Artif Intell Rev (2016) 46:1–39
DOI 10.1007/s10462-015-9454-6

Recent advances on multicue object tracking: a survey

Gurjit Singh Walia[1] · **Rajiv Kapoor**[2]

Published online: 9 January 2016

Abstract The performance of single cue object tracking algorithms may degrade due to complex nature of visual world and environment challenges. In recent past, multicue object tracking methods using single or multiple sensors such as vision, thermal, infrared, laser, radar, audio, and RFID are explored to a great extent. It was acknowledged that combining multiple orthogonal cues enhance tracking performance over single cue methods. The aim of this paper is to categorize multicue tracking methods into single-modal and multi-modal and to list out new trends in this field via investigation of representative work. The categorized works are also tabulated in order to give detailed overview of latest advancement. The person tracking datasets are analyzed and their statistical parameters are tabulated. The tracking performance measures are also categorized depending upon availability of ground truth data. Our review gauges the gap between reported work and future demands for object tracking.

Keywords Object tracking · Multicue · Data set · Tracking evaluation · Computer vision

1 Introduction

The object tracking aims at analysis of video sequences for localization of object in sub sequences frames. It is foremost important due to its myriad of applications in field of computer vision such as driver assistance systems, video surveillance, man-machine interaction, autonomous robot, medical imaging, activity analysis, augmented reality, video indexing, traffic control and much more. However, object tracking in video is very challenging due to: dynamic environment conditions, conversions from 3D to 2D world, real time requirements, full or partial occlusion, clutter background, pose variations, abrupt object motion, appear-

✉ Rajiv Kapoor
rajiv.kapoor@dce.ac.in

[1] Defense Research and Development Organization, Ministry of Defense, Delhi, India

[2] Department of Electronics and Communication, Delhi Technological University (Formerly DCE), Bawana Road, Delhi, India

ance variability, noise in video (snow, fog, dust, fire etc.). In last two decade, a number of object tracking algorithms were proposed with the aim to cater for one or more of these listed challenges during tracking. The object tracking methods using cues (color, motion, orientation, spatial energy, shape, texture, infrared, and position, etc.) was extensively explored and discussed at length (Yilmaz et al. 2006; Yang et al. 2011; Smeulders et al. 2014). But, tracker based upon single cue can limitedly address tracking challenges. Hence, in last few years, there is paradigm shift of research work from single cue to multicue tracking methods. Under multicue object tracking, complementary cues are added in order to cater for different challenges. The multiple cues are extracted either from single-modal (one sensor) or from multi-modal (more than one sensor). In literature, it was well argued that cues which are good at one point of time may degrade at other instance during tracking process. Color cue being computationally efficient, invariant to scaling and handle partial occlusion cannot provide spatial information of target. Although, semi overlapping regions were used for embedding spatial information into color space (Maggio and Cavallaro 2005), single color cue cannot handle illumination changes, full occlusion, and texture changes of target. HSV color model is invariant to illumination changes but it cannot handle low saturation videos where hue channel is affected to large extent with small change in saturation (Hong et al. 2009). Motion feature can be detected from consequent frames but could not provide shape information of object. In order to augment motion cue, spatial appearance along with motion was extracted as spatial-temporal motion energy which is rich in texture information and can handle illumination changes and background clutters (Cannons et al. 2010). Texture cue representing fine level details of object can be extracted with low computational power. But, due to changes in texture from fine to coarse, tracker using only texture cue degrades performance as object moves away from the camera. Gradient cue being extracted from chromatic contents cater for tilting and deformation of object but tracker based upon only gradient cue degrade under heavy clutter environment. Shape cue being one of prominent feature for rigid object fails to track non rigid object or object whose shape changes with time. In addition, it is also greatly affected by change in view angle of sensing device. Depth cue extracted from Kinect source or stereo vision is illumination and color invariant but unable to provide motion information and degrade its performance when object lack texture information or under occlusion (Wang et al. 2014; Munoz-salinas et al. 2008). Apart from vision sensors, thermal sensors were extensively explored for object tracking. The thermal sensors are based upon thermal radiation from human superficial blood vessels or from any object thermal profile. Thermal sensors are insensitive to illumination changes, disguise and pose variation of object. However, thermal cue is affected by ambient temperature and it is difficult to distinguish thermal profile of different people in crowd. Thermal sensor has high image noise and low resolution as compared to vision sensors (Kong et al. 2005). In addition, inherent hallo effect and their inability to provide texture information limit their usage as single modality for object tracking (Conaire et al. 2008). Audio sensors can faithfully provide direction of moving object and are very discriminant when present but they are unreliable for object far away from sensors. Laser scanner has potential advantages such as low computational requirement, easy projection of laser data to rectangular coordinate system and insensitive to illumination changes but could not distinguish objects once track of objects is lost under heavy cluttered environments (Cui et al. 2008). Radio frequency sensors (i.e. radar, RFID) can provide accurate range information and locate object behind wall but mostly needs to be compensated with other modularity with rich information of object. In sum, single cue trackers are in general less reliable and less robust to environment conditions. The complimentary cues either from same or different sensors are fused for building robust tracking systems. In human sensory system, different cues are either suppressed or boosted depending upon their reliability for achieving adaptive fusion of data from multisensory (Murphy 1996). The reliability of visual cues can be estimated either from correlation among cues or from ambiguity of cues (Jacobs 2002). Hence, in last few years, tracking researchers are motivated to shift their work towards either development of efficient cues or integration of multicue for robust tracking.

In 2006, review of object tracking techniques was addressed (Yilmaz et al. 2006). The review considered object representations for classification of object tracking methods. The object representation was classified into three categories as point correspondence, contour evolution or geometric models. In addition, object detection methods were briefly discussed. In similar line, various visual features descriptor for object representation along with visual tracking methods were reviewed (Yang et al. 2011). The feature descriptors for visual tracking were categorized as color, gradient, texture, spatio-temporal and multiple features. Further, the visual object tracking techniques were classified into Monte Carlo sampling, integration of context information and online learning. Sparse coding based visual tracking techniques were reviewed (Zhang et al. 2013c). But review was mainly focused on experimental analysis of some of representative work for sparse coding based appearance modeling and target tracking. In 2014, experimental survey for visual tracking was proposed (Smeulders et al. 2014). In this survey, algorithms that perform either matching or classification for object tracking were chosen for experiments. Nineteen trackers for which source code was available were evaluated for both qualitative and quantitative measures on large length of videos under various test scenarios. Also, survey on human detection techniques was proposed (Walia and Kapoor 2014a). But this review mainly focused on single-modal and multi-modal human detection methods. In last few years, object tracking work was extensively explored with main focus of research community on single-modal/multi-modal multicue tracking algorithms. Our survey mainly categorized and analyzed object tracking methods based upon single-modal and multi-modal multicue tracking.

With more than ten trackers being reported every year and also presence of flurry of work in literature, we feel time is ripe to summarize multicue tracking techniques as, in best of our knowledge, no such survey exists which exclusively categorizes multicue tracking methods. In order to gauge the gap in multi cue tracking work, we have limited our scope of this survey to object tracking based upon multiple different features extracted from either single-modal or multi-modal. We hope that our comprehensive survey on multicue object tracking will give seminal learning and new directions of research to academician, researchers and computer vision community. The remaining paper is organized as follows.

The aim of this survey is to characterize and summarize single-modal and multi-modal multicue techniques and analyse their merits and demerits. The single-modal multicue object tracking techniques are reviewed and tabulated in Sect. 2. In Sect. 3, multi-modal object tracking techniques are critically reviewed and tabulated. The different evaluation measures for tracking along with available online database for multicue object tracking are summarized in Sect. 4. The discussion of reported work is also followed in respective section in order to list out merit and demerits of work. In Sect. 5, concluding remarks and future directions in this field of multicue object tracking are sketched.

2 Multicue object tracking

The multicue object tracking aims at localization of object through processing of data acquired either from single-modal or from multi-modal sensors. The failure of single cue under different challenging scenarios is compensated by complementary cues. In general, multicue tracking framework can be illustrated as flow diagram Fig. 1. Where, sensors considered for

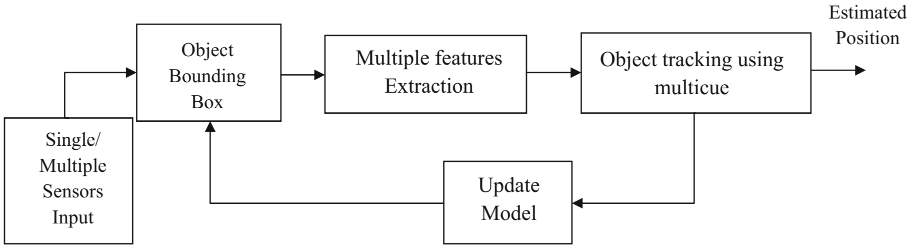

Fig. 1 Flow chart of multicue object tracking

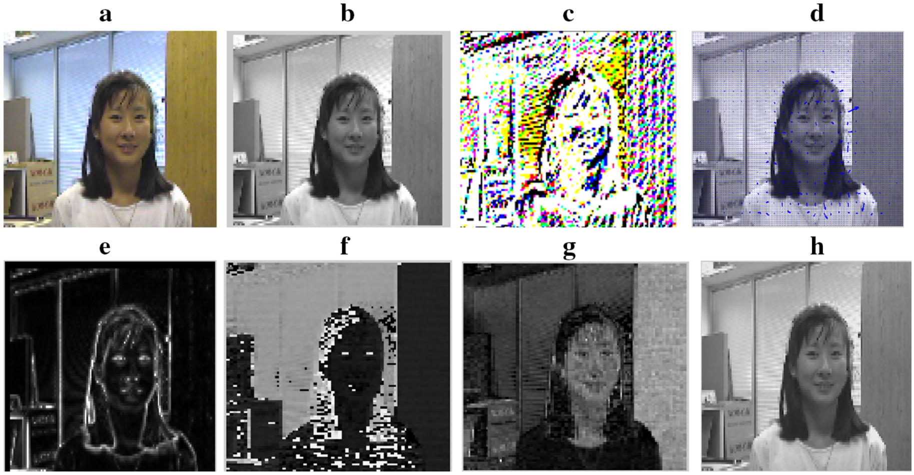

Fig. 2 Features for multicue tracker from sample image for HTS dataset (HTS): **a** original image, **b** gray intensity, **c** gradient, **d** optical flow, **e** texture, **f** hue channel, **g** S channel, **h** Y channel

tracking are either single sensor such as vision, infrared, Kinect etc. or different combination of sensors such as vision, laser, radar, microphone, infrared, Microsoft Kinect, and RFID.

In general, cues used for object tracking are color, motion, orientation, spatial energy, shape, texture, infrared, and position, depth, disparity etc.. The different extracted cues from image taken from HTS dataset (HTS) are shown in Fig. 2. In tracking framework, cues are further processed using either deterministic or stochastic methods for estimation of object location. The deterministic methods were focused on minimization of cost function representing relationship between estimated object and ground truth object and hence reduced the search space. On the other hand, in stochastic methods, stochastic factors were added during object search operation. In addition, methods for handling target occlusion and clutters were included. In general, target model may be updated with new estimated parameters in order to cater for variation in target shape during tracking process.

In our review work, we have classified multicue object tracking work based upon data acquisition sources. One group of researchers considered moving or stationary sensor for data acquisition and tracking was performed by extracting different cues from single sensor. Another group of researchers used multiple sensors for data acquisition and tracking was performed by extracting one or more cues from different modularity. The classification of multicue object tracking work is illustrated in Fig. 3.

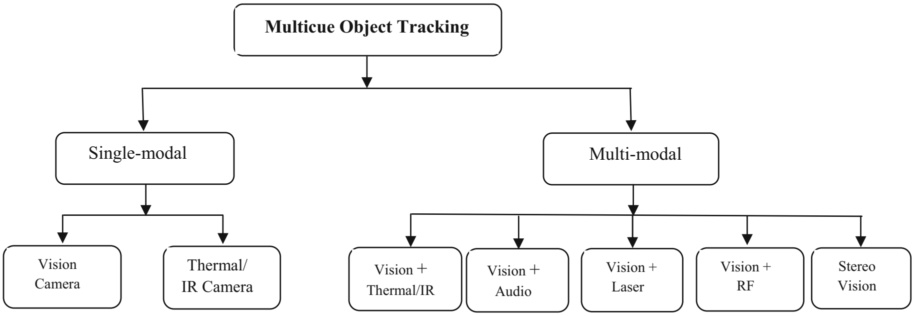

Fig. 3 Classification of multicue object tracking techniques

Being rich source of information, vision or thermal camera were exploited for extracting multiple complementary cues for video tracking. On the other hand, RFID, radar, laser, audio sensors provide limited information for vision tracking. However, mostly, this information aids vision camera for robust object tracking. The stereo vision camera provides disparity information that was exploited for visual target tracking using multicues. Taking mode of data acquisition and operation performed as reference, we have classified multi cue object tracking work and examined these methods.

2.1 Single-modal multicue object tracking

The vision and thermal camera has abundant information contents which were exploited for deducing number of complementary cues. In this section, we have reviewed tracking methods based upon different multiple features extracted from single sensory. The section is concluded with discussion on reviewed techniques. The different cues such as shape, texture, color, intensity, motion, position and orientation, etc. were extracted from single sensor and fused for final state estimation. The representative works under multicue tracking using single sensory are reviewed in turn.

2.1.1 Vision camera

In last decade, vision camera based multicue object tracking techniques were explored to a great extent. The cues extracted from vision camera were prone to degradation with changes in environment scenarios. However, orthogonal multiple cues were exploited for robust and reliable object tracking. In this review, we have categorized multicue object tracking work using single vision camera into techniques based upon operation performed on captured video sequences. The details of categorized work are discussed in turn.

