presence of facial expressions for example, partial or total occlusions, *etc.*). Multiple-face tracking can also be needed, and in such cases, the visual tracker should deal with multiple similar objects. Many datasets for face tracking have been proposed to evaluate face trackers' accuracy [10,44,17,84,39] or to estimate head movements or pose [39,19,38]. A goal of face tracking is also to extract the "best" face from a face tracklet to identify or reidentify people, *e.g.*, to recognize a person in the sequence that belongs to a dataset. Therefore, lots of datasets are dedicated to people identification or recognition [39,131,72,141,137], but also to reidentification [99,91]. Recently control access identity needs have emerged (for example in airports or for building entrance security) and a first dataset has recently been made publicly available [92].

Facial expression or emotion. Understanding facial expression means being able to track facial features and to analyze their movements. In this paper, we only deal with video sequences, but there exist many facial expression recognition datasets that only contain fix images. Note that the analysis of facial expressions is closer to motion estimation than to visual tracking. To be adapted to facial expressions, visual trackers have to consider facial constraints, because some facial local deformations have an influence on others. The first datasets dedicated to this problem focused on single feature point tracking [13], or on local face area tracking to decompose the motion into Action Units [12]. Other datasets proposed to analyze facial expressions in terms of level of expressiveness (from neutral to apex) [78], and the most recent datasets propose to recognize facial expressions in more complex situations (often referred to as "in the wild" situations) [95]. During verbal interactions, emotions can be induced from both facial expressions (*i.e.*, the deformation of the face) and sounds. As a consequence, datasets were constructed to recognize emotional states [73,28], some of which were even dedicated to more specific tasks, such as driver monitoring [84] or pain detection [35].

Body motion. The goal of visual tracking is to spatially localize a human in each frame of a video sequence. However, some applications need a finer level of precision to understand the human movements. They are often related to articulated tracking. To estimate human pose or movements, it is needed to consider not only the body as a whole but also the internal movements of the object. This is achieved by using more flexible models than those used traditionally in visual tracking (e.g., models for articulated objects). Of course, by their complex nature, these models increase the complexity of visual tracking. But, in addition, in these kinds of applications arise new problems such as self-occlusions, pose changes, shape deformations, *etc.* The first datasets related to body motion were dedicated to human 2D pose and movement estimation [39,74,52]. The most famous datasets were acquired with multi-cameras, thus allowing 3D pose or movement estimations in color images [39,114]. In such cases, camera calibration should be considered carefully in order to correctly synchronize spatially the different camera views. Moreover, some data are often not visible in some views whereas they are in others, therefore the fusion of all the information coming from all camera is necessary to correctly model the body in $3D$. The HumanEva dataset [31] proposed to use both color and motion capture data for 3D body motion estimation, and a similar protocol was used in [104] for multiple 3D body motion estimation. Finally, there exist applications in which the goal is not only to estimate the body motion but also to analyze it, e.g., walk and gait analysis in clinical analysis. For these, different datasets were proposed in constrained conditions [186,2,23,205,192], acquired by infrared [153] or depth [111] sensors.

Non verbal social signals (body movements, head pose, gestures, facial expressions, *etc.*) are used to understand more complex phenomena in video sequences, such as individual actions and activities, human behavior and small group interactions and social activities. All these themes have recently emerged and lots of datasets have been proposed to test algorithms, that are described in the next three paragraphs.

Individual Action/Activity. The recognition of individual actions or activities has become a major focus of the computer vision community. Lots of datasets have thus been proposed for human action or activity recognition in video sequences (see [172] for a survey). These were more and more difficult to address concerning the visual tracking point of view, notably because actions or activity require long-term visual tracking of the gestures, the body and/or the head. The first datasets proposed on this topic were grayscale video sequences acquired in indoor [20] or outdoor [43] unconstrained environments (homogeneous background, people just moving

their hands in a predefined way, only a small set of specific actions were addressed). Other datasets proposed scenarios for action recognition in colored outdoor environments with simple or cluttered backgrounds [26, 37,119], or with occlusions [33]. More complex situations were considered in [62] where video sequences were taken from two different Hollywood movies. Datasets on action/activity recognition thus seem to fall into one of the following two categories: either they consider only a very small set of actions or activities (such as "drinking" or "smoking"), or they correspond to real life actions/activities in daily environments. The Hollywood movies dataset clearly falls into the second one. In the HMDB51 dataset [206], the number of actions increase considerably (51 actions or activities, such as "handwaving", "drinking", "sword fighting", "diving", "running" or "kicking"), as well as the number of movies from which the dataset was constructed (around 7000 clips). The TRECVID workshop mainly focuses on video retrieval, but it also provides data sequences useful for human activity recognition [6]. Multi-modal datasets for action and activities recognition (grayscale, RGB, depth) are proposed in [63,113,187], or with motion capture data [2]. Multi-views is also considered in [45, 80,9], as well as both multimodal and multi-views in [136]. In all the last four cases, a fusion is necessary to include either multiple views or multiple modalities. Note that each modality has its own specificities: for example some of them are more noisy than others. Some datasets also propose video sequences in aerial views [103,37, 105]. In such cases, we face the problem of similar appearance of the objects. Finally, some datasets propose to recognize specific daily life actions or activities [69], e.g., cooking [154,133], dancing [90] or playing sport [55,82,56].

Behavior. Inferring the behavior of people refers to the analysis and understanding of their actions and activities. For instance, during the *activity* of talking with somebody else, one person can have an aggressive behavior observable by, e.g., sharp movements. Understanding behaviors is essential to analyze social interactions and communications and requires to take into account both actions and reactions of people, as well as the specific context. This is the reason why these datasets are really complicated to address, notably because of the large amount of information that needs be considered. Here again, this is an emergent topic for which different datasets have been made publicly available. These datasets are essentially multimodal, providing video, but also sound or physiologic signal data. Among them, we can cite a dataset dedicated to detect violence or violent behaviors in videos [128], or one dedicated to analyze dyadic behavior [4], or to determine the underlying emotions behind behaviors [109], to detect conflict behaviors [150] or to determine who gains the leadership in political debates [75].

