Neurocomputing 74 (2011) 3823–3831

Contents lists available at SciVerse ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Recent advances and trends in visual tracking: A review

Hanxuan Yang [a,c,1], Ling Shao [a,b,*], Feng Zheng [a], Liang Wang [d], Zhan Song [a,e]

[a] Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, China
[b] Department of Electronic and Electrical Engineering, The University of Sheffield, UK
[c] Department of Electronic Engineering, South China Agricultural University, China
[d] National Lab of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, China
[e] The Chinese University of Hong Kong, Hong Kong, China

ARTICLE INFO

Article history:
Received 3 April 2011
Received in revised form
25 July 2011
Accepted 28 July 2011
Communicated by M. Wang
Available online 24 August 2011

Keywords:
Visual tracking
Feature descriptor
Online learning
Contextural information
Monte Carlo sampling

ABSTRACT

The goal of this paper is to review the state-of-the-art progress on visual tracking methods, classify them into different categories, as well as identify future trends. Visual tracking is a fundamental task in many computer vision applications and has been well studied in the last decades. Although numerous approaches have been proposed, robust visual tracking remains a huge challenge. Difficulties in visual tracking can arise due to abrupt object motion, appearance pattern change, non-rigid object structures, occlusion and camera motion. In this paper, we first analyze the state-of-the-art feature descriptors which are used to represent the appearance of tracked objects. Then, we categorize the tracking progresses into three groups, provide detailed descriptions of representative methods in each group, and examine their positive and negative aspects. At last, we outline the future trends for visual tracking research.

1. Introduction

Visual tracking is an important task within the field of computer vision. The proliferation of high-end computers, the availability of high quality video cameras, and the increasing need for automated video analysis have generated a great deal of interest in visual tracking algorithms. The state of this art has advanced significantly in the past 30 years [1–8]. Generally speaking, the use of visual tracking is pertinent in the tasks of motion-based recognition, automated surveillance, video indexing, human–computer interaction and vehicle navigation, etc.

1.1. The problems in visual tracking

Visual tracking, in general, is a very challenging problem due to the loss of information caused by the projection of the 3D world on a 2D image, noise in images, cluttered-background, complex object motion, partial or full occlusions, illumination changes as well as real-time processing requirements, etc. In the early years, almost all visual tracking methods assumed that the object motion was smooth and no abrupt appearance change. However, tremendous progress has been made in recent years. Some algorithms can deal with the problems of abrupt appearance change, leaving out from scenes and drifting, etc. To build a robust tracking system, some requirements should be considered.

Robustness: Robustness means that even under complicated conditions, the tracking algorithms should be able to follow the interested object. The tracking difficulties may be cluttered background, partial and full changing illuminations, occlusions or complex object motion.

Adaptivity: Additional to various changes of the environment that an object is located in, the object itself also undergoes changes. This requires a steady adaptation mechanism of the tracking system to the actual object appearance.

Real-time processing: A system that needs to deal with live video streams must have high processing speed. Thus, a fast and optimized implementation as well as the selection of high performance algorithms is required. The processing speed depends on the speed of the observed object, but to achieve a smooth output video impression for human eyes, a frame-rate of at least 15 frames per second has to be established.

1.2. How does visual tracking work?

First, we need a description for the object to be tracked. This can, for example, be a template image of the object, a shape, texture or

* Corresponding author at: Department of Electronic and Electrical Engineering, The University of Sheffield, UK.
E-mail addresses: Hanxuan.Yang@hotmail.com (H. Yang), ling.shao@sheffield.ac.uk (L. Shao), feng.zheng@siat.ac.cn (F. Zheng), wangliang@nlpr.ia.ac.cn (L. Wang), zhan.song@siat.ac.cn (Z. Song).
[1] This work was done when the author is a visiting student with SIAT.

0925-2312/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2011.07.024

color model or something alike. Building such an initial object description is a very critical and hard task, because the quality of the description directly relates to the quality of the tracking process. Additionally, such a description is not always available to the tracking application beforehand and thus, it may need to be built up during runtime.

Second, objects are usually embedded into certain context. Visual context has been successfully studied in object detection tasks as well as the image understanding field. For instance, various parts-whole relations have been exploited by visual detectors. In the detection, only stable, long-term and statistically significant object-context relationships are easily incorporated, e.g., [10,11]. In visual tracking, many temporary, but potentially very strong links exist between the tracked object and the rest of the image. Appropriate integration of such context information into a tracking framework will substantially benefit the research of visual tracking.

Moreover, even having a good object description available a priori or established during runtime, adaptivity to appearance changes is necessary to achieve tracking robustness. These changes can arise from small rotations or geometrical transformations of the object, but also from changing texture. To handle such variations, the object model needs to be adjusted to the new circumstances from time to time. The major problem of building such an adaptive tracking system is the degradation of the appearance model caused by the inaccuracy in the estimation of the foreground and background. Most commonly the foreground and background are divided by a bounding box or a region around the location of the object. No matter how tight the region is, such a partition is too rough because some background regions are treated as a part of the foreground, especially when the location of the object is not precise or the object is occluded. This problem is called the *Drifting Problem* [9].

In a word, most visual tracking methods include image input, appearance feature description, context information integration, decision and modal update, as shown in Fig. 1. For different methods, emphasis is not the same, so their schemes will be different. Due to the great success of Particle Filtering [12], also known as sequential Monte Carlo methods (SMC), visual tracking has been formulated as a problem of Bayesian inference in state space. Compared with the regular exhaustive search-based methods, the main advantage of the use of a particle filter is the reduction of sampling patches during tracking. Another benefit of the particle filter is that the sampling effort can be kept constant, independent to the size of the object to track which is not the case with simply expanding the search region around the object with a

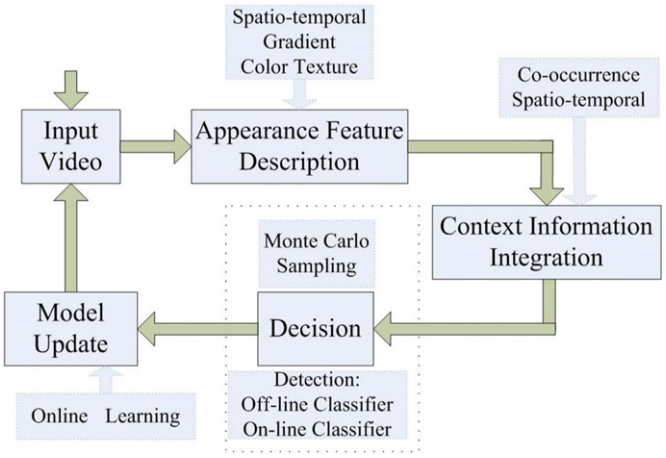

Fig. 1. The flowchart of visual tracking.