Firstly, in general, deterministic methods mainly aimed at minimization of cost function representing relationship between centroid of tracked object and that of ground truth. The representation works for multicue tracking under deterministic methods using single sensory are detailed in turn. Multicue object tracking work gain momentum with the introduction of head tracking algorithm (Birchfield 1998). The authors used color cue extracted from head interior and intensity cue extracted from head perimeter for tracking head motion. The local search decided based upon score of two cues was used for locating person head in subsequent frames. The proposed algorithm claimed to be robust to head rotation, tilting, movement of camera and partial occlusion. However, under this work cues score were not fused in adaptive manner for catering complex motion under variable environment. Democratic integration of five cues extracted from vision sensor was presented (Triesch and Malsburg 2001). Under democratic integration, adaptive mechanism boosts good cues and at the same time suppresses discordant cues. The motion, color, shape, position, contrast cues were used for locating object in subsequent frames. For each cue, saliency map was created by comparing each feature in current frame with reference frame. The reliability of each cue was estimated based upon its performance over previous frames. The saliency map of each cue was weighted with reliability factor for estimating overall saliency map and hence prediction of target position. However, two limitations of democratic integration: (a) incapability to multi hypothesis tracking, (b) false positive tracking was discussed (Spengler and Schiele 2003). In order to solve these problems, authors proposed a framework for robust object tracking using self-adaptions of cues and self-organization of integration method. The final estimated probability distribution of state was feedback to individual cues as well as integration unit for their adaptions. New quality measure along with occlusion handler for adaptive multicue integration using mean shift algorithm was proposed (Hong et al. 2009). Under this work, color cue along with orthogonal motion cue was extracted for handling low saturation object under full occlusion. In order to track objects which were lost in previous frames using mean shift algorithm, authors presented occlusion handler where size of search window changed if full occlusion was detected. Cue reliability in current frame was determined based upon its performance in previous frame and overall probability map was estimated as weighted sum of individual maps. The results were presented on thirteen sets of video sequence with different level of difficulties and revealed that proposed method can handle occlusion, clutter and low saturation video sequences. The edges along with color cues were exploited in centroid shifting framework for robust object tracking (Hong et al. 2015). The non-uniform quantization of color cue was performed for obtaining better image quality. The edges profile for object region was extracted using canny operator and motion based frame difference. The shift in edge color co-occurrence centroid was estimated without any iteration. The scaling of object was catered using edge-edge distance histogram which exploited smoothness of edges distance in consecutive frames. The proposed algorithm was tested on limited data set. The author claimed proposed algorithm to be robust to partial occlusion. As edge and color cues were both illumination dependent, performance of tracking under object rotation and illumination changes needs to be investigated. In sum, in most of deterministic methods, current state estimation depends upon estimated parameters from previous state which limit their ability to track occluded or lost targets.

Secondly, in another line of research, in recent past, multicue tracking is performed by introducing some stochastic factor during state prediction. The stochastic methods such as Kalman filter, Particle filter and its variants were preferred over contemporary methods due to their potential advantages such as fusion of multiple features, handling different level of uncertainty, and easy up-gradation for multiple object tracking. In order to reach global optimum location of targeted object, a stochastic factor was intentionally inserted during object search operation. In general, these techniques were covered under umbrella of Monte Carlo simulation where state PDF was represented by set of samples. The representation works for multicue tracking under stochastic methods using single sensory are detailed in turn. In order to compensate for variation in texture of object during tracking, texture cue and color cue in particle filter and Gaussian sum particle filter framework were combined (Brasnett et al. 2005). The color histogram was extracted from RGB color model and texture feature was recovered using discrete wavelet transform. The individual cue likelihood score in each frame was estimated by comparing current features with reference features using Bhattacharyya distance measure. The overall likelihood in the frame was determined as product of individual likelihoods. Inability of color histogram to provide spatial information, spatial-color mixture of Gaussian for object color and proposed spatial histogram for object shape was introduced (Wang and David 2006). In this work, shape was represented by spatial distribution of edges, edge points and gradient intensity along edges. In particle filter framework, individual similarity measures for color and shape were estimated and used for finding overall likelihood as product of individual likelihood. However, proposed techniques (Brasnett et al. 2005; Wang and David 2006) were not using any adaptive fusion mechanism, tested on limited data set under control environment with no consideration of resources utilization. In order to efficiently allocate resources among multiple cues, particle filter and cue processor frameworks for 3D object tracking in video sequences was proposed (Loy et al. 2002). The authors used depth, skin color and radial symmetry cues for head tracking using particle filter framework. In multi hypothesis tracking, generated particles from particle filter were handed over to cue processor which determined cues likelihood and assigned resources to cues based upon their performance. The individual likelihood score for active cues were fused as product of their likelihoods for final state estimation. For sample impoverishment and degeneracy solution in particle filter framework, particle filter based upon particle swarm optimization for robust object tracking was proposed (Zheng and Meng 2008). The color and Haar histograms of reference object were compared with propagated particles histograms using Bhattacharya coefficient. The individual likelihood scores were combined using linear combination of individual likelihood score for final state estimation. However, no adaptive fusion of cues was considered in proposed work. Color and foreground cue for multiple objects tracking in particle filter framework was considered (Kumar et al. 2007). Likelihood of each particle was determined through likelihoods product for color cue and foreground cue. This work was extended for tracking multiple objects in video sequences captured through moving camera (Kumar et al. 2008). Further, for non-overlapping cameras, facial and color information was exploited for person detection (Kumar and Dogancay 2011). In an attempt towards adaptive fusion at likelihood level, color cue along with shape cue in modified particle filter framework was considered for tracking objects (Shen et al. 2003). The inability of color cue to provide shape information of tracked object was compensated by iincorporating shape cue in their algorithm. The cues were adaptively integrated before posterior state estimation by estimating L2 norm between the centroid of final state estimation and that of individual cue state estimation. Also, to cater time varying objects, authors introduced target updation model based upon average likelihood of particles. However, the results were tested on single video sequence and no quantitative performance metrics were evaluated. For efficient movement of particles during interferences, Camshift algorithm in particle filter framework was embedded (Yin et al. 2011). The color and motion probability distribution maps were extracted and combined map was estimated. The adaptive integration of cues was performed by proposed combined likelihood estimation where cues weights were in concurrence with their average similarity coefficient. Due to Brownian motion and Camshift iteration of particles, authors claimed that effectiveness of particles increased and hence small numbers of particles were required for tracking object under illumination changes and background clutters. However, reliability of cues was estimated based upon previous measurements rather than only current measurements. Difficulties in integrating motion cue with color cue for hand tracking was discussed and a multi stage framework for tracking complex motion of hand was proposed (Zhang et al. 2015). The feature points were (1) generated through color cue, (2) selected through representation, compact and diverse criteria, (3) tracked using motion cue. The integration of cues was performed in feature point selection process and tested for hand video sequences. As extension of previous work (Brasnett et al. 2005), estimation of reliability of each cue based upon current measurement rather than earlier approach of estimation of cue reliability from previous observations was discussed

(Brasnett et al. 2007). The authors combined color, texture and edge cues in particle filter framework. The overall likelihood score was determined as product of individual cue likelihoods which were separately weighted by reliability factor. Also, solution for automatic tuning of noise parameter for cues was presented. However, contribution of all particles was not considered for deciding entire cue reliability score. Inability of color cue to provide spatial information of object was compensated by orientation cue (Maggio et al. 2007). The orientation histogram representing shape and edges of object catered for illumination and clutters variations. In particle filter framework, cues were adaptively weighted by measuring spatial uncertainty based upon weighted covariance matrix. More uncertainty leads to less reliability of cues and hence less contribution in overall likelihood estimation. The more features can be added to proposed model for better robustness in tracking. A new multi feature re-sampling algorithm was introduced where number of particles for each cues were re-sampled based upon performance of cue. In order to cater for texture variation during tracking, adaptive combination of three complementary cues i.e. color, local binary pattern and histogram of oriented gradient extracted from vision camera was proposed (Dou and Li 2014). The histogram of oriented gradient along with color cue can faithfully estimate shape of object. In addition, local binary pattern recover texture features of object and cater for rotational error. In particle filter framework, individual cue likelihood scores were estimated using Bhattacharya distance. In each frame, the adaptive cue weights were estimated through measurement of Euclidean distance between estimated target position using individual cue and estimated target position using all cues. Further, overall likelihood score was determined as weighted linear combination of individual cue likelihood scores. Multi-graph based ranking method for fusing HOG,intensity and LBP features in particle filter framework were introduced (Yang et al. 2015). The candidate objects having highest ranking score was labeled as tracked object for current frame. Robustness of solution was introduced through consideration of temporal consistency in video sequences. Evolutionary particle filter framework for adaptively combining color cue with texture cue was considered (Zhang et al. 2013b). The HSV color model was used for creation of color histogram and local binary pattern was used for extraction of texture histogram. The cues were fused adaptively using two fusion methodology being decided based on re-sampling requirement in particle filter. In order to improve accuracy, adaptive genetic algorithm was used for performing re-sampling and hence to overcome particle degeneracy. However, proposed algorithm was tested on three self-generated video sequences. Due to its ability to distinguish moving object from background clutters, object motion energy along with other cues was explored extensively for object tracking. A combination of color and spatial temporal motion energy for human tracking using particle filter framework was presented (Zhou et al. 2014). The spatial temporal motion energy was a three dimensional feature extracted from orientated selective band pass filter and edge detector. It can effectively handle clutter, and variation in illumination and shape. The cues adaptive contribution was determined through measurement of Euclidean distance between centroid estimated using cues and that of final estimation. The reliability score of cue was estimated by hyperbolic tangent function of mean Euclidean distances over set of measurements. The cues were adaptively combined using estimated reliability scores. The applicability of proposed work for far away multiple objects need to be explored. For multiple people tracking, independent tracker for each of object in cluttered scene was discussed (Sun and Bentabet 2010). The problem was modeled in DSmT framework where hyper power set with elements as individual tracker, ignorance and conflict were predefined. The location and color cue provided confidence for each of elements of hyper power set and any conflict among cues due to clutter or occlusion was resolved using DSmT fusion. The individual masses were combined using combination rule and final belief about each particle was determined. However, proposed method can only be applicable for resolving conflict among multiple targets not for single object tracking. Considering independent appearance and motion cues, multi hypothesis method for multiple objects tracking under closed vicinity was proposed (Ying et al. 2014). The overall likelihood was determined as weighted sum of dynamic and appearance confidences where optical flow confidence was deciding factor under constrained conditions. The authors also implemented dynamic updation of model and repulsion inertia model for avoiding local minimum traps. However, automatic adjustment of tracking parameters needs further investigations for scenarios where no data is available a prior.

Instead of adaptive fusion under single tracking model, switching among three interactive models for object tracking in video sequences was proposed (Dou and Jianxun 2014). In particle filter framework, authors considered three object models as: histogram of orientated gradient (HOG), completed local ternary patterns (CLTP), corrected background weighted histogram (CBWH). The switching between models was decided based upon model transition probability and model likelihood estimated in each frame. The results on three datasets revealed that proposed interactive multiple model particle filter (IMM_PF) was robust to clutter, rotation and partial occlusion.

Instead of fusing multiple cues, fusion of multiple likelihoods in multi feature PSO tracking algorithm was proposed (Ahmed et al. 2014). The authors performed study on different similarity measures and reported that Bhattacharya coefficient measure was rotation and scale invariant. On the contrary, similarity index measure (SIM) provided spatial information and easily computed. The likelihood measures were extracted sequentially and overall likelihood score was estimated as linear combination of individual scores. The authors claimed that proposed scheme was invariant to scaling, rotation and zooming. However, only single grey level cue was used for feature extraction.

In contrast to earlier approaches, feature level fusion for human detection and tracking was proposed (Liu et al. 2015). First, point ensemble image (PEI) was extracted from captured RGB-D data. In order to reduce search space, initial elimination of non-human candidates was performed using support vector machine trained using two features as histogram of height difference (HOHD) and joint height and color histogram (JHCH). After detecting plausible candidates, 3D JHCH was extracted and tracking was performed using Kalman filter where new similarity measure was used for associating each target with respective track. However, performance of proposed approach for non-human and maneuverings human needs further studies. Three cues i.e. gradient, color and local binary pattern for building robust appearance model was considered (Zeng et al. 2013). The authors claimed that using three independent cues lead to more discriminative appearance model which was used for updating classifier and hence estimation of object state. The multi feature joint descriptor (MFJD) based upon color and HOG was proposed as representation of object (Li et al. 2014c). The feature weights were adjusted using contrast and stability scores. The object tracking was performed using mean shift framework and author claimed that proposed joint feature could handle scale change and deformation of object.

Instead of using multiple cues in single Bayesian filer, multiple Bayesian filters tracking framework was introduced (Francesc et al. 2008). For each cue, separate Bayesian filter was used and integration among different filters was performed at hypothesis correction stage. The authors assumed dependence among estimation from different filters. The experiments were performed with proposed four cues tracker i.e. bounding box, fisher color map, color distribution and contour. The final object state at contour estimation was determined assuming its dependencies with previous filters outputs. However, as reported in Leichter et al.

(2014), dependencies of cues assumed (Francesc et al. 2008, 2005) were in fact independent measurement.

In sum, vision camera being extensively used for multicue object tracking due to their ability to provide rich source of information about targeted object. In this section, we have classified multicue tracking methods using vision camera into deterministic and stochastic methods. In addition to this, in order to explore other related domains, methods considering multiple models, multiple likelihood or cue with multiple features were reviewed. The multicue tracking methods using single vision camera are also tabulated in Table 1.