Interactions and social activities. The highest level of analysis of a scene with people concerns the understanding of their interactions or their social activities. Several datasets have been proposed for analyzing people interactions in indoor environments [15,34] or in outdoor environments [164,85]. They concern people interactions during meetings [70], in TV shows [87], during daily life conversations [102,93], while playing games [142] or during social activities [126]. Some datasets were acquired at a first-person viewpoint [145] (in such case, the camera is moving) or as motion capture data [2]. Finally, there also exist studies on group of people interactions [60], while playing games [101] or during meetings [89]. The visual tracking difficulties are multiple: multiple object tracking, occlusions, self-occlusions, similar appearance of objects, cluttered background, and so on.

5.2 Scene understanding

The spread of cameras in public areas has urged the development of algorithms for fully automated surveillance and monitoring systems. Scene understanding includes three questions [182]: (i) what are the actions present in the scene; (ii) how are the objects interacting; and (iii) which rules govern the scene. In the next paragraphs, we separate scene understanding into 4 groups of applications: video surveillance, crowd analysis, sport games description and event detection. Of course, some applications are related to others.

Video surveillance. Video surveillance is one of the most active research areas in computer vision. The goal is to efficiently extract relevant information from a large amount of videos collected by one or multiple cameras in specific areas. This necessitates to detect, track and recognize objects of interest and to understand and analyze their activities. Here again, visual tracking difficulties are multiple: multiple (and sometimes similar) objects to track, occlusion, appearance/disappearance, appearance changes, clutter, *etc.* Designing a visual tracking system that is able to monitor a scene during long periods of time without needing to be reinitialized is still an open problem on which lots of researchers are working. Numerous datasets have been proposed dedicated to video surveillance of indoor or outdoor public areas (airports, streets, halls, offices, ...) [41,112], vehicle traffic surveillance using one camera [65,149] or

multiple cameras [148,194] or plane movement surveillance in airports [134]. Some datasets also propose an aerial view of the scene [117].

Event. In video surveillance, some observed scenarios can give clues on the occurrence of specific events: video-surveillance and events are actually strongly related. In most cases, the event to detect is known in advance and the goal is just to detect it in the video sequence. Event detection scenarios that were taken into account by datasets are, for instance, abandoned object detection [32,42,1], detection of parked vehicles in streets [42], circulation of non pedestrians in walkways [8], suspect behavior detection in subways [48] or fall detection [81]. Some datasets, however, consider events such as anomalous behaviors as breaks with respect to the context [71]. The visual tracking difficulties resulting from such applications are identical to those of video-surveillance. The only difference is that, in cases where the kind of event we are looking for is known, it is possible to inject *prior* knowledge to guide the tracker.

Crowd. A crowd is a large and dense group of people or objects with varying features and goals. Developing computational methodologies for modeling and analyzing movements and behaviors of crowds has become an important task in computer vision. This involves an important problem for visual tracking, namely the similarity and spatial proximity of objects that compose the crowd. A first objective of several datasets concerns the analysis of the individual trajectories of many people when the scene is acquired by a fix camera [54,146], by a mobile one fixed in a car [51] or in a plane (giving aerial images) [76,127,115]. This is directly connected to a multiple object visual tracking problem, that includes an association step between the current track and the previously estimated trajectories. The crowd stability was studied on the dataset in [40], with respect to specific crowd behavior models, and collective motions extraction was studied in the dataset in [139]. A specific dataset has also been acquired with an infrared sensor [25,135]: such video sequences are noisier and this makes the visual tracking even more complicated. Other datasets have been proposed for people counting and crowd density estimation tasks, in a mono-camera context [140] or in a multi-camera one [143,91,67,46]. In [5], collective movements are addressed, in particular to compare normal and emergency behaviors in crowds. Finally, it is also important to analyze crowds to detect abnormal crowd behaviors [68]. Some datasets consider specific crowds, such as bees [53], to analyze their nuptial dances, or ants [98] to study their interactions and ways of communication. Note that for the two last kinds of datasets, the scenes are less structured than for other crowds.

Sport games. Other datasets, that, we think, should be cited in this section concern sport video sequences. These datasets are interesting because players have similar appearance, their trajectories are crossing, and for some sports, the motion can be strong and totally erratic. In general, the videos are acquired by multiple cameras. The key applications are the analysis of tactic games or the automatic point counting. Some specific areas of the scene are often focused on and, sometimes, tracking is performed following some specific motion models. The most famous datasets on this topic concern basket ball [58,27], handball [27,122], squash [27] and hockey [24].

6 Conclusion

The visual tracking community has grown a lot in the past ten years and it is necessary for people to compare their algorithms by testing them on a common base of video sequences. We have presented in this article some of the most used datasets, or the most original ones that have been made publicly available by researchers. These datasets have been classified w.r.t. the difficulties they induce for visual tracking. The goal was to help people choosing the video sequences that are the most appropriate to show the performances of their algorithms, as well as to choose the appropriate measures to evaluate and compare these algorithms to others. We also proposed an overview of the datasets designed to challenge algorithms on open research questions, notably (but not only) on visual tracking. We hope this will help guiding researchers toward the most appropriate applications for which their tracker is fit. Note that we mainly focused on datasets containing videos acquired using optical cameras. Of course, there also exist problems in which video sequences are acquired using other sensors. For instance, in fluorescent imaging, cells are tracked in videos obtained from microscopes, or in ultrasound imaging, stones can be tracked inside the human body using videos obtained from ultrasound devices. Unfortunately, very few such datasets exist: most of the time, researchers use their own videos and, due to ethic problems, are not authorized to make them available to the community.