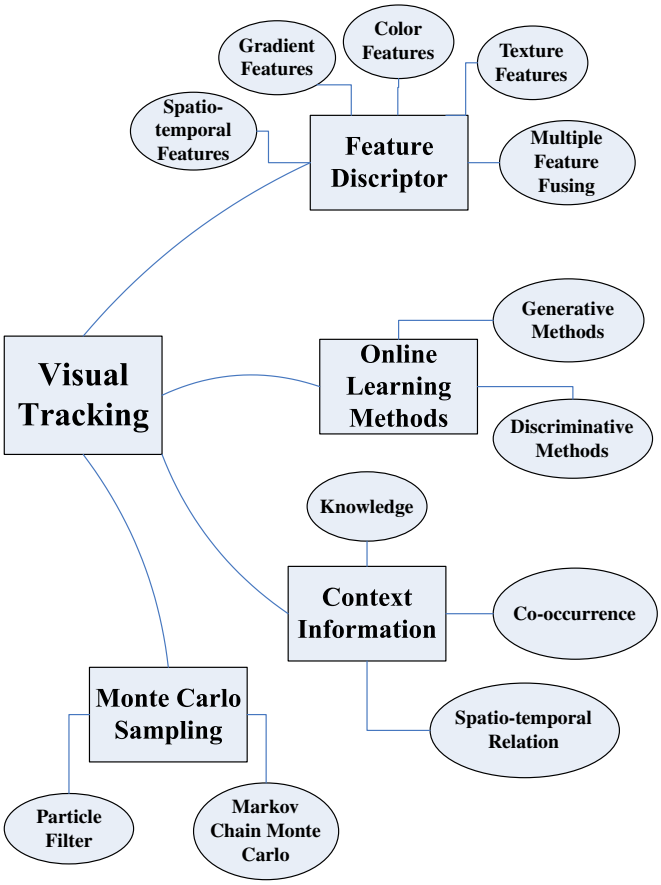

Fig. 2. An overview of the methodology of this paper.

fixed factor. Despite its great success, Particle Filtering often suffered from the sample impoverishment problem [12], which is due to the suboptimal sampling technique. Therefore, introducing more advanced Monte Carlo sampling methods would greatly elevate the visual tracking performance.

1.3. Aim and outline

Yilmaz et al. [13] reviewed the object tracking methods before 2006, presenting detailed analysis and comparisons of various representative methods. Our work aims at introducing recent advances in visual tracking field as well as identifying future trends. The methodology of this paper can be found at Fig. 2. In this paper, we first present representative feature descriptors for visual tracking in Section 2. Then we summarize recent advances in online learning based tracking methods in Section 3. Section 4 is dedicated to discussing the integration of context or knowledge information into visual tracking. In Section 5, we describe the recent progress on Monte Carlo sampling methods. Conclusion and future directions can be found in Section 6.

2. Feature descriptors for visual tracking

Selecting the right features plays a critical role in visual tracking. In general, the most desirable property of a visual feature is its uniqueness so that the objects can be easily distinguished in the feature space. During the last decade, detection of objects of a particular class, such as humans or cars, has absorbed increasing interest in the computer vision field. Visual detection is difficult because the object appearance may vary due

to many factors, including viewpoint, occlusion, illumination, texture, and articulation. This has motivated the invention of different image features that capture different characteristic properties. Some existing methods for object detection base their detectors on a single type of features. This enables a direct comparison of the detection performance of different features. Others try to integrate multiple feature types to improve performance. In fact, any feature descriptor used for visual detection can be adapted for visual tracking. Below, we present recent representative advances in feature descriptors motivated from recent innovations in the visual object detection area.

2.1. Gradient features

Recently, gradient features have been proved advantageous in human detection [14,15]. Numerous techniques dedicated to related research have been proposed. Generally speaking, there are two categories of gradient features.

One main category of gradient based methods is to use shape/contour to represent objects, such as the human body. Gavrila [16] presented a contour based hierarchical chamfer matching detector for pedestrian detection. Lin et al. [17] extended this work by decomposing the global shape models into parts and constructing a hierarchical tree for the part templates. Ferrari et al. [18] used the network of contour segments to represent the shape of an object in order to detect object in cluttered images. Wu and Nevatia [19] proposed edgelet features, which are a type of silhouette oriented features to represent the local silhouette of the human. The human detection problem is then formulated as a maximum a posteriori (MAP) estimation.

Another main category is to use the statistical summarization of the gradients. For example, in [20], Lowe introduced the well-known SIFT descriptor for object recognition. Later, Bay et al. proposed SURF [21], which is a much faster scale and rotation invariant interest point descriptor. Dalal and Triggs [14] used the Histogram of Oriented Gradient (HOG) descriptor in training SVM classifier for pedestrian detection. Zhu et al. [22] improved its computational efficiency significantly by utilizing a boosted cascade of rejectors. Maji et al. [23] also demonstrated promising results using the multi-resolution HOG descriptor and the faster kernel SVM classification. Felzenszwalb et al. [24] described a part based deformable model based on the multi-resolution HOG descriptor for pedestrian tracking. In Gao et al. [25], proposed a novel feature descriptor named Adaptive Contour Feature (ACF) for human detection and segmentation. This feature consists of a chain of a number of granules in Oriented Granular Space (OGS) that is learnt via the AdaBoost algorithm. In Liu et al. [26], proposed the granularity tunable gradients partition (GGP) descriptor for human detection. The concept granularity is used to define the spatial and angular uncertainty of the line segments in the Hough space. Mikolajczyk et al. [27] introduced position-orientation histogram features for human detection. Leibe et al. incorporated the SIFT descriptor into their implicit shape model (ISM) for human detection in [28]. In Gall et al. [29], extended the Hough-transform based class-specific ISM to construct a novel Hough forest detection method.

2.2. Color features

So far, intensity-based descriptors have been widely used for feature representation at salient points. To increase the discriminative power, color descriptors have been proposed, which are robust against certain photometric changes. The apparent color of an object is influenced primarily by two physical factors, (1) the spectral power distribution of the illuminant and (2) the surface reflectance properties of the object. Recent advances in color descriptors can be categorized into novel histogram-based color descriptors and SIFT-based color descriptors.

In the HSV color space, it is known that the hue becomes unstable near the grey axis. Van de Weijer et al. [30] applied an error propagation analysis to the hue transformation. The analysis shows that the certainty of the hue is inversely proportional to the saturation. Therefore, the hue histogram is made more robust by weighing each sample of the hue by its saturation. The H color model is therefore scale-invariant and shift-invariant with respect to light intensity. In Gevers et al. [31], proposed an rg-histogram descriptor, which is based on a normalized RGB color model. Because of the normalization, r and g are scale-invariant and therefore invariant to light intensity changes, shadows and shading.

The SIFT descriptor is not invariant to light color changes, because the intensity channel is a combination of the R, G and B channels. Van de Weijer et al. [30] introduced a concatenation of the hue histogram with the SIFT descriptor, which is scale-invariant and shift-invariant. In [32], color invariants had been first used as an input to the SIFT descriptor, which leads to a CSIFT descriptor that is scale-invariant with respect to light intensity. More detailed performance evaluation of color descriptors can be found in [33].

2.3. Texture features

Texture is a measure of the intensity variation of a surface which quantifies properties such as smoothness and regularity [34–36]. Gabor wavelet [37] is probably the most studied texture feature. The Gabor filters can be considered as orientation and scale tunable edge and line detectors, and the statistics of these micro-features in a given region are often used to characterize the underlying texture information. In recent years, increasing interest is paid on investigating image's local patterns for better detection and recognition. Especially, local patterns that are binarized with an adaptive threshold provide state-of-the-art results on various topics, such as face detection and image classification.

In Ojala et al. [38], developed a very efficient texture descriptor, called Local Binary Patterns (LBP). The LBP texture analysis operator is defined as a grayscale invariant texture measure, derived from a general definition of texture in a local neighborhood. The most important property of the LBP operator is its tolerance against illumination changes. Another equally important characteristic is its computational simplicity. Many variants of LBP have been recently proposed, including Local Ternary Patterns (LTP) [39] and multi-scale block LBP (MB-LBP) [40]. Zhang et al. [41] proposed the local Gabor binary pattern for face representation and recognition. In Mu et al. [42], proposed two variants of LBP: Semantic-LBP and Fourier LBP. These new features can work in perceptually color space and prove more suitable for the human detection task. In inspired by Weber's Law, Chen et al. [43], developed a new local descriptor called the Weber Local Descriptor (WLD). It is based on the fact that human perception of a pattern depends not only on the change of a stimulus (such as sound, lighting) but also on the original intensity of the stimulus.