2.1.2 Thermal/IR camera

With the advancement in thermal/infrared camera technologies, thermal/infrared cameras were considered to be a good alternative for night vision applications. The passive infrared and thermal cameras were insensitive to illumination changes, disguise and pose variation of targeted objects. In this section, we review multicue tracking methods using different multiple features extracted from thermal/infrared cameras (see Table 2). Pedestrian detection along with tracking in infrared video sequences was presented (Congxia et al. 2007). The pedestrian detection was performed at two stages: (a) shape cue discriminate pedestrian from non-pedestrian, (b) appearance cue aids in positioning of pedestrian in frame. The shot consisting of sequences of frames were segmented and within shot tracking was performed using graph matching approach. The shape and position cues were used for establishing associations among detected pedestrians. The weighted contributions of these cues were used in graph matching algorithm for final track establishment. The results were presented on OSU and WVU databases. The authors claimed that proposed framework can perform better under occlusion without prior knowledge of motion trajectory; however, tracking pedestrians missed during detection stage needed further investigation. In infrared video, inability of intensity cue to provide shape information was compensated by fusing edge information (Wang et al. 2012). The particle filter framework was used for tracking pedestrian in infrared video sequences. The cue discriminant ability measured based on position and size difference between reference and target model was introduced. At particle level, overall likelihood score was estimated as product of weighted likelihood score from both cues. Where weights were decided from new measure of discriminant ability named as relative discriminative coefficient. The results were presented on OTCBVS dataset. In similar line, joint grey and local binary pattern histogram was proposed for representation of object (Li et al. 2014b). The extracted joint feature was used in mean shift framework for efficiently object tracking in IR video sequences. The experiment results were presented on OTCBVS and VIVID database for tracking dim small objects. The use of evolutionary technique for optimum fusion of multiple cues was explored (Zhang et al. 2013a). The authors used intensity and gradient cues for tracking objects in infrared videos. In particle filter framework, particle swarm optimization was used not only for overcoming particle degeneracy but also for combining particle weights during particle updation stage. The performance was evaluated on limited OTCBVS and own infrared datasets.

Although, thermal camera found to be good alternative for night vision application, multicue tracking using only infrared sensor was limited to targets with thermal radiation profile. In addition, it was difficult to distinguish target having similar thermal profile under multicue tracking using single thermal/IR sensor. Hence, multicue tracking using infrared/thermal sensors needs to be augmented in Multi-modal framework where other modularity aids thermal/IR cues for building robust tracking platform.

Table 1 Recent advances on singlemodal multicue object tracking

Reference	Algorithm	Tracking cues	Integration of cues	Database	Description
(A) Vision camera					
Birchfield (1998)	Local search	Intensity gradient and color	Adding up percentage of gradient and color score	HTS	Robust to complex head motion, tilting and rotation
Triesch and Malsburg (2001)	Saliency map	Motion, color, shape, contrast, position	Weighted saliency map	Self-generated	Assume statistical dependencies of cues
Shen et al. (2003)	Modified particle filter	Color and shape	Adaptive integration where cue reliability by L2 norm	HTS	No quantitative analysis and tested on limited video
Brasnett et al. (2005)	PF, Gaussian sum PF	Color and texture	Product of individual likelihood but not adaptive	Synthetic and natural seq.	Robust and improved accuracy
Brasnett et al. (2007)	Particle filter	Color, texture and edge	Product of individual weighted likelihood	Six natural video seq.	Adaptive fusion with automatic tuning of parameter
Loy et al. (2002)	Particle filter	Depth, skin color and radial symmetry	Product of active cue likelihood	Self-generated	Adaptive allocation of resources for multi rate cues
Zheng and Meng (2008)	Particle filter	Color and local statistic feature	Linear combination of two likelihood	Self-generated	PSO overcome particle impoverishment
Yin et al. (2011)	Camshift and Particle filter	Color and motion	Adaptive based upon previous frame results	CAVIAR, PETS	CamShift aided particle filter for propagating particles
Hong et al. (2009)	Mean shift	Color and motion	Weighted combined probability map	Self-generated 13 video seq.	Quality measure based upon previous frame performance
Hong et al. (2015)	Modified Centroid shift	Color and edges	Centroid shift estimated from edge-color co-occurrence	CAVIAR, Supermarket	Proposed solution for scaling but no adaptive fusion
Zhang et al. (2013b)	Particle filter	Color, texture	Re-sampling adaptively decide fusion methods	Self-generated	Adaptive GA for re-sampling and multicue fusion
Maggio et al. (2007)	Particle filter	Color and orientation	Cue reliability used in linear combination of likelihood.	HTS,CAVIAR, PETS-2001	Eliminate manual selection of feature
Dou and Li (2014)	Particle filter	Color, HOG and LBP	Euclidean distance measure for combining individual cue	HTS	Cater for illumination changes, camera motion and clutter

Table 1 continued

Reference	Algorithm	Tracking cues	Integration of cues	Database	Description
Yang et al. (2015)	Particle filter	Intensity, LBP, HOG	Multi-graph ranking method	6 online video sequences	Could run at 2–3 frames per second
Zhou et al. (2014)	Particle filter	Color, spatial temporal motion energy	Hyperbolic tangent function of Euclidean distances	HTS and in house data set	Cater for pose/illumination variation and occlusion
Dou and Jianxun (2014)	IMM particle filter	CLTP, HOG, CBWH	Model likelihood or model transition probability	PETS,HTS, CAVIAR	Robust to rotation, partial occlusion and clutter
Sun and Bentabet (2010)	Particle filter	Color and location	Combination rules for finding maximum belief	Self-generated single video	DSmT for resolving conflict among multiple targets
Liu et al. (2015)	Kalman filter	Height and color	JHCH, similarity function with location and color.	Office, mobile clothing store	PEI representation for better human detection and tracking
Ying et al. (2014)	Kalman filter	LBP, Optical flow and dynamics	Weighted sum of appearance and dynamic confidence	TRECVID and PETS	Handle close vicinity problems of multiple person tracking
Francesc et al. (2008)	Kalman and Particle filter	Color, contour, fisher space, bounding box	Product with hypothesis correction from previous stage	Self-generated	Assumed dependencies among cues at measurement level
Kumar et al. (2007)	Particle filter	Color, foreground cues	Product of color and foreground likelihood	CAVIAR	Demonstrated tracker for tracking multiple objects
Li et al. (2014c)	Mean shift	Color and HOG	Joint histogram generated at object representation level	Freeman and PETS2009	Feature weights adaptive based upon contrast and stability
(B) Thermal/IR camera					
Congxia et al. (2007)	Graph matching	Shape and appearance and location	Shape and location cues are adaptively fused in tracking	OSU and WVU infrared	Pedestrian detection rate may affect accuracy
Wang et al. (2012)	Particle filter	Intensity and edges	Adaptive fusion at particle level using product of weighted likelihood	OTCBVS three thermal video seq.	Proposed new measure for cue discriminant ability
Zhang et al. (2013a)	PSO-particle filter	Intensity and gradient	PSO based optimum data fusion from cues	OTCBVS and self-generated	Demonstrated on infrared videos on limited set
Li et al. (2014b)	Mean shift	Gray and local binary pattern	Cues were fused at object model level as joint histogram	OTCBVS and VIVID	Efficiently track small dim object in IR sequences

Table 2 Recent advances on multi-modal multicue object tracking

Reference	Algorithm	Tracking cues	Integration of cues	Database	Description
(A) Vision + thermal/IR					
Conaire et al. (2008)	Mean shift	YUV, thermal, edge orientation	Product of individual tracker likelihood score	OTCBVS, PETS, self	No adaptive fusion of cues
Treptow et al. 2005	Particle filter	Color, thermal	Handover from thermal to vision sensor	Self-generated	Could detect and track human far away from robot
Erdem et al. (2012)	Particle filter	Color, motion and Infrared brightness	Weighted product of individual cue likelihood	OSU,BEHAVE, CAVIAR	Two stage integration of cues, adaptive for particle assign
Airouche et al. (2012)	Particle filter	Color, location and thermal	DSmH hybrid rules	Self-generated single video	Solve clutter and occlusion
Wang et al. (2014)	Part based matching	Optical flow, color and depth	Linear combination of depth, color and pixel count	Self-generated using Kinect	Illumination invariant
Susperregi et al. (2013)	Particle filter	Leg, vest and thermal detection	Weighted sum of three individual likelihoods	Self-generated video	Independent detections were combined in particle filter
Han et al. (2012)	ID matching	Depth and color	Linear combination of individual similarity score	Self-generated	Tested on indoor environment, cues given equal importance
Xu et al. (2014)	Level set evolution	Depth and color	Local adaptive weighted map at super pixel level	Self-generated	Demonstrated for active contour tracking in outdoor
Motai et al. (2012)	Kalman filter and optical flow	Infrared, optical flow, laser	Arbitration between two modularity	Self- generated from robot	Infrared for object detection and laser for obstacle detection
Stolkin et al. (2012)	Mean shift	Infrared and color	Adaptive fusion at pixel level	Self-generated	Cue importance was estimated from background relearning
Talha and Stolkin (2014)	Particle filter	Color and infrared	Adaptive fusion of cues at particle levels	Self-generated	background relearning decide modularity weight

Table 2 continued

Reference	Algorithm	Tracking cues	Integration of cues	Database	Description
(B) Vision + audio					
Chen and Rui (2004)	Particle filter	Contour, sound source location, color	Product of individual likelihoods	Self-generated two video	Sensors data used for proposal estimation and verification
Strobel et al. (2001)	Decentralized Kalman filter	Audio and vision sensors	Decentralized Kalman filter but not adaptive	Nil	Fusion was applicable for independent measurements
Perez et al. (2004)	Particle filter	Color, motion and sound	Partitioned sampling with two stage integration	Self-generated	Main color cue was aided by either motion or sound cue
Hu et al. (2002)	Kalman filter	Color and audio	LMS method	Self-generated	Automatic calibration of audio and video sensors
Megherbi et al. (2005)	Kalman filter	Color, position and audio parameters	Dempster's rule, and pignistic probability	Self-generated	Multiple person tracking with focus on data association
Nickel et al. (2005)	Particle filter	Face, upper body and TDOA of audio	Weighted sum of individual likelihood	Self-generated video	Solution for Audio + Visual log creation of lecturer
Kılıç et al. (2015)	Particle filter	Color and TDOA from audio sensor array	Two stage particle weight estimation with two cues	AV 16.3	Calibration and active speaker presence assumed
(C) Vision + laser					
Chakravarty and Jarvis (2006)	Particle filter	Color and laser data	Measurement from two sensor used for detection	Self-generated	Depth info. handle occlusion, can track multiple person
Spinello et al. (2009)	Extended Kalman Filter	Color and laser data	Kalman filter fuse independent measurements	Self-generated	Brownian and linear motion model for tracking
Cui et al. (2008)	Mean shift and Kalman filter	Laser data and color	Bayesian fusion at tracker level	Self-generated	Two stage framework, detection and tracking
Song et al. (2013)	Particle filter	Laser data, Intensity and Edge orientation	Laser for tracking and visual for online learning	Self-generated	Online learning aided during correlated target and merger

Table 2 continued

Reference	Algorithm	Tracking cues	Integration of cues	Database	Description
Scheutz et al. (2004)	Local search	Distance and color	Leg tracker and face tracker perform independently	Self-generated single video	Failure during leg occlusion handled by vision tracker
Song et al. (2008)	PD-PF	Distance and HSV color	Product of individual likelihood map	Self-generated single video	Better estimation of MAP than multiple and joint particle filter
(D) Vision + RF					
Kim and Moongu (2014)	Kalman Filter	Radar and vision	Kalman filter along with MIPDA	Self-generated single video	Low computational for real time implementation
Germa et al. (2010)	Icondensation	RFID, skin and face	Product for individual measurements	Self-generated single video	Perception systems for online detection and avoidance ability
(E) Stereo vision					
Nickel and Stiefelhagen (2008)	Condensation	Motion, color, detector, stereo correlation	Weighted likelihood by cue quality measure	Self-generated	Combine layered sampling with democratic integration
Munoz-salinas et al. (2008)	Particle filter	Depth, color and gradient	Product of individual distribution	Self-generated video seq.	Joint probalistic distribution for color and depth
Gavrila and Munder (2007)	Hungarian Matrix	Stereo, shape, texture	Weighted comb.: Euclidian and chamfer distance	Self-generated via Protector	Proposed automatic tuning of parameter using ROC
Zoidi et al. (2014)	Kalman filter	Texture, color, disparity	Maximum average similarity for texture	Self-generated nine videos	3D object tracking in a stereo video

2.2 Discussion

In last decade, multicue object tracking using single-modal was extensively explored with the aim to develop robust and reliable tracking solution. The vision research under single-modal multicue target tracking was mainly focused towards development of new techniques for extracting robust features and also towards adaptive fusion of multicues. In general, deterministic methods were more accurate in state estimation than stochastic methods. However, deterministic methods performance degraded for lost target or heavily occluded target. On the other hand, stochastic methods can effectively fuse multiple features, and easily up-gradate for multiple objects tracking. Although, particle filter established its niche in estimation of non-Gaussian, non-linear state, it suffers from sample degeneracy and impoverishment problems for which solutions were proposed (Li et al. 2014a; Walia and Kapoor 2014b, c). However, most of reviewed papers using particle filter did not discuss problem of sample impoverishment. Only, orthogonal cues need to be fused so that they can complement each other during tracking process. The color cue being computational efficient could not provide shape information of target. For incorporating shape information in color model, spatial histogram was proposed (Wang and David 2006). The feature level fusion, histogram of height difference (HOHD) and joint height and color histogram (JHCH) were also introduced for augmentation of color cue (Liu et al. 2015). Color cue was also fused with spatial temporal motion energy for object tracking (Zhou et al. 2014). Similarly in IR sequences, inability of grey cue to provide shape information was compensated by extracting joint gray and local binary pattern histogram (Li et al. 2014b). The cue orthogonality was also established by dynamically selecting same feature from different region of object (Nickel and Stiefelhagen 2008). Hence, it needs to further investigate newly developed cues and establish their complementary nature for development of robust tracking solution. The integration of multicue was also considered imperative for adaptive multicue tracking system development. The democratic integration (Triesch and Malsburg 2001) can efficiently boost good cues and suppress non performing cues but suffer from false positives tracking. In order to estimate cues contribution for overall likelihood for state estimation, cues reliabilities were evaluated considering either previous observations or current observation (Brasnett et al. 2007). For real time implementation of multicue framework, cues resource allocation was decided based on their reliability (Loy et al. 2002). In addition, switching between different models based upon cues reliability was presented (Dou and Jianxun 2014). Hence, estimation of cue reliability and adaptive fusion of cues is another line of research that needs further research work. The infrared/thermal cameras are useful for tracking object for night vision applications. But single intensity cues needs to be compensated with other features extracted from captured sequences. In addition, thermal sensors are unable to distinguish objects with similar thermal profile. For low vision application, vision camera may be compensated with other modularity for robust tracking.