For future works, it would be most helpful to propose a global benchmark dedicated to the comparison of all the algorithms that were proposed for visual tracking, an idea similar to [235]. A first important and difficult step should be to propose a universal measure of the tracking quality, that would take into account

all the difficulties (changing in appearance, variation of illuminations, occlusions, blur, *etc.*) involved in the visual tracking. The benchmark should then be decomposed into subsets of sequences, each one addressing a specific difficulty. A graduation of the difficulty in these sequences could also be proposed. The lack of such a benchmark has now become an important problem for the visual tracking community because it precludes comparing new tracking algorithms against previous ones in a principled way, in terms, for instance, of the capacity of an algorithm to address each aforementioned difficulty separately, or in a combined way.

References

1. CANDELA dataset. http://www.multitel.be/image/research-development/research-projects/candela/abandon-scenario.php
2. CMU Graphics Lab motion capture database. http://mocap.cs.cmu.edu
3. Computer vision datasets. http://clickdamage.com/sourcecode/cv_datasets.php
4. Multimodal dyadic behavior dataset (MMDB). http://www.cbi.gatech.edu/mmdb/dataset.php
5. Optical flow dataset. http://groups.inf.ed.ac.uk/vision/BEHAVEDATA/CROWDS/index.html
6. TREC video retrieval evaluation: TRECVID. http://www-nlpir.nist.gov/projects/trecvid/
7. TRECVID homepage. http://www-nlpir.nist.gov/projects/trecvid/
8. UCSD anomaly detection dataset. http://www.svcl.ucsd.edu/projects/anomaly/dataset.htm
9. WVU multi-view action recognition dataset. http://csee.wvu.edu/~vkkulathumani/wvu-action.html
10. The birchfield dataset. http://www.ces.clemson.edu/~stb/research/headtracker/seq/ (1998)
11. Tracking of abrupt motion project website. http://cv.snu.ac.kr/research/~wlmctracker/index.html (1998)
12. Cohn-Kanade AU-coded expression database. http://www.pitt.edu/~emotion/ck-spread.htm (2000)
13. Talking face video. http://www-prima.inrialpes.fr/FGnet/data/01-TalkingFace/talking_face.html (2000)
14. CAVIAR: Context Aware Vision using Image-based Active Recognition. http://homepages.inf.ed.ac.uk/rbf/CAVIAR/ (2002)
15. CAVIAR test case scenarios. http://groups.inf.ed.ac.uk/vision/CAVIAR/CAVIARDATA1/ (2002)
16. Thomas Moeslund's gesture recognition database. http://www-prima.inrialpes.fr/FGnet/data/12-MoeslundGesture/database.html (2002)
17. The Honda/UCSD video database. http://vision.ucsd.edu/~leekc/HondaUCSDVideoDatabase/HondaUCSD.html (2003)
18. Two-handed datasets. http://www-prima.inrialpes.fr/FGnet/data/04-TwoHand/main.html (2003)
19. Head pose image database. http://www-prima.inrialpes.fr/perso/Gourier/Faces/HPDatabase.html (2004)
20. Laptev's dataset. http://www.nada.kth.se/cvap/actions/ (2004)
21. Pointing gestures: video sequence database. http://www-prima.inrialpes.fr/FGnet/data/13-MoeslundHead/Pointing04/ (2004)
22. Sign language recognition dataset. http://www-i6.informatik.rwth-aachen.de/~dreuw/database.php (2004)
23. CASIA gait database. http://www.cbsr.ia.ac.cn/english/Gait%20Databases.asp (2005)
24. Hocker players dataset. http://www.cs.ubc.ca/~okumak/research.html (2005)
25. OTCBVS benchmark dataset collection. http://www.vcipl.okstate.edu/otcbvs/bench/ (2005)
26. Weisman's dataset. http://www.wisdom.weizmann.ac.il/~vision/SpaceTimeActions.html (2005)
27. CVBASE'06 dataset. http://vision.fe.uni-lj.si/cvbase06/downloads.html (2006)
28. Facial expressions and emotion database. http://cotesys.mmk.e-technik.tu-muenchen.de/isg/content/feed-database (2006)
29. FRAGtrack dataset. http://www.cs.technion.ac.il/~amita/fragtrack/fragtrack.htm (2006)
30. Haibin ling code and data website. http://www.dabi.temple.edu/~hbling/code_data.htm#L1_Tracker (2006)
31. HumanEva dataset. http://vision.cs.brown.edu/humaneva/ (2006)
32. International workshop on performance evaluation of tracking and surveillance website. http://www.cvg.rdg.ac.uk/PETS2006/index.html (2006)
33. IXMAS actions dataset. http://cvlab.epfl.ch/data/ixmas10 (2006)
34. PlacLab dataset. http://architecture.mit.edu/house_n/data/PlaceLab/PlaceLab.htm (2006)
35. The UNBC-McMaster shoulder pain expression archive database. http://www.pitt.edu/~emotion/um-spread.htm (2006)
36. Cambridge hand gesture data set. CambridgeHandGestureDataset (2007)
37. CASIA action database. http://www.cbsr.ia.ac.cn/english/Action%20Databases%20EN.asp (2007)
38. CAVA database. http://perception.inrialpes.fr/CAVA_Dataset/Site/ (2007)
39. CLEAR 2007 evaluation datasets. http://www.clear-evaluation.org/?The_Evaluation (2007)
40. Crowd segmentation data set. http://crcv.ucf.edu/data/crowd.php (2007)
41. ETISEO dataset. http://www-sop.inria.fr/members/Francois.Bremond/topicsText/etiseoProject.html (2007)
42. i-Lids bag and vehicle detection challenge. http://www.eecs.qmul.ac.uk/~andrea/avss2007_d.html (2007)
43. i3DPost multi-view human action datasets. http://kahlan.eps.surrey.ac.uk/i3dpost_action/ (2007)
44. Incremental learning for robust visual tracking project website. http://www.cs.utoronto.ca/~dross/ivt/ (2007)
45. INRIA Xmas motion acquisition sequence. http://4drepository.inrialpes.fr/public/viewgroup/6 (2007)
46. International workshop on performance evaluation of tracking and surveillance website. http://pets2007.net (2007)
47. Adaptive regression tracking project website. https://www.cvl.isy.liu.se/research/adaptive-regression-tracking/adaptive-regression-tracking/ (2008)