2.4. Spatio-temporal features

Local space-time features have recently become a popular representation for action recognition and visual detection. Local space-time features capture characteristic salient and motion patterns in video and provide relatively independent representation of events with respect to their spatio-temporal shifts and scales as well as background clutter and multiple motions in the scene. Several methods for feature localization and description have been proposed in the literature and promising results were demonstrated for action classification and various detection tasks [44].

Ke et al. [45] studied the use of volumetric features for event detection in video sequences. They generalized the notion of 2D box features to 3D spatio-temporal volumetric features, which is an extension of the Haar-like features [46]. Liu et al. [47] proposed a contour-motion feature descriptor for robust pedestrian detection. The space-time contours are used as the low level representation of the pedestrian. A 3D distance transform is then applied to extend the one-dimensional contour into three-dimensional space. Willems et al. [48] proposed the Hessian detector which is scale invariant both spatially and temporally as a spatio-temporal extension of the Hessian saliency measure used in [49] for blob detection in images. The detector measures the saliency with the determinant of the 3D Hessian matrix. The HOG/HOF descriptors were introduced by Laptev et al. [50]. To characterize local motion and appearance, the authors computed coarse histograms of oriented gradients (HOG) and optic flow (HOF) accumulated in space-time neighborhoods of detected interest points. Scovanner et al. [51] proposed a 3D SIFT descriptor. They used a bag of words approach to represent videos, and discovered relationship between spatio-temporal words. The HOG3D descriptor was proposed by Kläser et al. [52]. It is based on histograms of 3D gradient orientations. In [53], a transform based spatio-temporal descriptor was proposed for human action recognition. Willems et al. [48] proposed the extended SURF (ESURF) descriptor which extends the image SURF descriptor to videos. Zhao and Pietikäinen [54] proposed the dynamical Local Binary Patterns (DLBP) on three orthogonal planes, and used it for dynamic texture recognition.

2.5. Multiple features fusion

Since the emergence of various feature descriptors, feature fusion has become more and more important for image and video retrieval, visual tracking and detection. The feature fusion scheme typically achieves boosted system performance or robustness, which attracts much attention of researchers from multimedia, computer vision and audio–visual speech processing, etc.

Tuzel et al. [55] utilized the covariance matrix as the descriptor for human representation. Their method can encode the gradients' strength, orientation and position information in symmetric positive definite covariance descriptors which lie on a Riemannian manifold. The main disadvantage of the covariance matrix lies in that the operations through Riemannian geometry are usually time-consuming. Hong et al. [56] proposed a novel descriptor called Sigma Set. Compared with the covariance matrix, Sigma Set is not only more efficient in distance evaluation and average calculation, but also easier to be enriched with first order statistics.

Besides the invention of new multiple features, some works show that using the combination of existing features can also improve the performance. Wu and Nevatia [57] combined the existing heterogeneous features, e.g. edgelet, HOG and covariance matrix, into a boosting framework to improve both the accuracy and the speed. Han et al. [58] used generalized Swendsen–Wang cut to generate the composite of Haar-like features and their results showed that this composition leads to generic improvement for the Haar-like features. Shao and Ji [59] combined MHI and PCOG for human motion classification.

Wang et al. [60] proposed a new HOG-LBP descriptor for pedestrian detection, which can handle partial occlusions. Shotton et al. [61] proposed an efficient fusing of contour and texture cues based on the boosting algorithm for object recognition. Schwartz et al. [62] presented an efficient descriptor for pedestrian detection based on Partial Least Squares (PLS) analysis. Such a descriptor includes the combination of gradient, texture and color information. Alexe et al. [63] presented a generic objectness measure, which combines in a Bayesian framework several image cues, such as color contrast, edge density and multi-scale saliency. Recently, multiple kernel learning method has attracted increasing interest within researchers. Given multiple sources of information, one might calculate multiple basis kernels, one for each source. In such cases, the resultant kernel is often computed as a convex combination of the basis kernels. Kembhavi et al. [64] proposed an Incremental Multiple Kernel Learning (IMKL) approach for object recognition, which combines the Pyramidal Histogram of Oriented Gradients (PHOG) [65] and Geometric Blur [66] together. A feature detector and descriptor evaluation in human action recognition is given in [67].

Table 1
Recent advances on visual descriptors.

Descriptor	Representative	Methods
Gradient Features	HOG, SIFT, ISM	[14–29]
Color Features	CSIFT	[30–33]
Texture Features	LBP, WLD	[34–43]
Spatio-Temporal Features	3DSIFT, DLBP	[44–53]
Multiple Features Fusion	Sigma Set, HOG-LBP	[54–64]
Biological Features	EBIM, ARs	[65,66]

Recently, biological features also received a lot of attention such as Enhanced Biologically Inspired Model (EBIM) [68] and Attentional Regions (ARs) [69]. The biological features tried to mimic human beings' biological vision mechanism in order to achieve robust recognition. Feature description, in general, is a vital component of many visual research fields, such as visual tracking or detection. Table 1 has summarized the recent advances on visual descriptors. From the above categories, we can conclude that tremendous progress has been made in this area. However, no single feature descriptor is robust and efficient enough to deal with all kinds of situations. For instance, the HOG descriptor focuses on edges and structures, ignores flat areas, thus fails to deal with noisy edge regions. A possible drawback of the LBP operator is that the thresholding operation when comparing the neighboring pixels could make it sensitive to noise. Color features represent the global information of images, which are relatively independent of the viewing angle, translation, and rotation of the objects and regions of interest. However, objects with the same color histogram may be completely different in texture, thus color histogram cannot provide enough information. How to combine various kinds of features into a coherent framework needs much more study. Besides, deeper understanding of human vision principles would also enormously benefits feature descriptor research.

3. Online learning based tracking methods

For visual tracking, handling appearance variations of a target object is a fundamental and challenging task. In general, there are two types of appearance variations: intrinsic and extrinsic. Pose variation and/or shape deformation of a target object are considered as the intrinsic appearance variations while the extrinsic variations are due to the changes resulting from different illumination, camera motion, camera viewpoint, and occlusion. These variations can only be handled with adaptive methods which are able to incrementally update their representations. Thus there is an essential need for on-line algorithms that are able to learn continuously. Generally, on-line algorithms can be divided into two categories: *Generative methods* and *Discriminative methods.*

3.1. Generative online learning methods

Generative methods, which are used to learn the appearance of an object, have been exploited to handle the variability of a target. The object model is often updated online to adapt to appearance changes.

Jepson et al. [70] developed an elaborate mixture model with an online EM algorithm to explicitly model the appearance changes during tracking. Zhou et al. [71] embedded appearance adaptive models into a particle filter to achieve a robust visual tracking. Lee and Kriegman [72] presented an online learning algorithm to incrementally learn a generic appearance model from the video. In Ross et al. [73], proposed a generalized visual tracking framework based on the incremental image-as-vector subspace learning method with a sample mean update. It is noted that all the above tracking methods are unable to fully exploit the spatial redundancies within the image ensembles. Consequently, the focus has been made on developing the image-as-matrix learning algorithm for effective subspace analysis. Li et al. [74] employed a three-dimensional temporal tensor subspace learning (ITPCA) for visual tracking. In [75], an incremental learning algorithm is developed for the weighted tensor subspace (WTS) to adapt to the appearance changes during tracking. However, the appearance models adopted in the above mentioned tracking approaches are usually sensitive to the variations in illumination, viewpoint, and pose. This is because they lack a competent object description criterion that captures both statistical and spatial properties of object appearance. Motivated by the incremental Principal Component Analysis algorithm (IPCA) [73], Yang et al. [76] proposed an incremental PCA-HOG descriptor for visual hand tracking. Based on the Covariance Matrix descriptor and the Log-Euclidean Riemannian metric [77], Li et al. [78] presented an online subspace learning algorithm which models the appearance changes by incrementally learning an eigenspace representation for each mode of the target through adaptively updating the sample mean and eigenbasis. Considering the high computational complexity, Wu et al. [79] presented a tracking approach that incrementally learns a low-dimensional covariance tensor representation, efficiently adapting online to appearance changes.