3 Multi-modal multicue object tracking

The multi-modal multicue tracking solutions are based upon two or more modularities from which different complementary cues were extracted. In general, vision camera was aided by other sensors such as laser, audio, radar, RFID, sonar etc. for catering different challenges of visual world. In this section we reviewed different multi-modal multicue object tracking methods in turn. The details of different reported work are also summarized (see Table 2). This section is concluded with discussion on reviewed techniques.

3.1 Vision + thermal/IR sensor

The thermal/IR sensor along with vision sensor is required for 24 × 7 targets tracking for night vision applications where vision sensor alone fails. The passive IR/thermal sensor can recover objects information during night or bad weather conditions but unable to distinguish objects with similar thermal profiles. Hence, vision camera along with thermal/IR camera was exploited for robust 24 × 7 object tracking. For tracking human face using mobile robot, switching to vision camera after initial detection of human from thermal sequence was proposed (Treptow et al. 2005). In particle filter framework, efficient detection of human from thermal sequences aided vision camera for tracking human face. RGB sensor along with IR sensor for efficiently tracking moving objects was proposed (Kumar et al. 2014). Under this template based tracking using RGB sensor, false negative and false positive were overcome through track level fusion of IR sensor data with RGB data. In addition, solution for temporal and spatial alignment of IR and RGB data was proposed. For catering sudden turn during motion, arbitration between optical flow and Kalman filter model was proposed (Motai et al. 2012). In addition, initial target was efficiently detected from infrared camera and obstacles were overcome using laser sensors. Proposed algorithm was tested on data captured from mobile robot with three mode of target walking. However, proposed methods considered switching from thermal to vision camera rather than adaptive fusion. In order to develop solution with low computational requirement, tracker level fusion for multiple features extracted from vision and thermal camera was presented (Conaire et al. 2008). In this work, spectrograms were used instead of histogram for feature extraction incorporating spatial information of bins. The final likelihood score was determined through product of individual likelihood score. Mean shift framework for estimating new position of object and similarity measure from different spatiograms was claimed as novel contribution in this direction. The algorithm was tested on five dataset using different combination of five features i.e. YUV, edge orientation and thermal brightness. However, no automatic adaptive integration of features was incorporated and also tracker was evaluated with limited metrics. Pixel level fusion of color and infrared modularity was considered (Stolkin et al. 2012). In mean shift algorithm, importance of cues was adaptively adjusted through relearning of background information in each frame. The proposed framework could be extended for new modularity. In extension of earlier work (Talha and Stolkin 2012), particle level fusion of color and infrared modularity was proposed (Talha and Stolkin 2014). In particle filter framework, color and thermal information was adaptively fused through continuous relearning of background model for each particle. This method is useful for tracking object under heavy occlusion and rapidly varying background. Context-sensitive integration of cues at two stages in particle filter framework was considered (Erdem et al. 2012). In this work, color, motion and infrared brightness cues were extracted from calibrated video recorded using vision and CCD camera. In each frame, cues reliabilities were determined considering their performance in current estimation and also past performance. The reliability of cue was used not only for estimation of final likelihood but also for assigning particles to proposal functions. The algorithm was tested on three data set and authors claimed that their framework was open for new cues. In order to adapt to target shape and align images from thermal and vision camera, geometric fusion for image alignment and two stage background models for accurate tracking was proposed (Zhao and Sen-ching 2014). As extension of earlier work (Sun and Bentabet 2010), multiple objects tracking with online conflict resolving, due to occlusion and clutter, from color and thermal sensor was proposed (Airouche et al. 2012). The author used position cue as prior knowledge and color and thermal cues as independent measurement from properly aligned sensors. The thermal cue detected presence of pedestrian and aided by color feature which distinguish among pedestrian. The cues were fused in DSmT hybrid model which was defined using hyper power set. The DSm hybrid rules were used for combining measurements from individual cues and conflicts were resolved among cues. However, the algorithm was tested on single video sequence and also computational requirement increased with increase in number of tracked object. The Microsoft Kinect camera was also extensively explored for extracting depth information from video (Han et al. 2013; Kinect Camera). In general, for video tracking consistency of depth information was assumed in consecutive frames. Two phase framework for robust object tracking which was invariant to illumination changes and occlusion was presented (Wang et al. 2014). In first phase, optical flow was estimated from two adjacent frames which gave the approximate position of object. Further, the position of object was estimated with more precision through patch based search approach using color and depth information. The object search was divided into four parts. The color histogram of four parts was compared with reference target template using Bhattacharyya distance. Similarly, variation in depth mean and variance between target region and individual parts was exploited for estimating smoothness of depth information in consecutive frames. The experiments were performed on four video sequences which were captured using Kinect source having inbuilt color camera, infrared projector and IR camera. Authors claimed that proposed algorithm was robust to illumination changes due to fusion of depth information. Human tracking framework for indoor home environment was proposed (Han et al. 2012). Under this, from RGB-D camera, human presence and identification was carried out using complementary depth and color cues extracted from RGB-D camera. The tracking was performed through matching of detected ID with previous assigned human ID using depth and color cues. The similarity score from two cues were fused as linear combination of individual score. However, no adaptive fusion was discussed and both cues were given equal importance for final state estimation. Framework for estimating importance of depth cues from adaptive weighted map was proposed (Xu et al. 2014). The depth was considered as prominent cues when object was far away from background object. On the other hand, when object is close to background object, color cue dominates over depth cue and contribute more towards final state estimation in tracking framework based upon level set method. In similar line, vision and depth cues from Kinect source for 3D object tracking using extended Kalman filter was considered (Gedik and Alatan 2013). Cues adaptive fusion was proposed through variation of measurement noise variance in accordance with quality measure of cue. Recently, a mobile robot tracking solution using RGB-D Kinect camera, laser and thermal sensor was proposed (Susperregi et al. 2013). In this work, person leg detection, Kinect vest detection and thermal detection likelihoods were combined in particle filter as weighted sum of individual likelihood. The results revealed that weighted combination of modularity's improved angle and location estimation of person wearing yellow vest. However, results were presented on indoor environment on limited set. In sum, thermal/IR camera aids vision camera for tracking applications. Still, data fusion and calibration of independent source of information remains challenge for vision community.

3.2 Vision + audio sensor

Multi-modal tracking using audio and vision sensors are considered effective means of tracking during distributed meeting, automatic scene analysis and effective presentation during lectures. During lectures, Multi-modal solutions using audio and vision sensors were used for automatic zooming of object for effective view of presenter (Polycom). The audio sensors based tracking solutions were focused on person localization and limited to speaking persons. The vision camera based solutions were limited to object in camera field of view. Hence, in recent years, researchers were motivated to fuse audio sensors with vision sensors for robust object localization and tracking. The problem of fusing data from audio and video sensors was addressed (Strobel et al. 2001). The authors introduced decentralized Kalman filter for fusing data from different modularities. The audio sensors estimated posterior range and bearing information of target using extended Kalman filter. The Cartesian coordinates were extracted from video sensor and posterior estimation was performed using linear Kalman filter. The posterior estimates from independent sensors were given to fusion centre which determine global estimate of object. However, assumption of linear motion needs further investigation for manoeuvring objects. Using belief theory, a solution for data association between audio and vision sensor was discussed (Megherbi et al. 2005). The frame of discernment was assumed to be set of targets in the scene. For each target, three masses were determined from similarity measure in line with target association to target, not association to target or ignorance. Similarity measures were estimated using position and video or audio measurements separately. The masses were further combined using Dempster's rules and Pignistic probability was estimated for making final decision about the target. Calibration of audio and vision sensors was considered a prior and work was mainly focused on data association. In an attempt to automatically calibrate audio and vision sensors, multicue tracking system using two microphones and vision camera was proposed (Hu et al. 2002). In this work, object was localized using microphone array where fractional delay was determined for delay in arrival among microphone array. The direction of arrival estimated using vision camera was feedback to microphone array for necessary correction of phase distortion. The solution can automatically calibrate sensors without requirement of prior audio training samples. In order to increase localization accuracy, triangular array of three microphones along with vision camera was proposed for object tracking (Lim et al. 2007). The time delay in arrival of sound was used for estimation of sound likelihoods and standard deviation between microphones was changed dynamically. The overall likelihood was simple product of color and sound likelihood. Considering complexity of visual world and inherent property of particle filter to fuse different independent sources, framework where color cue acts as main cue and aided by intermittent motion cue or sound cue depending upon application scenario was presented (Perez et al. 2004). Under this, histograms were extracted for color and motion cues from video sequences. Also time difference of arrival (TDOA) of audio signals from pair of microphones placed on line orthogonal to optical centre was exploited for estimating bearing of object. The generalized cross correlation function used for estimating TDOA and likelihood was determined using multiple hypotheses. The independent likelihood were determined and used in partitioned sampling framework where motion or sound cues with mixture proposal distribution estimate the approximate search region. The estimated search region aided color cue for efficient estimation of state with minimum number of particles. Two stage frameworks for tracking object using audio, contour and color cues was proposed (Chen and Rui 2004). In proposed particle filter based framework, each sensor consisted of tracker and verifier where tracker estimated proposal for fuser and verifier determined likelihood score for individual sensor. The contour tracker exploited contour smoothness and based upon unscented Kalman filter (UKF). The color tracker assumed stability of object color in interior region and used mean shift algorithm for proposal estimation. Similarly, audio tracker exploited delay in arrival at two microphones for estimating proposal. These proposals were given to the fusers which generated combined proposal distribution considering tracker reliability and sent sampled particles to verifier. The three verifiers estimated likelihood for each cue and determined final likelihood score which was given to fuser for final state estimation. The approach was adaptive with feedback mechanism from fuser output to tracker input. Similarly, multiple audio and vision sensors under reverberant environment was used (Talantzis et al. 2008). The authors proposed independent tracker for audio and vision sensors and fused information for building robust solution. The vision sensor aided audio tracker for tracking person who stopped talking during tracking process. However, results were presented on limited set of data and performance for outdoor environment needs to be investigated for real time application. In order to create efficient log of lecture presentation, a solution using multiple video cameras along with multiple microphone for locating presenter during his/her presentation was proposed (Nickel et al. 2005). In this work, 3D projection of image from multi view was performed using particle filter instead of tribulation. The visual information was extracted using foreground segmentation and detectors for face, upper body. Similarly, from pair of microphones, TDOA was exploited for estimating audio observation likelihood. In particle filter, final likelihood was weighted sum of likelihoods where weights were adjusted in accordance to audio cue performance. However, multiple persons tracking along with adaptive selection of parameters were not addressed. In order to track multiple persons in indoor environments, audio sensors measurements at two stages in particle filter i.e. propagation and observation was considered (Kılıç et al. 2015). In addition, number of particles and noise variance were made adaptive depending upon tracker performance. The authors demonstrated that combining vision data with circular array of audio sensors measurements enhanced tracking performance over vision sensor based tracking for occluded active speaker. However, calibration of audio sensors and presence of active speaker was assumed for tracker proposed. In sum, audio sensors aided vision sensor for developing robust tracking solution, but most of studies were limited to coherent audio sources. Effect of coherent and non-coherent audio sources on image feature response was investigated and studies revealed that non-coherent sources degraded image features response (Chen et al. 2014). Hence, considering real world tracking scenarios, performance of tracking algorithms needs to be evaluated for both coherent and non-coherent audio sources.