48. BOSS dataset. http://www.multitel.be/image/research-development/research-projects/boss.php (2008)
49. Computer vision central. http://cvisioncentral.com/vision-resources/?frm_keyword=Dataset&Submit=Search (2008)
50. Computer vision online. http://www.computervisiononline.com/datasets (2008)
51. ETH-Person datasets. http://www.vision.ee.ethz.ch/~aess/dataset/ (2008)
52. ETHZ pascal stickmen dataset. http://groups.inf.ed.ac.uk/calvin/ethz_pascal_stickmen/ (2008)
53. Honey bee dance data. http://www.cc.gatech.edu/~borg/ijcv_psslds/ (2008)
54. TUD campus and crossing datasets. https://www.d2.mpi-inf.mpg.de/node/382 (2008)
55. UCF sports action data set. http://crcv.ucf.edu/data/UCF_Sports_Action.php (2008)
56. UIUC datasets. http://vision.cs.uiuc.edu/projects/activity/ (2008)
57. Visual tracking with integral histograms and articulating blocks project website. http://www.cise.ufl.edu/~smshahed/tracking.htm (2008)
58. APIDIS basket-ball dataset. http://www.apidis.org/Dataset/ (2009)
59. Cehovin dataset. http://www.vicos.si/User:Lukacu/Research/Tracking (2009)
60. Collective activity dataset. http://www.eecs.umich.edu/vision/activity-dataset.html (2009)
61. Computer vision online datasets. http://www.cvpapers.com/datasets.html (2009)
62. Hollywood datasets. http://www.di.ens.fr/~laptev/download.html#actionclassification (2009)
63. The LIRIS human activities dataset. http://liris.cnrs.fr/voir/activities-dataset/ (2009)
64. MILtrack dataset. http://vision.ucsd.edu/~bbabenko/miltrack.shtml (2009)
65. MIT traffic data set. http://www.ee.cuhk.edu.hk/~xgwang/MITtraffic.html (2009)
66. Tracking of a non-rigid object project website. http://cv.snu.ac.kr/research/~bhmctracker/index.html (2009)
67. UCSD pedestrian database. http://www.svcl.ucsd.edu/projects/peoplecnt/index.htm (2009)
68. UMN dataset. http://www.vision.eecs.ucf.edu/projects/rmehran/cvpr2009/Abnormal_Crowd.html (2009)
69. University of Rochester activities of daily living dataset. http://www.cs.rochester.edu/~rmessing/uradl/ (2009)
70. AMI meeting corpus. https://www.idiap.ch/dataset/ami/ (2010)
71. Anomalous behavior data sett. http://www.cse.yorku.ca/vision/research/anomalous-behaviour-data/ (2010)
72. Automatic naming of characters in tv video. http://www.robots.ox.ac.uk/~vgg/data/nface/ (2010)
73. Biwi 3D audiovisual corpus of affective communication - $B3D(AC)^2$. http://www.vision.ee.ethz.ch/datasets/b3dac2.en.html (2010)
74. Buffy stickmen dataset. http://www.robots.ox.ac.uk/~vgg/data/stickmen/ (2010)
75. Canal 9 political debates dataset. http://www.idiap.ch/scientific-research/resources/canal-9-political-debates (2010)
76. Edinburgh informatics forum pedestrian database. http://homepages.inf.ed.ac.uk/rbf/FORUMTRACKING/ (2010)
77. Gauglitz dataset. http://ilab.cs.ucsb.edu/tracking_dataset_ijcv/ (2010)
78. MMI facial expression database. http://www.mmifacedb.com (2010)
79. MSR action recognition datasets. https://research.microsoft.com/en-us/um/people/zliu/actionrecorsrc/default.htm (2010)
80. MuHAVi: Multicamera human action video data. http://dipersec.king.ac.uk/MuHAVi-MAS/ (2010)
81. Multiple cameras fall dataset. http://www.iro.umontreal.ca/~labimage/Dataset/ (2010)
82. Olympic sports dataset. http://vision.stanford.edu/Datasets/OlympicSports/ (2010)
83. PROST dataset. http://gpu4vision.icg.tugraz.at/index.php?content=subsites/prost/prost.php (2010)
84. RS-DMV dataset. http://www.robesafe.com/personal/jnuevo/Datasets.html (2010)
85. SDHA dataset. http://cvrc.ece.utexas.edu/SDHA2010/Human_Interaction.html (2010)
86. Tracking by sampling trackers project website. http://cv.snu.ac.kr/research/~vts/ (2010)
87. TV human interactions dataset. http://www.robots.ox.ac.uk/~vgg/data/tv_human_interactions/index.html (2010)
88. Visual tracking decomposition. http://cv.snu.ac.kr/research/~vtd/ (2010)
89. WOLF dataset. https://www.idiap.ch/dataset/wolf (2010)
90. 3DLife dataset. http://perso.telecom-paristech.fr/~essid/3dlife-gc-11/#dataset (2011)
91. 3DPeS dataset. http://www.openvisor.org/3dpes.asp (2011)
92. ChokePoint dataset. http://itee.uq.edu.au/~uqywong6/chokepoint.html (2011)
93. CoffeeBreak dataset. http://profs.sci.univr.it/~cristanm/datasets/CoffeeBreak/index.html (2011)
94. CVonline: Image databases. http://homepages.inf.ed.ac.uk/rbf/CVonline/Imagedbase.htm (2011)
95. Facial expressions in the wild (SFEW / AFEW). http://www.computervisiononline.com/dataset/facial-expressions-wild-sfew-afew (2011)
96. Godec dataset. http://lrs.icg.tugraz.at/research/houghtrack/index.php (2011)
97. INRIA video segmentation dataset. http://www.di.ens.fr/willow/research/videoseg/ (2011)
98. Multiple ant tracking dataset. http://fcl.uncc.edu/nhnguye1/anttrackingmcmc.html (2011)
99. SARC3D dataset. http://www.openvisor.org/sarc3d.asp (2011)
100. Superpixel tracking project website. http://ice.dlut.edu.cn/lu/Project/iccv_spt_webpage/iccv_spt.htm (2011)
101. TA2 database. https://www.idiap.ch/dataset/ta2 (2011)
102. TalkingHeads dataset. http://profs.sci.univr.it/~cristanm/datasets/TalkingHeads/index.html (2011)
103. UCF-ARG data set. http://crcv.ucf.edu/data/UCF-ARG.php (2011)
104. Ultrech multi-person motion benchmark. http://www.projects.science.uu.nl/umpm/data_description.html (2011)
105. VIRAT dataset. http://crcv.ucf.edu/data/VIRAT.php (2011)
106. Visual tracking datasets of York University. http://www.cse.yorku.ca/vision/research/visual-tracking/ (2011)