3.2. Discriminative online learning methods

Discriminative methods for classification have also been exploited to handle appearance changes during visual tracking, where a classifier is trained and updated online to distinguish the object from the background. This method is also termed as *tracking-by-detection,* in which a target object identified by the user in the first frame is described by a set of features. A separate set of features describes the background, and a binary classifier separates target from background in successive frames. To handle appearance changes, the classifier is updated incrementally over time. Motion constraints restrict the space of boxes to be searched for the target.

Collins and Liu [80] proposed a method to adaptively select color features that best discriminate the object from the current background. Avidan [81] used an adaptive ensemble of classifiers for visual tracking. Each weak classifier is a linear hyperplane in an 11D feature space composed of R,G,B color and a histogram of gradient orientations. In Wang et al. [82], proposed a tracking algorithm based on online selecting discriminative features from a large feature space with the Fisher discriminant method. In Li et al. [83], proposed a novel tracking method based on incremental 2D-LDA learning and Bayes inference. In Zhang et al. [84], proposed a graph embedding based discriminative learning method, in which the topology structures of graphs are carefully designed to reflect the properties of the sample distributions. Tian et al. [85] presented an online ensemble linear SVM tracker, which makes good usage of history information during tracking. Psychological and cognitive findings indicate that the human perception is attentional and selective. Inspired by this theory, Yang et al. [86] proposed a new visual tracking approach by reflecting some aspects of spatial selective attention, and presents a novel attentional visual tracking (AVT) algorithm. The algorithm dynamically identifies a subset of discriminative attentional regions through a discriminative learning on the historical data on the fly. Recently, Grabner et al. [87,88] designed an online boosting classifier that selects and maintains the best discriminative features from a pool of feature candidates. Later, Saffari et al. [89] proposed the online random forest (RF) algorithm based on an online decision tree growing procedure. Compared with the online boosting method [87,88], the RF method is more robust against label noise. Wang et al. [90] proposed a beyond distance measurement for video annotation. In [91], multi-graph learning was used to unify video annotation.

Despite its high efficiency, online adaption faces one key problem: Each update of the tracker may introduce an error which, finally, can lead to tracking failure (Drifting Problem). In order to deal with this problem, Grabner et al. [92] proposed a semi-supervised approach where labeled examples come from the first frame only, and subsequent training examples are left unlabeled. Although this method is well suited for scenarios where the object leaves the field of view completely, it is difficult to decide the exact object locations in the first frame. Therefore, Babenko et al. [93] proposed a novel tracking method based on the online multiple instance learning method, which resolves the uncertainties of where to take positive updates during tracking. Motivated by the merits of both semi-supervised method [92] and multiple instance learning method [93], Zeisl et al. [94] proposed an online semi-supervised learning algorithm which is able to combine both of these approaches into a coherent framework. This leads to more robust results than applying both approaches separately. More recently, Santner et al. [95] proposed a sophisticated tracking system called PROST that achieves top performance with a smart combination of three trackers: template matching based on normalized cross correlation, mean shift optical flow [96], and online random forests [89] to predict the target location. Other representative methods to deal with the Drifting Problem include [97,98]. In [97], classifiers with different confidence thresholds were applied. In Breitenstein et al. [98], proposed a multi-person tracking-by-detection algorithm with a cascade detection confidence threshold mechanism, which aims at avoiding the errors introduced by the online learning classifier. Considering the causes behind Drifting Problem, the most direct solution to it is to obtain accurate boundaries of the tracked object. In Aeschliman et al. [99], proposed a novel probabilistic framework for jointly solving segmentation and tracking, which achieved significantly improvement in tracking robustness. In Yin et al. [100], proposed a novel method to embed global shape information into local graph links in a Conditional Random Field (CRF) framework. Global shape information is an effective top-down complement to bottom-up figure-ground segmentation as well as a useful constraint to avoid drift during adaptive tracking.

Both the above generative methods and discriminative methods can be integrated into a multi-targets tracking framework. Recently, numerous multi-targets tracking algorithms have been proposed, which focused on resolving the data association problem. In Zhang et al. [101], proposed a network flow based optimization method for data association needed for multiple objects tracking. The optimal data association is formed by a min-cost flow algorithm in the network. Yuan et al. [102] proposed a learning-based hierarchical approach of multi-targets tracking by progressively associating detection responses into the desired target trajectories. Huang et al. [103] presented a detection-based three-level hierarchical association approach to robustly track multiple objects in crowded environments.

Table 2 has summarized the recent advances on online learning tracking methods. A major shortcoming of discriminative methods is their noise sensitivity, while generative methods would easily fail within cluttered background. Recently, [104,105] have tried

Table 2
Recent advances on online learning tracking methods.

Category	Representative	Methods
Generative Method	IPCA, ITPCA	[67–76]
Discriminative Method	OnlineRF, MILBoosting	[77–93]
Combined Method		[99,100]

Table 3
Recent advances on utilizing context information for tracking.

Category	Methods
Co-occurrence	[101,105]
Spatio-temporal relation	[102,106,107]
Knowledge	[103,104]

to combine these two methods and have achieved some progress. In fact, how to combine the generative machine learning methods and discriminative machine learning methods into a coherent framework is a classic question within machine learning field and needs more research. Besides, how to achieve a better balance between adaptivity and stability when using online learning methods is still an open problem.

4. Integration of context information

There is a broad agreement in the community on the valuable role that context plays in any video analysis and image understanding applications. Numerous psychophysics studies have shown the importance of context for human object recognition and detection. Recently, researchers are trying to integrate context information into the visual tracking framework and to achieve great improvements.

In Yang et al. [106], presented a novel tracking algorithm by integrating into the tracking process a set of auxiliary objects that are automatically discovered in the video on the fly by data mining. The collaborative tracking of these auxiliary objects leads to an efficient computation as well as a high robustness. Yuan et al. [107] addressed the problem of recognizing, localizing and tracking multiple objects of different categories in meeting room videos. They incorporated object-level spatio-temporal relationships into the framework. The contextual relationships were modeled by a dynamic Markov random field, in which recognition, localization and tracking were done simultaneously. In Stalder et al. [108], proposed a novel approach to increase the robustness of tracking-by-detection algorithms through a cascaded confidence filter which successively incorporates constraints on the size of the objects, on the preponderance of the background and on the smoothness of trajectories. Roth et al. [109] proposed a novel descriptor called Classifier Grids, which exploited the local context to split the generic detection task into easier sub-problems. Grabner et al. [110] proposed a method to learn supporters which are useful for determining the position of the object of interest, even when the object is not seen directly or when it changes its appearance quickly and significantly. In Kalal et al. [111,112], proposed robust visual tracking algorithms based on the spatio-temporal constraints.

Objects are always embedded into certain context. Table 3 has summarized the recent advances on integrating context information for visual tracking. A recent detailed study of context can be found in [113], which compared several kinds of context information within the object detection area. How to efficiently integrate contextual information into a tracking framework is a promising direction for future visual tracking research.