3.3 Vision + laser sensor

The high scanning rate laser scanners along with vision camera are used for target detection and tracking for video surveillance applications. The laser scanner has potential advantages such as low computational requirement, easy projection of laser data to rectangular coordinate system and insensitive to environment changes. But, it could not distinguish objects during merger/split, correlated objects or object with lost track under heavy cluttered environment (Cui et al. 2008; Song et al. 2013). On the other hand, vision data was complex, illumination dependent but rich in information and easily distinguished objects. Hence, in last decade, reliable and robust solutions was proposed for tracking people using laser and vision sensors. For tracking indoor people, multi-modal framework where independent laser and vision tracker perform tracking under constrained environment was proposed (Scheutz et al. 2004). In this work, person face tracking was performed on vision data using face detection module which confirmed presence of face using eye detection, horizontal projection and height to width ratio. The leg tracking was performed considering gap between two legs and if not detected properly information was passed to vision sensor for further processing. The mutual sharing of information between two trackers overcame failure of laser tracker under occlusion. Using stationary robot, fusion of data from panoramic vision camera and laser sensor for multiple people tracking was discussed (Chakravarty and Jarvis 2006). In this work, Laser sensors were focused on upper part of body rather than legs. If person was detected by vision and laser sensor, particle filter based tracker was initialized for detected person. However, results were presented on limited dataset. For robust tracking of people in outdoor environment, multiple laser scanners for fast detection of people along with single vision camera was considered (Cui et al. 2008). In this work, once people were detected by laser scanners, people tracking was performed using two independent trackers (a) mean shift algorithm for vision sensor using color histogram, (b) Kalman filter for laser sensors using people walking model. Also, authors recognized difficulty in fusing laser sensor with vision sensor at data and feature level and proposed decision level fusion of independent tracking results. However, swing feed model for people may change during tracking period. In addition, tracking of objects such as car, bus where swing was not available needs more research. For tracking multi class objects such as car, bus, person in outdoor environment, independent object detection in laser range scanner and vision camera and use of multiple motion models was proposed (Spinello et al. 2009). In this work, laser range scanner provided structural information of object using conditional random field where node features were enhanced using Adaboost. The object appearance features were extracted using modified implicit shape model. The tracking was performed using extended Kalman filter using Brownian and linear motion model in order to cater multi class object. Rather than placing laser scanner at feet height, laser scanner was placed at top of tripod and laser scanner based detection of human body was considered (Song et al. 2008). Once, vision tracker was initialized with laser detection center, HSV color histogram was extracted around body region and used for estimation of color likelihood in proposed probabilistic detection based particle filter (PD-PF). The individual likelihood score were determined from measurement and over all observation was determined through product of individual likelihood. Results revealed that occlusion and interaction among human could be handled by incorporating weighted detection information into proposal distribution. A probabilistic exemplar models for object appearance were trained independently from laser sensor and vision camera (Schulz 2006). During tracking, joint exemplar states were sampled using Rao-Blackwellized particle filter for estimation of state of object. However, approach used object contour feature which was affected by environment changes and occlusion. In order to handle object during merger/split or close interaction, a solution where tracking was performed along with online learning using vision and laser scanner was proposed (Song et al. 2013). During normal operation (non-correlated), independent laser scanner performed tracking using particle filter and vision data was used for online training of regression tree classifier. The intensity and edge orientated histograms were extracted from randomly located patches around laser tracker estimated center and used as features for training classifier. Once target merger/split was detected visual information aided laser tracker for classifying targets into respective class during split into non correlated objects. Also during correlated object detection, random patches were sampled around interacting region and subjected to respective classifier for finding score map being used as observation model for particle filter. The results were demonstrated on single video for tracking multiple interacting objects. However, rotation and scaling of object along with effect of environments on vision camera needs to be further explored.

3.4 Vision + radio frequency

The radio frequency based sensors provide accurate identification information about object but needs to be compensated by other sensors for robust tracking solution. Recently, due to availability of low cost radar, radar sensors were used for tracking object for video surveillance, robots and driver assistance. The radar signal provides radial information of object along with bearing that can be exploited for locating object on image plane. The radar signals were less affected by noises such as fog, dust, snow, wind etc. and can detect trapped objects. For tracking multiple objects, data association from low cost radar and vision camera was proposed (Kim and Moongu 2014). In this work, radar giving low resolution bearing information was compensated by vision camera for multiple object tracking. The measurements from independent sensors were projected on common plane using homography which mainly focus on calibration of two sensors. The tracks management and clutter handling was performed using multiple object integrated probabilistic data association (MIPDA). The tracking of multiple objects was performed using multiple Kalman filters where data fusion and tracks updation was performed. Results revealed that combining radial information with complementary image information enhanced performance over single modularity framework. Data association problem in vehicular networks for awareness of situation to driver or vehicle was discussed (Thomaidis et al. 2013). In this work, long range radar signals captured from first vehicle were fused with measurements such as yaw rate, position, velocity and acceleration from second vehicle. Once association between radar and VANET message was established, global tracker using multiple hypothesis algorithms was initialized for estimating further states. The authors claimed that fusing additional measurement to radar based perception system enhanced tracking accuracy and robustness. A perception system based upon 8-Radio Frequency ID (RFID) antennas and vision camera for accurate identification and tracking of person with passive tag was proposed (Germa et al. 2010). The vision cues, skin color and face, were fused with RF identification at measurement and important sampling in Icondensation algorithm. The results were presented on single video sequences and revealed that obstacle avoidance performance with proposed robot control scheme was better than single modularity. However, performance needs to be investigated for multiple persons carrying RFID tags in a crowd.

3.5 Stereo vision

Due to its ability to capture object gesture and illumination invariant, stereo vision is extensively explored for 3D object tracking in the real world scenario. The layered sampling (Perez et al. 2004) and democratic integration (Triesch and Malsburg 2001) were combined in single framework using condensation algorithm (Nickel and Stiefelhagen 2008). In this work, independent trackers were initialized for each targeted object using four cues such as color, motion, detector and stereo correlation. Apart from selection of orthogonal cues, authors proposed orthogonality by dynamically selecting same feature from different region of object. First, disparity map was extracted for stereo cue which gave rough approximate location of target. The estimated position was further refined using other cues for robust tracking in two stage. The cues were adaptively fused using proposed quality measure based upon position error. However, experimental results were briefly discussed on single self-generated video sequences. For mobile platform, 3D multiple people tracking without need of background model was proposed (Munoz-salinas et al. 2008). In this work, from stereo vision camera, depth information was extracted and its contribution in particle weight assignment was decided based upon amount of disparity information. First, Adaboost classifier along with depth information was used for efficient detection of people in the scene. The independent particle filter tracker was initialized for each detected person and interaction among them was determined for handling occlusion due to similar color person. Two ellipses at torso and head were considered for features extraction. The author proposed estimation of joint color and depth probability distribution for two portions where standard deviation of depth distribution was modified in accordance of confidence of disparity measure. Further, gradient probability distribution was extracted in order to estimation of matching of ellipsoidal object. The final observation was product of three probability distributions without considering adaptive contribution. In similar lien, disparity information from stereo video for 3D object tracking was exploited (Zoidi et al. 2014). The targeted object was first passed through Kalman filter for new state prediction. The search region was established around new predicted centre. In order to reduce search space, sub sampling was performed by measuring similarity of 2D color disparity histogram of candidate ROI with that of reference ROI. From reduced search space, texture cue was extracted and similarity index was calculated for both left and right channel of video by comparing candidate texture with that of reference model. The final position of object was determined based upon maximum of average similarity determined using cosine similarity. The results revealed that stereo tracker perform better than monocular tracker for slowly deforming articulated objects. In order to automatic tuning of various tracking and detection parameters, receiver operator characteristics (ROC) usage for close interaction among different detection modules was discussed (Gavrila and Munder 2007). In this work, pedestrian detection was performed in various stages such as stereo for coarse search region generation, shape based detection followed by classification by texture cue and finally verification by stereo module. The tracking of pedestrian was performed using Hungarian method where cost matrix was extracted using two similarity measures as Euclidian and chamfer distance. The results were presented using proposed mobile protector system.

3.6 Discussion

In recent past, in order to exploit potential advantages of different sensing technology, multi-modal multicue object tracking has been extensively explored. In general, vision camera was aided by one of sensing technology for development of robust multi-modal tracking solutions. Being invariant to illumination changes, insensitive to pose variation, thermal/IR cameras could detect object for night vision applications and aided vision camera for continuous video surveillances. The fusion and synchronization of vision and thermal camera was major area of research which was addressed differently for robust tracking solution. The tracker level fusion of vision and thermal camera was addressed (Conaire et al. 2008). The fusion based on context sensitive reliability was discussed (Erdem et al. 2012). Cues from color and thermal cameras were fused using DSmT framework for resolving conflict from independent measurements, but, model could not handle tracking of single object (Airouche et al. 2012). For tracking single object using vision and thermal camera, DSmT based framework was proposed considering conflict resolving using proportional conflict redistribution rules, PCR-5 rules (Walia and Kapoor 2015). The Microsoft Kinect camera having synchronized infrared and vision camera could provide depth information used for handling occlusion (Susperregi et al. 2013). The audio sensors can locate person through efficiently extraction of range and bearing information of target, but, fail to locate silent person/object. Hence, combining audio sensors with vision sensor found its niche for applications such as distributed meeting, automatic scene analysis and effective presentation during lectures. Being insensitive to environment changes and computationally efficient, laser scanners combined with vision camera for handling occlusion, and merge/split for multicue tracking framework. The laser scanner were placed either above upper body or at legs positions for extracting different patterns of objects. The patterns such as walking pattern (Scheutz et al. 2004), swing model (Cui et al. 2008), motion models (Spinello et al. 2009) were extracted using laser scanner and combined with vision camera for distinguishing objects. In addition, detection of whole body using laser scanner was considered (Song et al. 2008) which aided vision camera for robust tracking solution. Due to its ability to locate hidden object, RF sensing such as radar and passive RF tags along with vision camera were explored for multi-modal tracking solution. However, tracking multiple persons with similar passive RFID tag needs further investigated. The stereo camera being able to provide disparity information about target was exploited extensively for 3D multicue object tracking. The depth information extracted from disparity map was combined with other cues for multicue tracking. In sum, similar to single-modal, multi-modal tracking work was mainly focused towards either extraction of efficient features or fusion of different sensors data in efficient manner. In addition, calibration and synchronization of different sensors is another domain which needs further research for accurate tracking solution development. Our analyses of single-modal and multi-modal multicue techniques revealed that technology in this field has matured enough for real applications of these solutions.

4 Datasets and evaluation metrics for multicue object tracking

The object tracking is challenging task due to complex nature of real world and number of environments scenarios. In each year, in addition to work reported by different authors, Performance Evaluation of Tracking and Surveillance (PETS) workshop introduced new datasets and performance measures for evaluation of tracking algorithms. On the contrary, most of reported methods were evaluated on limited dataset for specific performance measures. Hence, in order to give reader a quick overview and new trends and also for sake of completion of multicue tracking survey, we have briefly summarized various reported multicue datasets and performance measures for object tracking.

4.1 Datasets for evaluation of multicue object tracking techniques

For evaluation of tracking algorithms, standard datasets are available publically. Some of these datasets provide ground truth data in the form of separate *.xml file. In addition, algorithms are tested on synthesis video sequences for catering different test scenarios. The overview of various datasets for object tracking is summarized in Table 3 and also sample image of each listed set is included in Fig. 4. The brief description of multicue datasets follows in turn.

(a) TUD-Stadmittee dataset sequences contained 178 images with resolution of 640 × 480 (Andriluka et al. 2010). The data was captured using stationary monocular camera. The dataset had multiple persons being self or fully occluded by different objects. The ground truth data was available which consist of different object ID along with corresponding coordinates in different frames.

(b) PETS 2014 workshop introduced ARENA dataset which had 22 scenarios categorized as 'something is wrong' 6 video sequences, 'potentially criminal' 6 video sequences, 'criminal behaviour' 7 video sequences and 'extra criminal scenarios' 3 video sequences (Patino and Ferryman 2014; PETS 2014). The data set was collected at university using four cameras mounted on vehicle. The aim was to provide benchmark data set for evaluation of algorithms for behaviour and abnormality analysis in a scene captured using multiple cameras.

(c) Head tracking sequences (HTS) included various challenges such as occlusion, rotation, zooming and distractions. The video sequences along with ground truth data representing coordinate of ellipse position were available at HTS. The resolution of bmp images are 128 × 96 pixels.

d) Audio visual people (AVP) dataset consisting of three vision sequences along with two microphone outputs was available online (MOTINAS project). The dataset consist of 3271 frames of 360 × 288 pixels size at 25 fps. The two audio channels were at 44.1 kHz. The data was recorded in closed rooms having reverberations. The ground truth data was also provided in the form of *.xml file. In addition, AV16.3 dataset consisted of 8 annotated sequences captured indoor with three cameras and 2 arrays (8 each) of microphones (Lathoud et al. 2004).

Table 3 Recent advances on datasets for object tracking

Data set	Web address	GT	Frame size, Number of videos, rate etc.	Description
(A) Vision camera dataset				
Andriluka et al. (2010)	https://www.d2.mpiinf.mpg.de/node/428#name:cvpr10_videos	Yes	640 × 480, 178 frames	Multi persons with occlusion, captured from stationary camera
PETS (2014)	Pets2014.net	No	1280 × 960, 30 fps, 22 video sequences	Multi Camera capturing of scene classified into four levels
HTS	http://www.ces.clemson.edu/_stb/research/headtracker/seq/	No	128 × 64, 17 video sequences	Included Occlusion, Rotation, Zooming and Distractions
Blunsden and Fisher (2010)	http://groups.inf.ed.ac.uk/vision/BEHAVEDATA/INTERACTIONS/	Yes	640 × 480, 25 fps, 90000 frames with 4 video	Interaction of people in 10 types of group behaviour with limited GT
EC-CAVIAR	http://groups.inf.ed.ac.uk/vision/CAVIAR/CAVIARDATA1/	Yes	384 × 288, 25 fps, 54 video sequences in two sets	Included scaling, occlusion, illumination changes and interactions
Klein et al. (2010)	http://www.iai.uni-bonn.de/~kleind/tracking/index.htm	Yes	320 × 240, 25 fps, 5 human video sequences in mpeg2 format	Included rotation, occlusion and captured using moving camera
Fleuret et al. (2008)	http://cvlab.epfl.ch/data/pom	Yes	5 video sequences captured using multiple cameras, 25 fps	Multiple people with partial or full occlusion, Scaling indoor/ outdoor environment
PETS (2006a, b)	http://www.cvg.reading.ac.uk/PETS2006/data.html	Yes	7 video sequences, 768 × 576 pixels, 25 fps and JPEG.	Captured using four cameras, different difficulty level with multiple persons
Ess et al. (2007)	http://www.vision.ee.ethz.ch/~aess/iccv2007/	Yes	Two video sequences with 804 frames, 640 × 480 at 13–15 fps	Mobile chariots with two AVT Marlins camera mounted on it
Baltieri et al. (2011)	http://imagelab.ing.unimore.it/visor/3dpes.asp	Yes	Resolution 704 × 576, 8 cameras sequences, 15 fps, JPEG image	Large sequences provided in six parts included person captured using multi cameras
(B) Thermal/IR camera dataset				
Davis and Keck (2005)	http://www.vcipl.okstate.edu/otcbvs/bench/Data/01/download.html	Yes	360 × 240, 10 sequences 284 frames, 8 bit grey scale	Video under different environments such as cloudy, rainy, fair etc, with occlusion
Zheng et al. (2014)	http://csr.bu.edu/BU-TIV/BUTIV.html	Yes	1024 × 640 and 512 × 512, 3 pedestrian sequences with 16 bit	Multiple camera used for capturing multi view for real world scenario