107. Zdenek Kallal's website. http://personal.ee.surrey.ac.uk/Personal/Z.Kalal/ (2011)
108. 3D iconic gesture dataset (3DIG). http://projects.ict.usc.edu/3dig/ (2012)
109. BINED dataset: Belfast natural induced emotion datasets. http://www.psych.qub.ac.uk/BINED/ (2012)
110. Caltech pedestrian detection benchmark. http://www.vision.caltech.edu/Image_Datasets/CaltechPedestrians/ (2012)
111. DGait database. http://www.cvc.uab.es/DGaitDB/Summary.html (2012)
112. ETH Zurik datasets. http://www.vision.ee.ethz.ch/datasets/ (2012)
113. G3D: A gaming action dataset. http://dipersec.king.ac.uk/G3D/ (2012)
114. ICPR'12 contest. http://www.wide-baseline-camera-network-contest.org/?page_id=35 (2012)
115. IPSU HUB dataset. http://vision.cse.psu.edu/data/data.shtml (2012)
116. Kinect gesture data set. http://research.microsoft.com/en-us/um/cambridge/projects/msrc12/ (2012)
117. KIT AIS dataset. http://www.ipf.kit.edu/downloads_data_set_AIS_vehicle_tracking.php (2012)
118. KITTY vision benchmark suite. http://www.cvlibs.net/datasets/kitti/ (2012)
119. KTH multi view football dataset. http://www.csc.kth.se/cvap/cvg/?page=software (2012)
120. Locally orderless tracking project website. http://www.eng.tau.ac.il/~oron/LOT/LOT.html (2012)
121. Locally orderless tracking project website. http://www.eng.tau.ac.il/~oron/LOT/LOT.html (2012)
122. Multi-camera and virtual PTZ dataset. http://lrs.icg.tugraz.at/download.php#vptz (2012)
123. PaFiSS dataset. http://campar.in.tum.de/Chair/PaFiSS (2012)
124. Robust object tracking via sparsity-based collaborative model project website. http://faculty.ucmerced.edu/mhyang/project/cvpr12_scm.htm (2012)
125. TME motorway dataset. http://cmp.felk.cvut.cz/data/motorway/ (2012)
126. USAA dataset. http://www.eecs.qmul.ac.uk/~yf300/USAA/download/ (2012)
127. VABENE dataset. http://www.ipf.kit.edu/downloads_People_Tracking.php (2012)
128. Violent scenes dataset. https://research.technicolor.com/rennes/vsd/ (2012)
129. Visual tracking via adaptive structural local sparse appearance model project. http://faculty.ucmerced.edu/mhyang/project/cvpr12_jia_project.htm (2012)
130. Visual tracking via adaptive structural local sparse appearance model project website. http://faculty.ucmerced.edu/mhyang/project/cvpr12_jia_project.htm (2012)
131. YouTube celebrities face tracking and recognition dataset. http://seqam.rutgers.edu/site/index.phpoption=com_content&view=article&id=64&Itemid=80 (2012)
132. Zhang dataset. http://www4.comp.polyu.edu.hk/~cslzhang/CT/CT.htm (2012)
133. 50 salads dataset. http://cvip.computing.dundee.ac.uk/datasets/foodpreparation/50salads/ (2013)
134. Airport dataset. http://www.vision.ee.ethz.ch/~dragonr/943/ (2013)
135. Bedre brug af hallen dataset. http://www.create.aau.dk/bbh/dataset/ (2013)
136. Berkeley multimodal human action database (MHAD). http://tele-immersion.citris-uc.org/berkeley_mhad (2013)
137. The big bang theory dataset. https://cvhci.anthropomatik.kit.edu/~mtapaswi/projects/personid.html (2013)
138. ChAirGest'13 dataset. https://project.eia-fr.ch/chairgest/Pages/CorpusInformation.aspx (2013)
139. Collective motion database. http://mmlab.ie.cuhk.edu.hk/archive/project/collectiveness/dataset.htm (2013)
140. Crowd counting dataset. http://crcv.ucf.edu/data/crowd_counting.php (2013)
141. Hannah dataset. https://research.technicolor.com/rennes/hannah-home/ (2013)
142. IGCLab 6 dataset. http://lrs.icg.tugraz.at/download.php#lab6 (2013)
143. International workshop on performance evaluation of tracking and surveillance website. http://pets2013.net (2013)
144. ISBI'13 grand challenge dataset. http://www.codesolorzano.com/celltrackingchallenge/Cell_Tracking_Challenge/Datasets.html (2013)
145. JPL first-person interaction dataset. http://cvrc.ece.utexas.edu/mryoo/jpl-interaction.html (2013)
146. Mall dataset. http://www.eecs.qmul.ac.uk/~ccloy/downloads_mall_dataset.html (2013)
147. Multi-modal gesture recognition. http://www-prima.inrialpes.fr/FGnet/data/03-Pointing/ (2013)
148. Public dataset of traffic video. http://www.tft.lth.se/video/co_operation/data_exchange/ (2013)
149. QMUL junction dataset. http://www.eecs.qmul.ac.uk/~ccloy/downloads_qmul_junction.html (2013)
150. SSPNet conflict corpus. http://www.dcs.gla.ac.uk/vincia/?p=270 (2013)
151. Structure preserving object tracker project website. http://visionlab.tudelft.nl/spot (2013)
152. Vojir tracking dataset repository. http://cmp.felk.cvut.cz/~vojirtom/dataset/index.html (2013)
153. WVU outdoor SWIR gait dataset. http://community.wvu.edu/~bmd024/WOSG/WOSG.html (2013)
154. YouCook dataset. http://www.cse.buffalo.edu/~jcorso/r/youcook/ (2013)
155. Litiv dataset. http://www.polymtl.ca/litiv/en/vid/index.php (2014)
156. Visual Tracking: An Experimental Survey. IEEE Transactions on Pattern Analysis and Machine Intelligence **36**(7), 1442–1468 (2014)
157. Multiple object tracking benchmark. http://motchallenge.net (2015)
158. Adam, A., Rivlin, E., Shimshoni, I.: Robust fragments-based tracking using the integral histogram. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 798–805 (2006)
159. Babenko, B., Yang, M.H., Belongie, S.: Visual tracking with online multiple instance learning. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 983–990 (2009)
160. Babenko, B., Yang, M.H., Belongie, S.: Robust object tracking with online multiple instance learning. IEEE Transactions on Pattern Analysis and Machine Intelligence **33**(8), 1619–1632 (2011)
161. Bao, C., Wu, Y., Ling, H., Ji, H.: Real time robust L_1 tracker using accelerated proximal gradient approach. In: International Conference on Computer Vision, pp. 1830–1837 (2012)