5. Monte Carlo sampling

Visual tracking usually can be formulated as a graphical model and involves a searching process for inferring the motion of an object from uncertain and ambiguous observations. If the state posterior density is a Gaussian, Kalman Filter [114], Extended Kalman Filter [114] or Unscented Kalman Filter [115] can be used to find the optimal/suboptimal solution. However, most real tracking problems are usually nonlinear and non-Guassian, and thus Particle Filtering [12] is proposed to deal with this situation by Monte Carlo simulation. The key idea of Particle Filtering is to represent the required posterior density function by a set of random samples with associated weights. The Markov Chain Monte Carlo method is well applied to multi-object tracking problems while rigorously formulating the entrance and exit of an object [116,117].

Although Particle Filtering has achieved considerable success in tracking literature, it is faced with a fatal problem due to its suboptimal sampling mechanism in the importance sampling process and thus leads to the well-known sample impoverishment problem. In Zhang et al. [118], proposed an improved unscented particle filter algorithm by the singular value decomposition (SVD) based sigma points calculation method. In [119], particles were generated from a two-stage procedure: at the first stage, simulate the particles with large predictive likelihoods; at the second stage, reweigh the particles and draw the final states. In Zhang et al. [120], proposed a swarm intelligence based particle filter algorithm with a hierarchical importance sampling process which is guided by the swarm intelligence extracted from the particle configuration, and thus greatly overcome the sample impoverishment problem suffered by particle filters. In Kwon et al. [121], proposed a geometric method for visual tracking with a geometrically defined optimal importance function, obtained via Taylor expansion of a principal component analysis based measurement function on the 2D affine group. Schindler et al. [122] represented an object as the constellations of parts to accurately track a bee with the Rao–Blackwellized Particle Filter with fixed topology of the constellation. The cascade particle filter addresses tracking in low frame rate videos [97]. In this approach, the detection algorithm is well combined with particle filter to deal with abrupt motions. It demonstrates efficiency in a face tracking case. In order to deal with abrupt motion, Kwon and Lee [123] proposed a novel tracking algorithm based on the Wang–Landau Monte Carlo (WLMC) sampling method, which can alleviate the motion smoothness constraint utilizing both the likelihood term and the density of states term. The Basin Hopping Monte Carlo (BHMC) sampling method was introduced in [124] to construct a novel tracking algorithm for the target of which geometric appearance changes drastically over time. The BHMC method efficiently reduces the computational complexity and deals with the problem of getting trapped in local minima. A great breakthrough of Monte Carlo sampling methods is [125], which achieved high tracking robustness in extremely complicated scenarios. The algorithm is based on a visual tracking decomposition scheme. Specifically, the observation model is decomposed into multiple basic observation models that are constructed by sparse principal component analysis of a set of feature templates. The motion model is also represented by the combination of multiple basic motion models, each of which covers a different type of motion.

In recent years, numerous innovations have been made in the Monte Carlo sampling based stochastic tracking area. Table 4 has summarized the recent advances on Monte Carlo sampling methods for visual tracking. A detailed description of Monte Carlo sampling methods can be found in [126]. Compared with regular exhaustive search-based methods, the main advantage of Monte Carlo sampling methods is the reduction of sampling patches

Table 4
Recent advances on Monte Carlo sampling methods.

Particle filter	Markov chain Monte Carlo methods
[92,113–117]	[111,112,118–121]

Table 5
Comparison between well-known methods. P.: partial. M.: multi-feature.

System	Domain	Speed	Rotation	P. occlusion	Drifting
TLD [94]	Online Learning	Fast	Not Robust	Not Robust	Robust
CCF [108]	M. fusion	General	Not Robust	Robust	Robust
TI [110]	Context	General	Not Robust	Robust	Robust
VTD [125]	Monte Carlo	Slow	Robust	Robust	Robust

during tracking. Another benefit is that the sampling effort can be kept constant, independent to the size of the object to track which is not the case with simply expanding the search region around the object with a fixed factor. Many state-of-the-art discriminative online learning based tracking methods [85,87–89,93] estimated target's position directly from exhaustive search-based methods. It would be reasonable and inspiring to integrate these methods into a stochastic inference framework.

6. Conclusion and future directions

Visual tracking is a kind of motion analysis at the object level, which consists of two major components: object representation and temporal filtering. In this paper, we survey the recent progress on visual tracking, including feature descriptors, online learning methods, integration of context information as well as Monte Carlo sampling methods. One well-known system in each domain is selected to show its performance, and the comparison between them is given in Table 5.

Although various kinds of feature descriptors have been invented in recent years, such as HOG, LBP and SURF, no single one is robust and fast enough to deal with all tracking situations due to severe appearance or motion changes. With the development of ensemble machine learning methods like Boosting as well as multiple kernel learning methods, a promising direction is combining various complementary features such as contour and texture or multi-modal sensory data like video information and audio information into a coherent framework to capture both statistical and spatial properties. Many research works have been done in this direction, such as the classic cascaded boosted face detector [127] and multiple kernels for object detection [128]. However, the selection of weak classifiers within Boosting methods is still an open problem, poor weak classifiers do not perform better than random guess, thus cannot help decrease the training error during the Boosting process. A promising direction is to discovery compositional features from a given and possibly small feature pool based on data mining techniques. On the other hand, multiple kernel learning methods (MKL) often introduce complexity cost and thus result in quite a heavy algorithm in both training and testing. To overcome it, efficient search techniques such as cascaded method [127] and Efficient Subwindow Search [129] can be integrated with MKL framework. Online learning methods have been extensively studied for the last decade in the visual tracking field. Many state-of-the-art algorithms have been proposed to deal with complex situations during tracking. However, the inherent Drifting Problem of online learning tracking methods still needs more discussions. The existing methods such as semi-supervised Boosting [92], multiple instance learning based Boosting [93] as well as PROST [95], are all based on the "Anchoring Mechanism", which is a general strategy for drift avoidance that can make sure the interested objects do not stray too far from the initial appearance models. Yet, these methods all suffer from the fixed prior appearance model, which either can be too generic to drift to the similar objects or too restrictive to fails in dramatic changes.

A promising solution is to construct an adaptive prior online which can adapt to changes incrementally during tracking. Such an adaptive prior can achieve a good balance between adaptivity and stability. Considering the reason behind the Drifting Problem, new advances on segmentation can also contribute directly to solve the Drifting Problem. The drawback of this direction is that it is quite difficult to obtain precise segmentation within complicated background. Moreover, current generative based online learning methods and discriminative based online learning methods are not robust enough to operate in realistic scenarios. A promising direction is to combine the merits of both generative based online learning methods and discriminative based online learning methods into a coherent framework in order to achieve more robust results than applying both approaches separately. Such a combination has been a classic question within machine learning area. Besides, novel findings from the Monte Carlo sampling community would also greatly benefit the visual tracking research. Smarter Monte Carlo sampling methods can greatly reduce the searching space thus result in reduced computational complexity. Contextual information has been widely studied in image and video understanding. While only recently, contextual information has been exploited effectively in visual tracking. In fact, many psychophysics studies have shown the importance of context for human beings' vision system. With the advances of machine learning methods such as transfer learning [130] and graphical models [131], contextual information will play an increasingly important role in future visual tracking research.

Acknowledgments

The work described in this article was supported partially by the grants from the National Natural Science Foundation of China (NSFC, grant no. 61002040, 60903115), NSFC-GuangDong (grant no. 10171782619-2000007).