Table 3 continued

Data set	Web address	GT	Frame size, Number of videos, rate etc.	Description
Portmann et al. (2014)	http://projects.asl.ethz.ch/datasets/doku.php?id=ir:iricra2014	Yes	324 × 256, 9 sequences at 20 Hz rate	Handheld camera from elevated position to capture different scenarios
(C) Thermal/IR + vision camera dataset				
Torabi et al. (2012)	http://www.polymtl.ca/litiv/en/vid/	Yes	320 × 240, 9 sequences 6,236 frames upto 88 sec	Captured with vision (Sony XCD-710CR) and thermal (FLIR T A40M) camera
INO's video	http://www.ino.ca/en/examples/video-analytics-dataset/	Yes	Variable resolution for 13 video sequences	Sequences with variable challenges captured under different weather
Davis and Sharma (2007)	http://www.vcipl.okstate.edu/otcbvs/bench/	No	340 × 240, 6 sequences at 30 fps	Having occlusion, shadow and number of objects to be targeted
(D) Vision camera + audio sensor dataset				
MOTINAS project	ftp://motinas.elec.qmul.ac.uk/pub/av_people	Yes	3271 frames of 360 × 288 pixels size at 25 fps	One vision and two microphones were placed in rooms with reverberation
Lathoud et al. (2004)	http://www.idiap.ch/dataset/av16-3	Yes	Avi video 288 × 360, 25 fps, audio at 16 khz, 8 annotated sequences	One sequences consisted of three video and 16 audio files
(E) Vision camera + laser scanner dataset				
Strigel et al. (2014)	http://www.uniulm.de/in/mrm/forschung/datensaetze.html	Yes	Monochrome 656 × 494, 25 fps, laser 12.5 Hz	Three sequence for public interaction with object label and reference data
(F) Stereo vision camera				
Ess et al. (2007)	http://www.vision.ee.ethz.ch/~aess/iccv2007/	Yes	Three video sequences, 640 × 480, 10–15 fps	Pair of AVT Marlins used for recording in platform, calibration data is given
(G) Kinect sensor				
Klein	http://www.iai.unibonn.de/~martin/tracking.html	Yes	Five video sequences, 640 × 480, RGB 8 bit and depth 16 bit	Contains rotation, occlusions, illumination changes
Munaro et al. (2012)	http://www.dei.unipd.it/~munaro/KTP-dataset.html	Yes	640 × 480 resolutions, 30 Hz and in total 8475 frames in 4 Seq.	2D and 3D ground truth data, five people in each sequence

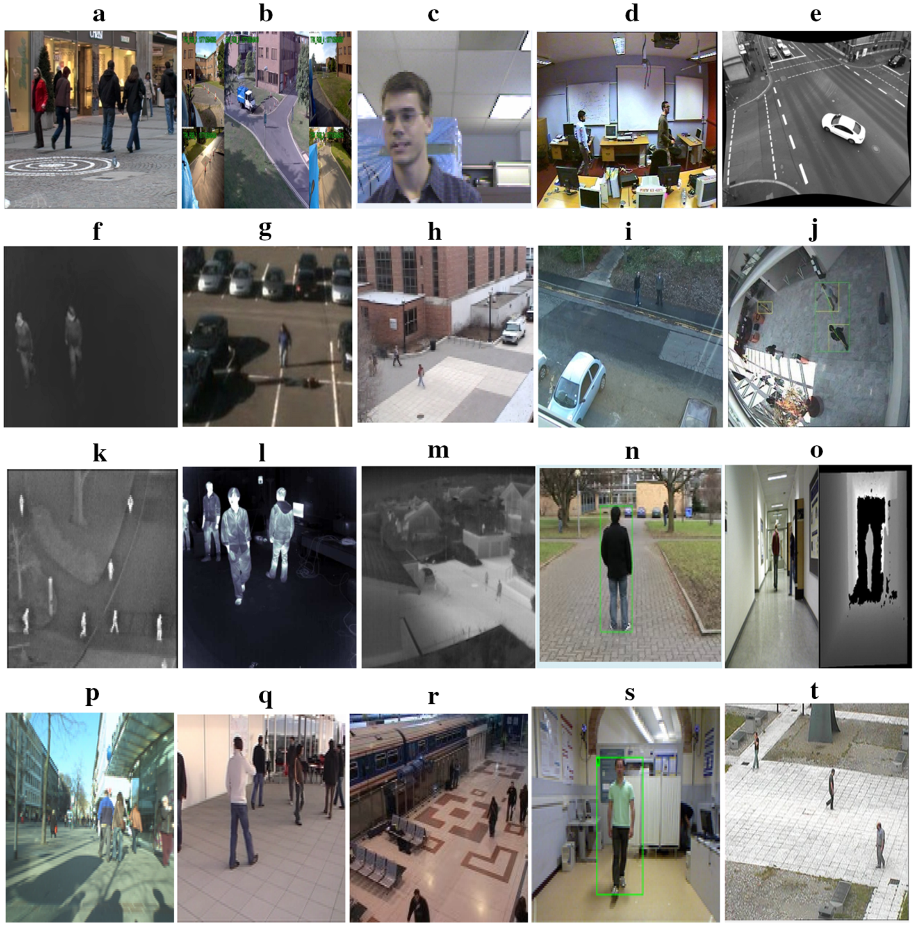

Fig. 4 Sample image of datasets: **a** TUD-Stadmittee (Andriluka et al. 2010), **b** ARENA (PETS 2014), **c** HTS (HTS), **d** AVP (MOTINAS project), **e** Ko-PER (Strigel et al. 2014), **f** LITIV (Torabi et al. 2012), **g** INO (INO's Video), **h** OSU (Davis and Sharma 2007), **i** BEHAVE (Blunsden and Fisher 2010), **j** CAVIAR (EC CAVIAR), **k** OSU (Davis and Keck 2005), **l** TIV (Zheng et al. 2014), **m** ASL-TID (Portmann et al. 2014), **n** BoBoT (Klein et al. 2010), **o** BoBoT-D (Klein), **p** MSA (Ess et al. 2007), **q** EPFL (Fleuret et al. 2008), **r** PETS2006 (PETS 2006a, 2006b), **s** KTP (Munaro et al. 2012), **t** 3DPes (Baltieri et al. 2011)

(e) Under Ko-FAS, Ko-PER road interaction dataset was generated as benchmark dataset for evaluation of algorithms (Strigel et al. 2014). The dataset consist of three video sequences captured using eight monochrome cameras TXG-04 and laser scanner 14 SICK LD–MRS installed at height of 5m above ground level. The sequence 1 included object label and sequences 2 and 3 had reference data which was useful for evaluation of multiple object detection and tracking algorithms.

(f) LITIV Dataset included nine video sequences of visible and thermal video sequences captured under different conditions such zooming, position and occlusion (Torabi et al. 2012). The sequences were having resolutions of 320 × 240, variable frame rate and length ranging from 11 to 88 s. The ground truth data were provided separately for these sequences.

(g) INO's video analytics dataset consisted of video sequences captured using VIRxCam platform installed in outdoor with different weather conditions (INO's Video). The sequences with different resolution and duration were captured from calibrated thermal and color camera. The ground truth data was provided in form of images with extracted foreground.

(h) OSU Color-Thermal Database was created in three different locations at Ohio State University campus (Davis and Sharma 2007). The dataset consisted of 6 video sequences with 17,089 images captured from thermal and vision camera located on tripod above ground level. The sequences were captured in outdoor with different level of challenges. The tracking results were available in the form of *.cvml file.

(i) BEHAVE Interactions dataset (Blunsden and Fisher 2010) consisted of 4 video sequences captured at two different sites. The sequences capture 10 interaction behavior as in group, ignore, chase, fight, walk together, run together, approach, split, following, and meet. However, ground truth data for few video sequence was available online in the form of *.xml file. In totality, videos had approximately 90,000 frames with resolution of 640 × 480 at 25 fps.

(j) CAVIAR dataset consisted of 2 sets of video sequences captured at resolution of 384 × 288 with 25 fps. Set1: contained 28 video sequences representing six scenarios at INRIA lab. Set2: contained 26 video sequences, each with two views, representing 26 scenarios at shopping center Lisbon (EC CAVIAR). The sizes of videos were upto 21MB with MPEG2 compression. The ground truth data was available in the form of *.xml file.

(k) OSU dataset consisted of 10 thermal video sequences, 284 frames, captured at Ohio State University campus (Davis and Keck 2005). The ground truth (GT) data was provided for each sequence. The sequences were captured under different environmental conditions such as cloudy, rainy, fair etc.

(l) TIV dataset included three pedestrian thermal infrared (IR) video sequences at high resolution of 1024 × 640 and 512 × 512 (Zheng et al. 2014). The data was collected at variable frame rate with multiple cameras. The annotation of each sequence was provided along with Matlab code for reading data file.

(m) ASL-TID dataset had nine sequences captured at different outdoor locations (Portmann et al. 2014). The sequences were captured at fixed rate 20 Hz with resolution 324 × 256 with handheld FLIR Tau320 camera from elevated location.

(n) BoBoT benchmark dataset contained nine video sequences out of which five had presence of human (Klein et al. 2010). The sequences were captured at resolution of 320 × 240, 25 frame rate in mpeg2 format. The camera was in motion and scene had different video challenges such as distractions, occlusion, and rotation.

(o) BoBoT-D dataset included five RGB and depth sequences marked as 'Milk', 'Ball', 'Person', 'Tank', and 'Lunch Box'. The sequences were captured using Kinect sensor with color 3 channel 8 bit each and depth one channel 16 bit (Klein). The ground truth data is also available online.

(p) Mobile scene analysis (MSA) dataset consisted of two sequences, 804 frames, and recorded using pair of cameras mounted on mobile chariots (Ess et al. 2007). The images with resolution 640 × 480 were captured at rate 13–14 fps using AVT Marlins. The annotation along with calibration data was included for testing.

(q) Multiple people tracking EPFL dataset was created using multiple cameras (Fleuret et al. 2008). The dataset consisted of five video sequences captured at 25 fps with 2–4 camera mounted at height of 2 m above ground. The various challenges such as bad lighting, occlusion, scaling had been incorporated in these sequences. The ground truth data was provided for three video sequences.

(r) PETS series contained different sets of dataset depending upon scenarios and surveillance application domains. PETS2006 (PETS 2006a, 2006b) dataset consisted of 7 (S1–S7) video sequences captured using four cameras. The sequences had resolution of 768 × 576 pixels, 25 fps and JPEG compression. The different levels of difficulty were marked for each sequence and ground truth along with calibration data was provided.

(s) Kinect tracking precision (KTP) dataset consisted of 4 sequences captured at different position of mobile robot carrying Microsoft Kinect camera (Munaro et al. 2012). The images had 640 × 480 pixel resolutions, 30 Hz and in total 8475 frames. The 2D and 3D ground truth data was provided.

(t) 3DPes dataset contained video sequences in six parts with multiple cameras for surveillance (Baltieri et al. 2011). The eight cameras were installed with different zooming at different locations at University of Modena and Reggio Emilia. The resolution of cameras was fixed at 704 × 576 pixels with 15 frames/s. The sequences captured objects multiple times in one or more cameras.

In addition, there are datasets for which either public access is not present or cited in a few of multicue research papers. Additional details about datasets are available (CANTATA). In literature, it was well acknowledged that publically available datasets were preferred over self-generated datasets. However, for using public domain dataset, individuals are directed to read the license agreement/download instruction/reference citation for acknowledgement of sources. In addition to selection of datasets, performance measures are considered as critical parameters for evaluation of tracking algorithms. The brief introduction of various performance measures for multicue object tracking is discussed in following section.