162. Belagiannis, V., Schubert, F., Navab, N., Ilic, S.: Segmentation based particle filtering for real-time 2D object tracking. In: European Conference on Computer Vision, pp. 842–855 (2012)
163. Birchfield, S.: Elliptical head tracking using intensity gradients and color histograms. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 232–237 (1998)
164. Blunsden, S., Fisher, R.B.: The BEHAVE video dataset: ground truthed video for multi-person. `http://groups.inf.ed.ac.uk/vision/BEHAVEDATA/INTERACTIONS/` (2009)
165. Bouachir, W., Bilodeau, G.A.: Structure-aware keypoint tracking for partial occlusion handling. In: IEEE Winter conference on Applications of Computer Vision, p. To appear (2014)
166. Cannons, K.: A review of visual tracking. Tech. rep., York University (2008)
167. Cannons, K., Gryn, J., Wildes, R.: Visual tracking using a pixelwise spatiotemporal oriented energy representation. In: European Conference on Computer Vision, pp. 511–524 (2010)
168. Cannons, K., Wildes, R.: Spatiotemporal oriented energy features for visual tracking,. In: Asian Conference on Computer Vision, pp. 532–543 (2007)
169. Cehovin, L., Kristan, M., Leonardis, A.: An adaptive coupled-layer visual model for robust visual tracking. In: International Conference on Computer Vision, pp. 1363–1370 (2011)
170. Cehovin, L., Kristan, M., Leonardis, A.: Robust visual tracking using an adaptive coupled-layer visual model. IEEE Transactions on Pattern Analysis and Machine Intelligence **35**(4), 941–953 (2013)
171. Cehovin, L., Kristan, M., Leonardis, A.: Robust visual tracking using an adaptive coupled-layer visual model. IEEE Transactions on Pattern Analysis and Machine Intelligence **35**(4), 941–953 (2013)
172. Chaquet, J., Carmona, E., Fernández-Caballero, A.: A survey of video datasets for human action and activity recognition. Computer Vision and Image Understanding **117**(6), 1–49 (2013)
173. Collins, R., Liu, Y., Leordeanu, M.: Online selection of discriminative tracking features. IEEE Transactions on Pattern Analysis and Machine Intelligence **27**(10), 1631–1643 (2005)
174. Di Lascio, R., Foggia, P., Percannella, G., Saggese, A., Vento, M.: A real time algorithm for people tracking using contextual reasoning. Computer Vision and Image Understanding pp. 1–42 (2013)
175. Dick, A., Kumar, P.: Adaptive earth movers distance-based Bayesian multi-target tracking. IET Computer Vision **7**(4), 246–257 (2013)
176. Dinh, T.B., Vo, N., Medioni, G.: Context tracker: Exploring supporters and distracters in unconstrained environments. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1177–1184 (2011)
177. Ellis, L., Dowson, N., Matas, J., Bowden, R.: Linear regression and adaptive appearance models for fast simultaneous modelling and tracking. International Journal of Computer Vision **95**(2), 154–179 (2011)
178. Ellis, L., Matas, J., Bowden, R.: Online learning and partitioning of linear displacement predictors for tracking. In: British Machine Vision Conference, pp. 33–43 (2008)
179. Erdem, E., Dubuisson, S., Bloch, I.: Fragment based tracking with adaptive cue integration. Computer Vision and Image Understanding **116**(7), 827–841 (2012)
180. Erdem, E., Dubuisson, S., Bloch, I.: Visual tracking by fusing multiple cues with context-sensitive reliabilities. Pattern Recognition **45**(5), 1948–1959 (2012)
181. Escalera, S., Gonzàlez, J., Baró, X., Reyes, M., L., O., Guyon, I., Athitsos, V., Escalante, H.: Multi-modal gesture recognition challenge 2013: Dataset and results. In: ACM International Conference on Multimodal Interaction, pp. 445–452 (2013)
182. Fu, W., Wang, J., Lu, H., Ma, S.: Dynamic scene understanding by improved sparse topical coding. Pattern Recognition **46**(7), 1841–1850 (2013)
183. Gao, M.L., Luo, D.S., Teng, Q.Z., He, X.H., Jiang, J.: Object tracking using firefly algorithm. IET Computer Vision **7**(4), 227–237 (2013)
184. Gauglitz, S., Hollener, T., Turk, M.: Evaluation of interest point detectors and feature descriptors for visual tracking. International Journal of Computer Vision **94**(3), 335–360 (2011)
185. Godec, M., Roth, P., Bischof, H.: Hough-based tracking of non-rigid objects. In: International Conference on Computer Vision, pp. 81–88 (2011)
186. Gross, R., Shi, J.: The CMU motion of body (MoBo) database. Tech. Rep. CMU-RI-TR-01-18, Robotics Institute, Pittsburgh, PA (2001)
187. Hadfield, S., Bowden, R.: Hollywood 3D: Recognizing actions in 3D natural scenes. In: IEEE conference on Computer Vision and Pattern Recognition, pp. 3398 – 3405 (2013)
188. Hailin, J., Favaro, P., Cipolla, R.: Visual tracking in the presence of motion blur. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 18–25 (2005)
189. Hare, S., Saffari, A., Torr, P.: Struck: structured output tracking with kernels. In: International Conference on Computer Vision, pp. 263–270 (2011)
190. He, S., Yang, Q., Lau, R.W., Wang, J., Yang, M.H.: Visual tracking via locality sensitive histograms. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2427–2434 (2013)
191. Henriques, J., Caseiro, R., Martins, P., Batista, J.: High-speed tracking with kernelized correlation filters. IEEE Transactions on Pattern Analysis and Machine Intelligence (To appear)
192. Iwama, H., Okumura, M., Makihara, Y., Yagi, Y.: The ou-isir gait database comprising the large population dataset and performance evaluation of gait recognition. IEEE Transactions on Information Forensics and Security **7**(5), 1511–1521 (2012)
193. Jalal, A.: The state-of-the-art in visual object tracking. Informatica **36**, 227–248 (2012)
194. Jaynes, C., Kale, A., Sanders, N., Grossmann, E.: The Terrascope dataset: scripted multi-camera indoor video surveillance with ground-truth. In: IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, pp. 309–316 (2005)
195. Jia, X., Lu, H., Yang, M.H.: Visual tracking via adaptive structural local sparse appearance model. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–7 (2012)
196. Jia, X., Lu, H., Yang, M.Y.: Visual tracking via adaptive structural local sparse appearance model. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–6 (2012)
197. Jin, B., Jing, Z., Xiao, G., Tang, Y., Zhang, C.: Locally discriminative stable model for visual tracking with clustering and principle component analysis. IET Computer Vision **7**(3), 151–162 (2013)