References

[1] B. Lucas, T. Kanade, An iterative image registration technique with an application to stereo vision, in: IJCAI, 1981.
[2] J. Shi, C. Tomasi, Good features to track, in: CVPR, 1994.
[3] M. Isard, A. Blake, A smoothing filter for condensation, in: ECCV, 1998.
[4] D. Comaniciu, V. Ramesh, P. Meer, Real-time tracking of non-rigid objects using mean shift, in: CVPR, 2000.
[5] A. Adam, E. Rivlin, I. Shimshoni, Robust fragments-based tracking using the integral histogram, in: CVPR, 2006.
[6] S. Avidan, Ensemble tracking, IEEE Trans. Pattern Anal. Mach. Intell. 29 (2) (2007) 261–271.
[7] Yuan Li, Haizhou Ai, T. Yamashita, Shihong Lao, M. Kawade, Tracking in low frame rate video: a cascade particle filter with discriminative observers of different life spans, IEEE Trans. Pattern Anal. Mach. Intell. 30 (10) (2008) 1728–1740.
[8] M. Ozuysal, M. Calonder, V. Lepetit, P. Fua, Fast keypoint recognition using random ferns, IEEE Trans. Pattern Anal. Mach. Intell. 32 (3) (2010) 448–461.
[9] L. Matthews, T. Ishikawa, S. Baker, The template update problem, IEEE Trans. Pattern Anal. Mach. Intell. 26 (6) (2004) 810–815.
[10] Antonio Torralba, Contextual priming for object detection, Int. J. Comput. Vision 53 (2) (2003) 169–191.
[11] C. Desai, D. Ramanan, C. Fowlkes, Discriminative models for multi-class object layout, in: ICCV, 2009.
[12] S. Maskell, N. Gordon, A Tutorial on Particle Filters for On-line Nonlinear/Non-Gaussian Bayesian Tracking, Target Tracking: Algorithms and Applications IEE, Workshop, 2001.
[13] A. Yilmaz, O. Javed, M. Shah. Object tracking: A survey, in: IPCV, 2006.
[14] N. Dalal, B. Trigges, Histograms of oriented gradients for human detection, in: CVPR, 2005.
[15] P. Sabzmeydani, G. Mori. Detecting pedestrians by learning shapelet features, in: CVPR, 2007.