4.2 Evaluation of multicue object tracking techniques

The tracker performance needs to be evaluated both quantitatively and qualitatively over a length of video sequences. In literature, tracking performance measures considered different aspects of video with aim to gauge robustness and efficiency of proposed technique. The aim of tracking performance measures is to estimate either deviation from track or false positive/false negative during tracking. In our review, we have categorized quantitative performance measures for single and multiple objects tracking depending upon availability of ground truth (GT) data. The brief introduction of various performance measures are examined in this section in turn:

4.2.1 Performance measure without using ground truth data

Tracking performance measure without using GT data exploits abnormality in regular pattern of motion, appearance, shape etc. (Spampinato et al. 2012; Chau et al. 2009; Erdem et al. 2001b) or uses some prior knowledge of object trajectory (Wu et al. 2010). These measures are easily computed on real time but prone to errors under different tracking scenarios. Performance measure for evaluation of video segmentation and tracking without GT data was proposed (Erdem et al. 2001b). The spatial difference extracted from color, motion difference along boundary and temporal difference extracted from color histogram difference were used for deducing evaluation metrics. The color metrics were determined with the assumptions that intra frame color boundary coincides with object boundary and also that inter frame color histogram of object did not change. The motion metric was also determined with the assumption that motion vector of object does not coincide with background motion vector at object boundary. The individual metrics were combined with weighted contribution to get final score in the range [0, 1]. As an extension of previous work, feedback mechanism for adjusting the weights assigned to features was considered (Erdem et al. 2001a). However, proposed technique was investigated for contour tracking. Similarly, appearance and motion smoothness for extracting performance measure scores was assumed (Wu and Zheng 2004). In this work, five features such as trajectory complexity, scale, motion smoothness, shape, and appearance similarity were considered to be consistent/smooth during tracking process. The individual scores were weighted to get final performance evaluation score. In order to make performance measure more adaptive and robust to variation in object scale and shape during tracking, seven features such as temporal length, exit zone, shape ratio, area, color along with their feature score at each frame were extracted (Chau et al. 2009). The final performance score of tracker was determined through weighted contribution of individual feature. The most of tracking performance metrics without GT were either application specific or algorithm specific. In order to test performance irrespective of tracker algorithms, physical motion reversibility of targeted object was exploited (Wu et al. 2010). The association between prior distribution at $t = 0$ and posterior distribution of time reversed Markov chain was used for evaluating performance of tracking algorithm. Also, fast approximation at different time interval was suggested for real time performance evaluation. Automatic exclusion of temporal state when target was lost and finally estimation of tracker performance when target was present was proposed (SanMiguel et al. 2012). In this work, tracker accuracy was measured by finding similarity between estimated target positions and time reversed tracker positions at each state. In similar lien, online evaluation of tracking algorithms using shape, appearance and motion features and Bayesian classifier was considered (Spampinato et al. 2012). In this work, Bayesian classifier classified tested frame features into good or bad tracking. The classifier was trained through features set extracted from training dataset. The final matching score was estimated from all features and used for deciding tracker performance metrics. In sum, performance measure without ground truth could evaluate tracker on real time during tracking but most of performance measures gave less reliable results due to variation in features with environment.

4.2.2 Performance measures using ground truth (GT) data

The tracker performance measures using GT are evolved with the start of Performance Evaluation of Tracking and Surveillance (PETS) workshop. These measures gave more precise results and also independent of tracker algorithms. We have further categorized these metrics based upon how the ground truth data was used for comparison with tracker output.

First in trajectory based approach, ground truth trajectory was compared with estimated tracker trajectory. Apart from complete trajectory comparison, comparison of some discrete events in trajectory in order to reduce the effort for ground truth extraction was proposed (Pingali and Segen 1996). In similar line, establishment of correspondence between ground truth track and tracker estimated track before estimation of performance metrics was introduced (Senior et al. 2001). In this work, association of ground truth track with estimated track was established through threshold of $N_g \times N_t$ distance matrix. Where, N_g was count of ground truth tracks and N_t was count of estimated tracks. Distance matrix $D_T(K_1,\ K_2)$ between two tracks K_1 and K_2 was determined using Eqs. (1)–(3).

$$D_T(K_1, K_2) = \frac{1}{N_{12}^2} \left\{ \sum_{\substack{i:\,\exists K_1(t_i) \\ \&\,\exists K_2(t_i)}} \sqrt{p_x^2(i)} + p_v^2\,(i) \right\} \tag{1}$$

$$p_x(i) = |X_1(i) - X_2(i)| \tag{2}$$

$$p_v(i) = |V_1(i) - V_2(i)| \tag{3}$$

where N_{12} is total number of states in two tracks, p_x (i) and p_v (i) were difference in centroid and velocity of object at time instant t_i. Using estimated $N_g \times N_t$ distance matrix, association between ground truth track and estimated tracks was established. Further, performance metrics for object tracking such as position error, area error, detection lab, incompleteness factor and track error rates were determined. Apart from track matching using association criteria, counting index was used and performance metrics such as object tracking accuracy, occlusion success rate, tracker detection rate, and tracking success rate, false alarm rate, track detection rate were considered (Black et al. 2003). Issues of spatial and temporal fragmentation/merging during object tracking were discussed (Brown et al. 2005). In this work, two stage evaluation of performance was proposed. Apart from measuring track false positive and negative, authors introduced metrics for measurement of fragmentation and merger errors. Temporal correspondence between GT track and tracker estimated track was measured (Yin et al. 2007) using Eq. (4).

$$\frac{L(GT_m \cap TT_n)}{L(GT_m \cup TT_n)} > \text{Threshold} \tag{4}$$

Under this work, function L () finds number of frames between GT track (GT_m) and target track (TT_n). When the above ratio was greater than threshold value, target track was associated with ground truth track. After establishment of association, various performance metrics such as track detection failure (TDF) track fragmentation (TF), latency of system track (LT), closeness of track (CT), false alarm track(FAT), track matching error (TME), ID change (IDC), correct detected track (CDT) and track completeness (TC) were determined for evaluation of tracker. Correspondence between ground truth track and tracker estimated track through estimation of Euclidean distance between centroid of ground truth bounding box and estimated bounding box was established (Bashir and Porikli 2006). After establishment of correspondence, different frame based metrics such as positive prediction, false alarm rate, tracker detection rate, detection rate etc. were determined based upon counting of number of true negative, true positive, false negative and false positive frames. In addition, authors proposed object tracking error (OTE), Eq. (5), as measure of average displacement of centroid of ground truth and estimated bounding box.

$$OTE = \frac{1}{N_{gt}} \left\{ \sum_{n \in T_g \cap T_t} \sqrt{(X_{ng} - X_{nt})^2 + (Y_{ng} - Y_{nt})^2} \right\} \tag{5}$$

where X_{ng} and X_{nt} were centroid of object in frame n for ground truth and tracker estimation. N_{gt} was total number of frame where correspondence was established. For evaluating performance of mobile robot, visual contact rate (VCR) as measure of ratio of number of frames target was present in robot field of view to total number of frames processed was proposed (Germa et al. 2010). In (Bernardin and Stiefelhagen 2008; Zhang et al. 2015) CLEAR MOT matrix and multiple objects tracking accuracy (MOTA) was determined, Eq. (6), by counting number of false positive, miss detection and mismatches.

$$MOTA = 1 - \frac{\sum_{t=1}^{T} (M(t) + FP(t) + MM(t))}{\sum_{t=1}^{T} N(t)} \tag{6}$$

where at time t, M(t), FP(t), MM(t), and N(t) were number of miss, false positive, mismatch and object present respectively.

In order to measure more precise positional accuracy, performance measures considered area of overlap between ground truth bounding box and tracker estimated bounding box in subsequence frames. An area based measure of precision, recall, F measure and success rate was proposed (Kwon and Lee 2009). In this work, If GT bounding box area was A_{GT} and that of tracker output as A_{TO}. The precision, recall, and F measure were determined using Eqs. (7)–(9) respectively. The F measure estimate area of overlap between ground truth bounding box and tracker estimated bounding box at each frame. Maximum value of F measure (max at 1) indicated best estimation results. Also, author defined success rate based upon estimated value of F measure at each frame. If F measure value was greater than 0.5, object was considered to be tracked correctly.

$$r = \frac{A_{GT} \cap A_{TO}}{A_{GT}} \tag{7}$$

$$p = \frac{A_{GT} \cap A_{TO}}{A_{TO}} \tag{8}$$

$$F = \frac{2\,r\,p}{r + p} \tag{9}$$

The success rate was determined as ratio of correctly tracked frame to total number of frame in video sequences.

The performance metrics based upon statistics of pixels in area of detected object and GT were introduced (Karasulu and Korukoglu 2011). Under this work, in each frame, pixel based precision and recall metrics were determined using Eqs. (10)–(11). Where, symbol || gives number of pixels in the area, $SUM_{GT}(k)$ and $SUM_{TO}(k)$ represent spatial union of bounding boxes of GT and detected object for kth fame in video sequence.

$$P_{pa} = \begin{cases} \text{undefined}, & \text{if } SUM_{TO}(k) = \text{empty} \\ 1 - \dfrac{\left|SUM_{TO}(k) \cap \overline{SUM_{GT}(k)}\right|}{SUM_{TO}(k)}, & \text{otherwise} \end{cases} \tag{10}$$

$$R_{pa} = \begin{cases} \text{undefined}, & \text{if } SUM_{GT}(k) = \text{empty} \\ 1 - \dfrac{\left|SUM_{GT}(k) \cap \overline{SUM_{TO}(k)}\right|}{SUM_{GT}(k)}, & \text{otherwise} \end{cases} \tag{11}$$

These measures for complete video sequences were estimated as weighted average over N_{frame} where tracking was performed. Also, the authors introduced F1 measure as performance metric, Eq. (12), which was estimated from individual calculated values of pixel based precision and recall.

$$F1 = \frac{(1+\beta)P_{pa} \times R_{pa}}{\beta(P_{pa} + R_{pa})} \tag{12}$$

where parameter β was determined by user depending upon application scenario. In order to give more weightage to frames with greater intersection area between ground truth object and tracker output, an error measure for estimation of tracker accuracy was proposed (Maggio et al. 2007). The error measure was determined using Eq. (13).

$$E(k) = 1 - \frac{2 \times TP(k)}{|A_{TO}(k)| + |A_{GT}(k)|} \tag{13}$$

where in frame k, TP(k) number of pixels matched in both ground truth and tracker output. $|A_{TO}(k)|$, $|A_{GT}(k)|$ were cardinality in the area of tracker output and ground truth. The error was estimated for all frames where target was present. Also, authors proposed estimation of standard deviation of error under different run for probalistic tracker. A real time performance metric for object tracking was presented (Denman et al. 2009). The different object parameters such as position and area were explored for dynamically estimation of object detection and tracking metrics. For each of these metric, final performance score was determined while considering contribution of each frame with appropriate learning factor. Apart from quantitative measures, various qualitative measures such as ability to handle occlusion, multiple targets, rotation, scaling, deformation were investigated in different reported work. In sum, both quantitative and qualitative measures need to be considered for testing robustness and reliability of proposed algorithms.

4.3 Discussion

The different datasets for object tracking are summarized in Table 3 with listing of various parameters. The datasets are generated with incorporation of various environment challenges such as moving camera, clutter, occlusion, illumination changes etc. In general, for vision camera, tracking research works are published using standard dataset so that results can be reproduced for further study. But, for other sensing technologies, most of work was reported on self-generated data for which public access was not available. In addition, multi-modal synchronization and calibration of independent sensors is imperative for faithful tracking results. In line with this, experiment analysis was performed (Smeulders et al. 2014). Therefore, in our opinion newly reported methods need to be thoroughly evaluated on different sets of video sequences. Also, in our opinion video datasets may be assigned tracking difficulty levels which may be considered for performance metric. Mostly for vision dataset ground truth data is provided along with video sequences, but for most of multi-modal datasets ground truth needs to be extracted. The manual extraction of GT data is cumbersome task and subjected to human errors. Various challenges in extraction of ground truth data were investigated (Milan et al. 2013). A number of semi-automatic online tools such as VATIC (Vondrick et al. 2012) and multi-view annotation tool (Utasi and Benedek 2012), GTVT (Ambardekar et al. 2009), ViPER-GT (ViPER) etc. are available for extraction of GT. In another approach, in order to ease extraction of ground truth data, pseudo synthetic video sequences were generated for testing the robustness of tracking algorithms (Black et al. 2003; Schlogl et al. 2004). For prerecorded video, different challenges and complexities are intentionally introduced for generation of pseudo synthetic video.

In this review, we briefly categorized performance measures based upon availability of ground truth data. The performance measures without ground truth can be applied on real time. Performance metrics using ground truth data are more accurate and invariant to algorithm application domain. The measures based upon track similarity (Pingali and Segen 1996; Senior et al. 2001), object count (Yin et al. 2007; Bashir and Porikli 2006), area based (Kwon and Lee 2009) or pixel based performance metrics (Karasulu and Korukoglu 2011) may be considered for tracking performance evaluation. In an attempt to standardize evaluation methodology, various online tools such as ETISEO (Nghiem et al. 2007), VIVID (Collins et al. 2005), Video Analysis and Content Extraction (VACE; Kasturi et al. 2005), PETS (PETS 2006a, b) etc. are available for evaluation of performance of newly developed algorithm. Mostly, users need to submit online results and these tools provide complete performance reports. The CLEAR (CLEAR) performance evaluation workshop was mainly focused towards bringing VACE and CHIL (CHIL) program together under a common platform. The goal was to generate universal data set along with performance metrics which will help vision community for pursuing further research. However, most of performance measures do not consider system level changes in actual real system. Hence, in our view, considering overall tracking scenarios, system level performance metric needs to be investigated in future.

5 Concluding remarks and future direction

In this manuscript, we have reviewed recent advancement in single-modal and multi-modal multicue object tracking techniques. In general, combining orthogonal complementary cues not only enhance tracking accuracy over single cue techniques but also make algorithms more robust to environment challenges. Vision and thermal camera being rich source of information were extensively exploited for extracting different cues which were adaptively fused in multicue tracking framework. The thermal sensors are considered to be good alternative for vision sensor but cannot distinguish object with similar thermal profile. The laser, radar, sonar sensors are limited in extracting fine granularities of visual world. The audio sensors are good in locating object but unable to track silent person/object. Hence, vision camera was augmented by different modularity such as thermal, radar, sonar, laser, and RFID sensors for robust and reliable tracking solution development. The synchronization and calibration of different modularity is discussed to a limited extent and need further investigation. The Kinect sensor can give synchronized images from vision and thermal camera. The recent work on multicue tracking was mainly focused on either development of robust cues or integration of multiple cues in adaptive manner. Although, number of dataset for single-modal multicue are available online for testing of single-modal multicue techniques, multi-modal datasets are limitedly discussed in open literature. Although, it is well established that different independent modularity enhance tracking performance in multicue framework, sensors modularity such as seismic sensors, sonar etc. needs to be explored for tracking distance object. Considering our analysis of reviewed techniques, we conclude our survey with the following remarks:

1. Multi-modal and single-modal multicue techniques attained their advancement for real applications
2. Single-modal and multi-modal multicue methods can handle various tracking challenges more efficiently than single cue methods
3. Although, It was well established that multi-modal methods improve tracking performance, usage of more than two modularity needs more research
4. Calibrated and synchronized dataset for evaluating multi-modal techniques needs further research
5. Features with multiple cues for incorporating different tracking scenarios have promising research directions
6. Adaptive integration of cues with online cues conflict resolving solutions can be investigated

In sum, in this review we have investigated some of key single-modal and multi-modal multicue object tracking methods. Most recent developments are analyzed and compared in tabular form. We believe that our review will clear direction of research in ever growing field of multicue object tracking.