198. Kalal, Z., Mikolajczyk, K., Matas, J.: Tracking-learning-detection. IEEE Transactions on Pattern Analysis and Machine Intelligence **34**(7), 1409–1422 (2012)
199. Karasulu, B., Korukoglu, S.: Performance evaluation software: moving object detection and tracking in videos. SpringerBriefs in Computer Science. Springer (2013)
200. Karavasilis, V., Nikou, C., Likas, A.: Visual tracking using the Earth Mover's distance between Gaussian mixtures and Kalman filtering. Image and Vision Computing **29**, 295–305 (2011)
201. Klein, D.: BoBot - Bonn benchmark on tracking. `http://www.iai.uni-bonn.de/~kleind/tracking/index.htm` (2010)
202. Klein, D., Cremers, A.: Boosting scalable gradient features for adaptive real-time tracking. In: International Conference on Robotics and Automation, pp. 4411–4416 (2011)
203. Klein, D., Schulz, D., Frintrop, S., Cremers, A.: Adaptive real-time video-tracking for arbitrary objects. In: International Conference on Intelligent Robots and Systems, pp. 772–777 (2010)
204. Konigs, A., Schulz, D.: Fast visual people tracking using a feature-based people detector. In: International Conference on Intelligent Robots and Systems, pp. 3614–3619 (2011)
205. Kuchi, P., Hiremagalur, R., Huang, H., Carhart, M., He, J., Panchanathan, S.: DRAG: a database for recognition and analysis of gait. In: Proceedings SPIE 5242, Internet Multimedia Management Systems, pp. 1–10 (2003)
206. Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., Serre, T.: HMDB: A large video database for human motion recognition. In: International Conference on Computer Vision, pp. 2556–2563 (2011)
207. Kwon, J., Lee, K.: Tracking of abrupt motion using wang-landau monte carlo estimation. In: European Conference on Computer Vision, pp. 387–400 (2008)
208. Kwon, J., Lee, K.: 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, pp. 1208–1215 (2009)
209. Kwon, J., Lee, K.: Visual tracking decomposition. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1269–1276 (2010)
210. Kwon, J., Lee, K.: Tracking by sampling trackers. In: International Conference on Computer Vision, pp. 1195–1202 (2011)
211. Kwon, J., Lee, K.: Tracking by sampling trackers. In: International Conference on Computer Vision, pp. 1195–1202 (2011)
212. Kwon, J., Lee, K.: Highly non-rigid object tracking via patch-based dynamic appearance modeling. IEEE Transactions on Pattern Analysis and Machine Intelligence **35**(10), 2427–2441 (2013)
213. Kwon, J., Lee, K.: Wang-Landau Monte Carlo-based tracking methods for abrupt motions. IEEE Transactions on Pattern Analysis and Machine Intelligence **35**(4), 1011–1024 (2013)
214. Kwon, J., Lee, K.M.: Minimum uncertainty gap for robust visual tracking. IEEE Conference on Computer Vision and Pattern Recognition pp. 1–8 (2013)
215. Li, J., Wang, Y., Wang, Y.: Visual tracking and learning using speeded up robust features. Pattern Recognition Letters **33**(16), 2094–2101 (2013)
216. Li, Z., Wang, W., Wang, Y., Chen, F., Wang, Y.: Visual tracking by proto-objects. Pattern Recognition **46**(8), 2187–2201 (2013)
217. Ling, H.: BLUT dataset. `http://www.dabi.temple.edu/~hbling/code_data.htm#L1_Tracker` (2011)
218. Liu, B., Huang, J., Yang, L., Kulikowsk, C.: Robust tracking using local sparse appearance model and k-selection. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–6 (2011)
219. Lu, X., Yuan, Y., Yan, P.: Robust visual tracking with discriminative sparse learning. Pattern Recognition **46**(7), 1762–1771 (2013)
220. Maggio, E., Cavallaro, A.: Video tracking: theory and practice. SpringerBriefs in Computer Science. Wiley (2013)
221. Nejhum, S.S., J.H., Yang, M.H.: Visual tracking with histograms and articulating blocks. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8 (2008)
222. Nejhum, S.S., J.H., Yang, M.H.: Online visual tracking with histograms and articulating blocks. Computer Vision and Image Understanding **114**(8), 901–914 (2010)
223. Oron, S., Bar-Hillel, A., Levi, D., Avidan, S.: Locally orderless tracking. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1940–1947 (2012)
224. Pantrigo, J., Montemayor, A., Sànchez, A.: Heuristic particle filter: applying abstraction techniques to the design of visual tracking algorithms. Expert Systems **28**(1), 49–69 (2011)
225. Pavlovic, V., Sharma, R., Huang, T.: Visual interpretation of hand gestures for human-computer interaction: A review. IEEE Transactions on Pattern Analysis and Machine Intelligence **16**, 677–695 (1997)
226. Pérez, P., Hue, C., Vermaak, J., Gangnet, M.: Color-based probabilistic tracking. In: European Conference on Computer Vision, pp. 1–6 (2002)
227. Ross, D., Lim, J., Lin, R.S., Yang, M.H.: Incremental learning for robust visual tracking. International Journal of Computer Vision **77**(1-3), 125–141 (2008)
228. Santner, J., Leistner, C., Saffari, A., Pock, T., Bischof, H.: PROST: parallel robust online simple tracking. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 723–730 (2010)
229. Sevilla-Lara, L., Learned-Miller, E.: Distribution fields for tracking. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1910–1917 (2012)
230. Vojir, T., Noskova, J., Matas, J.: Robust scale-adaptive mean-shift for tracking. In: J.K. Kämäräinen, M. Koskela (eds.) Image Analysis, *Lecture Notes in Computer Science*, vol. 7944, pp. 652–663. Springer Berlin Heidelberg (2013)
231. Wahab, M., Abas, F.: Target lock: robust real time adaptive visual tracker. In: International Conference on Digital Image Processing, pp. 1–7 (2012)
232. Wang, D., Lu, H., Yang, M.H.: Least soft-thresold squares tracking. IEEE Conference on Computer Vision and Pattern Recognition pp. 1–8 (2013)
233. Wang, S., Lu, H., Yang, F., Yang, M.H.: Superpixel tracking. In: International Conference on Computer Vision, pp. 1323–1330 (2011)
234. Wang, Y., Tang, X., Cui, Q.: Dynamic appearance model for particle filter based visual tracking. Pattern Recognition **45**(12), 4510–4523 (2012)
235. Wu, Y., Lim, J., Yang, M.H.: Online object tracking: a benchmark. IEEE Conference on Computer Vision and Pattern Recognition pp. 1–8 (2013)
236. Wu, Y., Ling, H., Li, J.Y.F., Mei, X., Cheng, E.: Blurred target tracking by blur-driven tracker. In: International Conference on Computer Vision, pp. 1100–1107 (2011)