[16] D. Gavrila, Pedestrian detection from a moving vehicle, in: ECCV, 2000.
[17] Z. Lin, L.S. Davis, D. Doermann, D. DeMenthon, Hierarchical part-template matching for human detection and segmentation, in: ICCV, 2007.
[18] V. Ferrari, T. Tuytelaars, L. Van, Gool, Object detection by contour segment networks, in: ECCV, 2006.
[19] B. Wu, R. Nevatia, Detection of multiple, partially occluded humans in a single image by Bayesian combination of edgelet part detectors, in: ICCV, 2005.
[20] D.G. Lowe, Object recognition from local scale-invariant features, in: ICCV, 1999.
[21] H. Bay, T. Tuytelaars, L. Van Gool, SURF: Speeded up robust features, in: ECCV, 2006.
[22] Q. Zhu, S. Avidan, M. Yeh, K. Cheng, Fast human detection using a cascade of histograms of oriented gradients, in: CVPR, 2006.
[23] S. Maji, A. Berg, J. Malik, Classification using intersection kernel support vector machines is efficient, in: CVPR, 2008.
[24] P. Felzenszwalb, D. McAllester, D. Ramanan, A discriminatively trained, multiscale, deformable part model, in: CVPR, 2008.
[25] W. Gao, H. Ai, S. Lao, Adaptive contour features in oriented granular space for human detection and segmentation, in: CVPR, 2009.
[26] Y. Liu, S. Shan, W. Zhang, X. Chen, W. Gao, Granularity-tunable gradients partition (GGP) descriptors for human detection, in: CVPR, 2009.
[27] K. Mikolajczyk, C. Schmid, A. Zisserman, Human detection based on a probabilistic assembly of robust part detectors, in: ECCV, 2004.
[28] B. Leibe, E. Seemann, B. Schiele, Pedestrian detection in crowded scenes, in: CVPR, 2005.
[29] J. Gall, V. Lempitsky, Class-specific hough forests for object detection, in: CVPR, 2009.
[30] J. Van de Weijer, T. Gevers, A.D. Bagdanov, Boosting color saliency in image feature detection, IEEE Trans. Pattern Anal. Mach. Intell. 28 (1) (2006) 150–156.
[31] T. Gevers, J. van de Weijer, H. Stokman, Color Image Processing: Methods and Applications: Color Feature Detection: An Overview, CRC Press, 2006.
[32] A. Abdel-Hakim, A. Farag, CSIFT: A SIFT descriptor with color invariant characteristics, in: CVPR, 2006.
[33] G.J. Burghouts, J.M. Geusebroek, Performance evaluation of local color invariants, Comput. Vision Image Understanding 113 (1) (2009) 48–62.
[34] R. Fergus, P. Perona, A. Zisserman, Object class recognition by unsupervised scale-invariant learning, in: CVPR, 2003.
[35] J. Shotton, J. Winn, C. Rother, A. Criminisi, TextonBoost for image understanding: multi-class object recognition and segmentation by jointly modeling texture, layout, and context, Int. J. Comput. Vision 81 (1) (2008) 2–23.
[36] J. Winn, A. Criminisi, T. Minka, Categorization by learned universal visual dictionary, in: ICCV, 2005.
[37] B. Manjunath, W. Ma, Texture features for browsing and retrieval of image data, IEEE Trans. Pattern Anal. Mach. Intell. 18 (8) (1996) 837–842.
[38] T. Ojala, M. Pietikainen, T. Maenpaa, Multiresolution gray-scale and rotation invariant texture classification with local binary patterns, IEEE Trans. Pattern Anal. Mach. Intell. 24 (7) (2002) 972–987.
[39] X. Tan, B. Triggs, Enhanced local texture feature sets for face recognition under difficult lighting conditions, AMFG, 2007.
[40] S. Liao, S. Li, Learning multi-scale block local binary patterns for face recognition, in: ICB, 2007.
[41] W. Zhang, S. Shan, W. Gao, X. Chen, H. Zhang, Local gabor binary pattern histogram sequence (LGBPHS): a novel non-statistical model for face representation and recognition, in: CVPR, 2005.
[42] Y. Mu, S. Yan, Y. Liu, T. Huang, B. Zhou, Discriminative local binary patterns for human detection in personal album, in: CVPR, 2008.
[43] Jie Chen, Shiguang Shan, Chu He, Guoying Zhao, M. Pietikainen, Xilin Chen, Wen Gao, WLD: a robust local image descriptor, IEEE Trans. Pattern Anal. Mach. Intell. 32 (9) (2010) 1705–1720.
[44] H. Wang, M. Ullah, A. Klaser, I. Laptev, C. Schmid, Evaluation of local spatio-temporal features for action recognition, in: BMVC, 2009.
[45] Y. Ke, R. Sukthankar, M. Hebert, Efficient visual event detection using volumetric features, in: ICCV, 2005.
[46] P. Viola, M. Jones, Rapid object detection using a boosted cascade of simple features, in: CVPR, 2001.
[47] Yazhou Liu, Xilin Chen, Contour-motion feature (CMF): a space–time approach for robust pedestrian detection, Pattern Recognition Lett. 30 (2) (2009) 148–156.
[48] G. Willems, T. Tuytelaars, L. Van Gool, An efficient dense and scale-invariant spatio-temporal interest point detector, in: ECCV, 2008.
[49] T. Lindeberg, Feature detection with automatic scale selection, Int. J. Comput. Vision 30 (2) (1998) 77–116.
[50] I. Laptev, M. Marszalek, C. Schmid, B. Rozenfeld, Learning realistic human actions from movies, in: CVPR, 2008.
[51] P. Scovanner, S. Ali, M. Shah, A 3-Dimensional SIFT descriptor and its application to action recognition, ACM Multimedia, 2007.
[52] A. Kläser, M. Marszalek, C. Schmid, A spatio-temporal descriptor based on 3D gradients, in: BMVC, 2008.
[53] L. Shao, R. Gao, Y. Liu, H. Zhang, Transform based spatio-temporal descriptors for human action recognition, Neurocomputing 74 (6) (2011) 962–973.
[54] Guoying Zhao, M. Pietikainen, Dynamic texture recognition using local binary patterns with an application to facial expressions, IEEE Trans Pattern Anal. Mach. Intell. 29 (6) (2007) 915–928.
[55] O. Tuzel, F. Porikli, P. Meer, Human detection via classification on Riemannian mani-folds, in: CVPR, 2007.
[56] X. Hong, H. Chang, S. Shan, X. Chen, W. Gao, Sigma set: a small second order statistical region descriptor, in: CVPR, 2009.
[57] B. Wu, R. Nevatia, Optimizing discrimination-efficiency tradeoff in integrating heterogeneous local features for object detection, in: CVPR, 2008.
[58] F. Han, Y. Shan, H.S. Sawhney, R. Kumar, Discovering class specific composite features through discriminative sampling with Swendsen–Wang cut, in: CVPR, 2008.
[59] L. Shao, L. Ji, A descriptor combining MHI and PCOG for human motion classification, in: CIVR, Xi'an, China, July 2010.
[60] X. Wang, T. Han, S. Yan, A HOG-LBP human detector with partial occlusion handling, in: ICCV, 2009.
[61] J. Shotton, A. Blake, R. Cipolla, Efficiently combining contour and texture cues for object recognition, in: BMVC, 2008.
[62] W. Schwartz, A. Kembhavi, D. Harwood, L. Davis, Human detection using partial least squares analysis, in: ICCV, 2009.
[63] B. Alexe, T. Deselaers, V. Ferrari, What is an object? in: CVPR, 2010.
[64] A. Kembhavi, B. Siddiquie, R. Miezianko, S. McCloskey, L. Davis, Scene it or not? Incremental Multiple Kernel Learning for Object Detection, in: ICCV, 2009.
[65] A. Bosch, A. Zisserman, X. Munoz, Representing shape with a spatial pyramid kernel, in: CIVR, 2007.
[66] A. Berg, J. Malik, Geometric blur for template matching, in: CVPR, 2001.
[67] L. Shao, R. Mattivi, Feature detector and descriptor evaluation in human action recognition, in: CIVR, Xi'an, China, July 2010.
[68] Y. Huang, K. Huang, L. Wang, D. Tao, X. Li, T. Tan, Enhanced biologically inspired model, in: CVPR, 2008.
[69] J. Fan, Y. Wu, S. Dai, Discriminative spatial attention for robust tracking, in: ECCV, 2010.
[70] A.D. Jepson, D.J. Fleet, T.F. El-Maraghi. Robust online appearance models for visual tracking, in: CVPR, 2001.
[71] Shaohua Kevin Zhou, R. Chellappa, B. Moghaddam, Visual tracking and recognition using appearance-adaptive models in particle filters, IEEE Trans. Image Process. 13 (11) (2004) 1491–1506.
[72] K. Lee, D. Kriegman, Online learning of probabilistic appearance manifolds for video-based recognition and tracking, in: CVPR, 2005.
[73] David A. Ross, Jongwoo Lim, Ruei-Sung Lin, Ming-Hsuan Yang, Incremental learning for robust visual tracking, Int. J. Comput. Vision 77 (3) (2008) 125–141.
[74] X. Li, W.M. Hu, Z.F. Zhang, Robust visual tracking based on incremental tensor subspace learning, in: ICCV, 2007.
[75] J. Wen, X. Gao, Incremental learning of weighted tensor subspace for visual tracking, in: SMC, 2009.
[76] H. Yang, Z. Song, R. Chen, An incremental PCA-HOG descriptor for robust visual hand tracking, in: ISVC, 2010.
[77] V. Arsigny, P. Fillard, X. Pennec, N. Ayache, Geometric means in a novel vector space structure on symmetric positive definite matrices, SIAM J. Matrix Anal. Appl. 29 (2006) 328–347.
[78] X. Li, W. Hu, Z. Zhang, X. Zhang, G. Luo, Visual tracking via incremental log-Euclidean Riemannian subspace learning, in: CVPR, 2008.
[79] Y. Wu, J. Cheng, J. Wang, H. Lu, Real-time visual tracking via incremental covariance tensor learning, in: ICCV, 2009.
[80] R.T. Collins, Yanxi Liu, M. Leordeanu, Online selection of discriminative tracking features, IEEE Trans. Pattern Anal. Mach. Intell. 27 (10) (2004) 1631–1643.
[81] S. Avidan, Ensemble tracking, in: CVPR, 2005.
[82] J. Wang, X. Chen, W. Gao, Online selecting discriminative tracking features using particle filter, in: CVPR, 2005.
[83] G. Li, D. Liang, Q. Huang, S. Jiang, W. Gao, Object tracking using incremental 2d-lda learning and Bayes inference, in: ICIP, 2008.
[84] X. Zhang, W. Hu, S. Maybank, X. Li, Graph based discriminative learning for robust and efficient object tracking, in: ICCV, 2007.
[85] M. Tian, W. Zhang, F. Liu, On-line ensemble SVM for robust object tracking, in: ACCV, 2007.
[86] M. Yang, J. Yuan, Y. Wu, Spatial selection for attentional visual tracking, in: CVPR, 2007.
[87] H. Grabner, H. Bischof, On-line boosting and vision, in: CVPR, 2006.
[88] H. Grabner, H. Bischof, Real-time tracking via on-line boosting, in: BMVC, 2006.
[89] A. Saffari, C. Leistner, J. Santner, M. Godec, H. Bischof, On-line random forests, in: OLCV, 2009.
[90] M. Wang, X.-S. Hua, J. Tang, R. Hong, Beyond distance measurement: constructing neighborhood similarity for video annotation, IEEE Trans. Multimedia 11 (3) (2009).
[91] M. Wang, X.-S. Hua, R. Hong, J. Tang, G.-J. Qi, Y. Song, Unified video annotation via multi-graph learning, IEEE Trans. Circuits Syst. Video Technol. 19 (5) (2009).
[92] H. Grabner, C. Leistner, H. Bischof, Semi-supervised on-line boosting for robust tracking, in: ECCV, 2008.
[93] B. Babenko, M.-H. Yang, S. Belongie, Visual tracking with online multiple instance learning, in: CVPR, 2009.
[94] B. Zeisl, C. Leistner, A. Saffari, H. Bischof, On-line semi-supervised multiple-instance boosting, in: CVPR, 2010.
[95] J. Santner, C. Leistner, A. Saffari, T. Pock, H. Bischof, PROST Parallel Robust Online Simple Tracking, in: CVPR, 2010.
[96] M. Werlberger, W. Trobin, T. Pock, A. Wedel, D. Cremers, H. Bischof, Anisotropic huber-l1 optical flow, in: BMVC, 2009.

[97] Yuan Li, Haizhou Ai, T. Yamashita, Shihong Lao, M. Kawade, Tracking in low frame rate video: a cascade particle filter with discriminative observers of different life spans, IEEE Trans. Pattern Anal. Mach. Intell. 30 (10) (2008) 1728–1740.
[98] M. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, L. Gool, Robust tracking-by-detection using a detector confidence particle filter, in: ICCV, 2009.
[99] C. Aeschliman, J. Park, A. Kak, A probabilistic framework for joint segmentation and tracking, in: CVPR, 2010.
[100] Y. Zhao, R.T. Collins, Shape constrained figure-ground segmentation and tracking, in: CVPR, 2009.
[101] L. Zhang, Y. Li, R. Nevatia, Global data association for multi-object tracking using network flows, in: CVPR, 2008.
[102] Y. Li, C. Huang, R. Nevatia, Learning to associate: hybridboosted multi-target tracker for crowded scene, in: CVPR, 2009.
[103] C. Huang, B. Wu, R. Nevatia, Robust object tracking by hierarchical association of detection responses, in: ECCV, 2008.
[104] T. Woodley, B. Stenger, R. Cipolla, Tracking using online feature selection and a local generative model, in: BMVC, 2007.
[105] H. Grabner, P. Roth, H. Bischof, Eigenboosting: combining discriminative and generative information, in: CVPR, 2007.
[106] M. Yang, Y. Wu, S. Lao, Intelligent collaborative tracking by mining auxiliary objects, in: CVPR, 2006.
[107] Y. Li, R. Nevatia, Key object driven multi-category object recognition, localization and tracking using spatio-temporal context, in: ECCV, 2008.
[108] S. Stalder. H. Grabner, L. Gool, Cascaded confidence filtering for improved tracking-by-detection, in: ECCV, 2010.
[109] P. Roth, S. Sternig, H. Grabner, H. Bischof, Classifier grids for robust adaptive object detection, in: CVPR, 2009.
[110] H. Grabner, J. Matas, L. Gool, P. Cattin, Tracking the invisible: learning where the object might be, in: CVPR, 2010.
[111] Z. Kalal, J. Matas, K. Mikolajczyk, Online learning of robust object detectors during unstable tracking, in: OLCV, 2009.
[112] Z. Kalal, K. Mikolajczyk, J. Matas, Forward–backward error: automatic detection of tracking failures, in: ICPR, 2010.
[113] S. Divvala, D. Hoiem, J. Hays, A. Efros, M. Hebert, An empirical study of context in object detection, in: CVPR, 2009.
[114] Y. Bar-Shalon, T. Fortmann, Tracking and Data Association, Academic Press, 1988.
[115] Peihua Li, Tianwen Zhang, Bo Ma, Unscented Kalman filter for visual curve tracking, Image and Vision Computing 22 (2) (2004) 157–164.
[116] K. Smith, D.G. Perez, J. Odobez, Using particles to track varying numbers of interacting people, in: CVPR, 2005.
[117] Zia Khan, T. Balch, F. Dellaert, MCMC-based particle filtering for tracking a variable number of interacting targets, IEEE Trans. Pattern Anal. Mach. Intell. 27 (11) (2005) 1805–1819.
[118] X. Zhang, W. Hu, Z. Zhao, Y. Wang, X. Li, Q. Wei, SVD based kalman particle filter for robust visual tracking, in: ICPR, 2008.
[119] S. McKenna, H. Nait-Charif, Tracking human motion using auxiliary particle filters and iterated likelihood weighting, IVC, 2007.
[120] X. Zhang, W. Hu, S. Maybank, A smarter particle filter, in: ACCV, 2009.
[121] J. Kwon, K.M. Lee, F.C. Park, Visual tracking via geometric particle filtering on the affine group with optimal importance functions, in: CVPR, 2009.
[122] G. Schindler, F. Dellaert, A rao-blackwellized partsconstellation tracker, in: ICCV workshop, 2005.
[123] J. Kwon, K. Lee, Tracking of abrupt motion using Wang-Landau Monte Carlo estimation, in: ECCV, 2008.
[124] J. Kwon, K. Lee, Tracking of a non-rigid object via patch-based dynamic appearance modeling and adaptive basin hopping Monte Carlo sampling, in: CVPR, 2009.
[125] J. Kwon, K. Lee, Visual tracking decomposition, in: CVPR, 2010.
[126] J. Liu, Monte Carlo Strategies in Scientific Computing, Springer, 2001.
[127] P. Viola, M. Jones, Robust real-time object detection, in: CVPR, 2001.
[128] A. Vedaldi, V. Gulshan, M. Varma, A. Zisserman, Multiple kernels for object detection, in: ICCV, 2009.
[129] C.H. Lampert, M.B. Blaschko, T. Hofmann, Efficient subwindow search: a branch and bound framework for object localization, IEEE Trans. Pattern Anal. Mach. Intell. 31 (12) (2009) 2129–2142.
[130] Sinno Jialin Pan, Qiang Yang, A survey on transfer learning, IEEE Trans. Knowl. Data Eng. 22 (10) (2010) 1345–1359.
[131] D. Koller, N. Friedman, Probabilistic Graphical Models: Principles and Techniques, MIT Press, 2009.

Hanxuan Yang got his master degree of engineering from South China Agricultural University in 2011. During 2009–2010, he was a visiting student with Laboratory for Culture Integration Engineering, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, doing research related to visual hand tracking and recognition. His research interests are machine learning and computer vision.

Ling Shao received the B.Eng. degree in Electronic Engineering from the University of Science and Technology of China (USTC), the M.Sc. degree in Medical Image Analysis and the Ph.D. (D.Phil.) degree in Computer Vision from the University of Oxford. Dr. Ling Shao is currently a Senior Lecturer (Associate Professor) in the Department of Electronic and Electrical Engineering at the University of Sheffield. Before joining Sheffield University, he worked for 4 years as a Senior Research Scientist in the Video Processing and Analysis Group, Philips Research Laboratories, Eindhoven, The Netherlands. His research interests include computer vision, pattern recognition and video processing. He has published over 60 academic papers in refereed journals and conference proceedings and has filed over 10 patent applications. Ling Shao is an associate editor of the International Journal of Image and Graphics, the EURASIP Journal on Advances in Signal Processing, and Neurocomputing, and has edited several special issues for journals of IEEE, Elsevier and Springer. He has been serving as Program Committee member for many international conferences, including ICIP, ICASSP, ICME, ICMR, ACM MM, CIVR, BMVC, etc. He is a senior member of the IEEE.

Feng Zheng received the Master degree of applied mathematics from Hubei University in 2009. He is currently with the Shenzhen Institutes of Advanced Technology (SIAT), Chinese Academy of Sciences (CAS), as a research assistant. His research interests include machine learning, computer vision and human-computer interaction.

Liang Wang received both the B. Eng. and M. Eng. degrees from Anhui University in 1997 and 2000 respectively, and the Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences (CAS) in 2004. From 2004 to 2010, he has been working as a Research Assistant at Imperial College London, United Kingdom and Monash University, Australia, a Research Fellow at the University of Melbourne, Australia, and a lecturer at the University of Bath, United Kingdom, respectively. Currently, he is a Professor of Hundred Talents Program at the National Lab of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, P.R. China. His major research interests include machine learning, pattern recognition and computer vision. He has widely published at highly ranked international journals such as IEEE TPAMI and IEEE TIP, and leading international conferences such as CVPR, ICCV and ICDM. He has obtained several honors and awards such as the Special Prize of the Presidential Scholarship of Chinese Academy of Sciences. He is currently a Senior Member of IEEE, as well as a member of BMVA. He is an associate editor of IEEE Transactions on Systems, Man and Cybernetics – Part B, International Journal of Image and Graphics, Signal Processing, Neurocomputing and International Journal of Cognitive Biometrics. He is a guest editor of four special issues, a co-editor of five edited books, and a co-chair of seven international workshops.

Zhan Song received the Ph.D. degree in Mechanical and Automation Engineering from the Chinese University of Hong Kong, Hong Kong, in 2008. He is currently with the Shenzhen Institutes of Advanced Technology (SIAT), Chinese Academy of Sciences (CAS), as an assistant researcher. His current research interests include structured-light based sensing, image processing, 3-D face recognition, and human–computer interaction.