237. Wu, Y., Shen, B., Ling, H.: Online robust image alignment via iterative convex optimization. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–6 (2012)
238. Zhang, C.L., Jing, Z.L., Pan, H., Jin, B., Li, Z.X.: Robust visual tracking using discriminative stable regions and K-means clustering. Neurocomputing **111**(C), 131–143 (2013)
239. Zhang, K., Song, H.: Real-time visual tracking via online weighted multiple instance learning. Pattern Recognition **46**(1), 397–411 (2013)
240. Zhang, K., Zhang, L., Yang, M.H.: Real-time compressive tracking. In: European Conference on Computer Vision, pp. 864–877 (2012)
241. Zhang, L., van der Maaten, L.: Structure preserving object tracking. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–6 (2013)
242. Zhong, W., Lu, H., Yang, M.H.: Robust object tracking via sparsity-based collaborative model. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1838–1845 (2012)
243. Zhou, Q.H., Lu, H., Yang, M.H.: Online multiple support instance tracking. In: IEEE International Conference on Automatic Face & Gesture Recognition and Workshops, pp. 545–552 (2011)
244. Zhou, X., Li, Y.F., He, B.: Game-theoretical occlusion handling for multi-target visual tracking. Pattern Recognition **46**(10), 2670–2684 (2013)