Acknowledgments The authors are grateful Defence Research and Development Organization and Delhi Technological University for financial support to this work. We would also like to thank anominous reviewers for their valuable suggestions.

References

Ahmed M, Tawab A, Abdelhalim MB, Habib SED (2014) Efficient multi-feature PSO for fast gray level object tracking. Appl Soft Comput 14:317–337

Airouche M, Bentabet L, Zelmat M, Gao G (2012) Pedestrian tracking using color, thermal and location cue measurements: a DSmT-based framework. Mach Vis Appl 23:999–1010

Ambardekar A, Niclescu M, Dascalu S (2009) Ground truth verification Tool (GTVT) for video surveillance system. In: ACHI'09 second international conferences on advances in computer–human interactions

Andriluka M, Roth S, Schiele B (2010) Monocular 3D pose estimation and tracking by detection. In: IEEE conference on computer vision and pattern recognition (CVPR 2010), USA

Baltieri D, Vezzani R, Cucchiara R (2011) 3DPes: 3D people dataset for surveillance and forensics. In: Proceedings of international ACM workshop MA3HO, Scottsdale, AZ, USA, pp 59–64

Bashir F, Porikli F (2006) Performance evaluation of object detection and tracking systems. In: IEEE international workshop on performance evaluation of tracking and surveillance (PETS)

Bernardin K, Stiefelhagen R (2008) Evaluating multiple object tracking performance: the clear MOT metrics. EURASIP J IVP 2008(1):246–309

Birchfield S (1998) Elliptical head tracking using intensity gradients and color histogram. In: Proceedings of IEEE conference on computer vision and pattern reorganization, pp 232–237

Black J, Ellis T, Rosin P (2003) A novel method for video tracking performance evaluation. In: Joint IEEE international workshop on visual surveillance and performance evaluation of tracking and surveillance (VS-PETS), pp 125–132

Blunsden SJ, Fisher RB (2010) The BEHAVE video dataset: ground truth video for multi-person behavior classification. Ann BMVA 4:1–12

Brasnett P, Mihaylova L, Bull D, Canagarajah N (2007) Sequential Monte Carlo tracking by fusing multiple cues in video sequences. Image Vis Comput 25(8):1217–1227

Brasnett P, Mihaylova L, Canagarajah N, Bull D (2005) Particle filtering with multiple cues for object tracking invideo sequences. In: Proceedings of SPIE's 17th annual symposium on electronic imaging, science and technology, vol 5685, pp 430–441

Brown LM, Senior WA, Tian Y-l, Connell J, Hampapur A, Shu C-F, Merkl H, Max L, (2005) Performance evaluation of surveillance systems under varying conditions. In: IEEE internationl workshop on performance evaluation of tracking and surveillance, Colorado

Cannons K, Gryn J, Wildes R (2010) Visual tracking using a pixel wise spatio-temporal oriented energy representation. In: Proceedings of the11th European conference on computer vision, pp 511–524

CANTATA. http://www.multitel.be/cantata/

Chakravarty P, Jarvis R (2006) Panoramic vision and laser range finder fusion for multiple person tracking. In: Proceedings of IEEE/RSJ international conference on intelligent robots and systems, pp 2949–2954

Chau DP, Bremond F, Thonnat M (2009) Online evaluation of tracking algorithm performance. In: The 3rd international conference on imaging for crime detection and prevention

Chen Y, Nguyen TV, Kankanhalli M, Yuan J, Shuicheng Yan, Meng Wang (2014) Audio matters in visual attention. IEEE Trans Circuits Syst Video Technol 24(11):1992–2003

Chen Y, Rui Y (2004) Real-time speaker tracking using particle filter sensor fusion. In: Proceedings of the IEEE, vol 92(3)

CHIL—computers in the human interaction loop. http://chil.server.de/

CLEAR: classification of events, activities and relationships. http://www.clear-evaluation.org, 2008

Collins R, Zhou X, Teh SK (2005) An open source tracking test bed and evaluation web site. In: IEEE international workshop on performance evaluation of tracking and surveillance (PETS2005)

Conaire CO, O' Connor NE, Smeaton A (2008) Thermo-visual feature fusion for object tracking using multiple spatiogram trackers. Mach Vis Appl 19:483–494

Congxia Dai, Yunfei Zheng, Xin Li (2007) Pedestrian detection and tracking in infrared imagery using shape and appearance. Comput Vis Image Underst 106:288–299

Cui J, Zha H, Zhao H, Shibasaki R (2008) Multi-modal tracking of people using laser scanners and video camera. Image Vis Comput 26:240–252

Davis J, Keck M (2005) A two-stage approach to person detection in thermal imagery. In: Proceedings of IEEE workshop on applications of computer vision

Davis J, Sharma V (2007) Background-subtraction using contour-based fusion of thermal and visible imagery. IEEE OTCBVS WS Ser Bench Comput Vis Image Underst 106(2–3):162–182

Denman S, Fookes C, Sridharan S, Lakemond R (2009) Dynamic performance measures for object tracking systems. In: Proceeding of advanced video and signal based surveillance, IEEE

Dou JF, Jianxun Li (2014) Robust visual tracking based on interactive multiple model particle filter by integrating multiple cues. Neurocomputing 135:118–129

Dou J, Li J (2014) Robust visual tracking base on adaptively multi-feature fusion and particle filter. Optik 125:1680–1686

EC Caviar project/IST 2001 37540, found at http://homepages.inf.ed.ac.uk/rbf/CAVIAR/

Erdem CE, Sankur B, Tekalp AM (2001a) Non-rigid object tracking using performance evaluation measures as feedback. In: Proceedings of IEEE internationl conference on computer vision and pattern recognition, pp II-323–II-330

Erdem CE, Tekalp AM, Sankur B (2001b) Metrics for performance evaluation of video object segmentation and tracking without ground truth. In: Proceedings of international conference on image processing, vol 2, pp 69–72

Erdem E, Dubuisson S, Bloch I (2012) Visual tracking by fusing multiple cues with context-sensitive reliabilities. Pattern Reorgan 45:1948–1959

Ess A, Leibe B, van Gool L (2007) Depth and appearance for mobile scene analysis. In: Proceedings of ICCV

Fleuret F, Berclaz J, Lengagne R, Fua P (2008) Multi-camera people tracking with a probabilistic occupancy map. IEEE Trans Pattern Anal Mach Intell 30(2):267–282

Francesc MN, Sanfeliu A, Samaras D (2005) Integration of conditionally dependent object features for robust figure/background segmentation. In: Proceedings of IEEE international conference on computer vision (ICCV), vol 2, pp 1713–1720

Francesc MN, Sanfeliu A, Samaras D (2008) Dependent multiple cue integration for robust tracking. IEEE Trans Pattern Anal Mach Intell 30(4):670–685

Gavrila DM, Munder S (2007) Multicue pedestrian detection and tracking from a moving vehicle. Int J Comput Vis 73:41–59

Gedik OS, Alatan A (2013) 3-D rigid body tracking using vision and depth sensors. IEEE Trans Cybern 43(5):1395–1405

Germa T, Lerasle F, Quadah N, Cadenat V (2010) Vision and RFID data fusion for tracking people in crowds by a mobile robot. Comput Vis Image Underst 114:641–651

Han J, Pauwels EJ, de Zeeuw Paul M, de With Peter HN (2012) Employing a RGB-D sensor for real-time tracking of humans across multiple re-entries in a smart environment. IEEE Trans Consum Electron 58(2):1318–1334

Han J, Shao L, Xu D, Shotton J (2013) Enhanced computer vision with microsoft kinect sensor: a review. IEEE Trans Cybern 43(5):1318–1334

Hong Liu, Ze Yu, Hongbin Zha, Yuexian Zou, Lin Zhang (2009) Robust human tracking based on multicue integration and mean shift. Pattern Recognit Lett 30:827–837

Hong Lu, Zou WL, Li HS, Zhang Y, Fei SM (2015) Edge and color contexts based object representation and tracking. Optik 126:148–152

HTS. http://www.ces.clemson.edu/_stb/research/headtracker/seq/

Hu J, Su TM, Cheng CC, Liu, WH, Wu TI (2002) A self-calibrated speaker tracking system using both audio and video data. In: Proceedings of the 2002 IEEE international conference on control applications

INO's Video Analytics Dataset, found at URLwww.ino.ca/en/video-analytics-dataset/

Jacobs RA (2002) What determines visual cue reliability? Trends Cogn Sci 6(8):345–350

Karasulu B, Korukoglu S (2011) A software for performance evaluation and comparison of people detection and tracking methods in video processing. Multimed Tools Appl 55(3):677–723

Kasturi R, Goldgof D, Soundararajan P, Manohar V, Boonstra M, Korzhova V (2005) Performance evaluation protocol for text, face, hands, person and vehicle detection & tracking in video analysis and content extraction (VACEII). Technical report, University of South Florida

Kılıç V, Barnard M, Wang W, Kittler J (2015) Audio assisted robust visual tracking with adaptive particle filtering. IEEE Trans Multimed 17(2):186–200

Kim DY, Moongu Jeon (2014) Data fusion of radar and image measurements for multi-object tracking via Kalman filtering. Inf Sci 278:641–652

Kinect camera. http://www.xbox.com/en-US/kinect/default.htm

Klein DA BoBoT—Bonn benchmark on tracking. http://www.iai.uni-bonn.de/~kleind/tracking/index.htm

Klein DA, Schulz D, Frintrop S, Cremers AB (2010) Adaptive real-time video-tracking for arbitrary objects. In: Proceedings of IEEE IROS, Taipei, Taiwan, pp 772–777

Kong S, Heo BAJ, Paik J, Abidi M (2005) Recent advances in visual and infrared face recognitio—a review. Comput Vis Image Underst 97(1):103–135

Kumar P, Brooks MJ, Dick A (2007) Adaptive multiple object tracking using colour and segmentation cues. In: Asian conference on computer vision (ACCV 2007), Tokyo, Japan, Lecture notes in computer science, Springer, vol 4844, pp 853–863

Kumar P, Dick A, Brooks MJ (2008) Integrated Bayesian multicue tracker for objects observed from moving cameras. In: International conference on image and vision computing, New Zealand

Kumar P, Dogancay K (2011) Fusion of colour and facial features for person matching in a camera network. In: Seventh international conference on intelligent sensors, sensor networks and information processing, Adelaide

Kumar S, Marks TK, Jones M (2014) Improving person tracking using an inexpensive thermal infrared sensor. In: IEEE conference on computer vision and pattern recognition workshops

Kwon J, Lee KM (2009) Tracking of a non-rigid object via patch-based dynamic appearance modelling and adaptive basin hopping Monte Carlo sampling. In: Proceedings of IEEE CVPR, Miami, FL, USA

Lathoud G, Odobez JM, Gatica-Perez D (2004) AV16.3: an audio-visual corpus for speaker localization and tracking. In: Proceedings of the MLMI'04 workshop

Leichter I, Lindenbaum M, Rivlin E (2014) The cues in "dependent multiple cue integration for robust tracking" are independent. IEEE Trans Pattern Anal Mach Intell 36(3):620–621

Li T, Sun S, Sattar TP, Corchado JM (2014a) Fight sample degeneracy and impoverishment in particle filters: a review of intelligent approaches. Expert Syst Appl 41:3944–3954

Li Ying, Liang S, Bai B, Feng D (2014b) Detecting and tracking dim small targets in infrared image sequences under complex backgrounds. Multimed Tools Appl 71:1179–1199

Li Z, He S, Hashem M (2014c) Robust object tracking via multi-feature adaptive fusion based on stability: contrast analysis. Vis Comput 31:1432–2315

Lim YS, Jong S, Kim M (2007) Particle filter algorithm for single speaker tracking with audio-video data fusion. In: 16th IEEE international conference on robot and human interactive communication

Liu J, Liu Y, Zhang G, Zhu P, Chen YQ (2015) Detecting and tracking people in real time with RGB-D camera. Pattern Recognit Lett 53:16–23

Loy G, Fletcher L, Apostoloff N, Zelinsky A (2002) Adaptive fusion architecture for target tracking. In: IEEE international conference on automatic face and gesture recognition (FGR)

Maggio E, Smeraldi F, Cavallaro A (2007) Adaptive multi-feature tracking in a particle filtering framework. IEEE Trans Circuits Syst Video Technol 17(10):1–12

Maggio E, Cavallaro A (2005) Multi-part target representation for color tracking. In: Proceedings of IEEE international conference on image processing, Genoa, Italy, vol 1, pp 729–732

Megherbi N, Ambellouis S, Colot O, Cabestaing F (2005) Joint audio–video people tracking using belief theory. In: Proceeding of IEEE conference on advanced video and signal based surveillance

Milan A, Schindler K, Roth S (2013) Challenges of ground truth evaluation of multi-target tracking. In: Proceeding of IEEE conference on computer vision and pattern recognition workshops (CVPRW), pp 735–742

Motai Y, Jha SK, Kruse D (2012) Human tracking from a mobile agent: optical flow and Kalman filter arbitration. Signal Process Image Commun 27:83–95

MOTINAS project: courtesy of EPSRC funded MOTINAS project (EP/D033772/1). www.eecs.qmul.ac.uk/~andrea/avss2007_d.html

Munaro M, Basso F, Menegatti E (2012) People tracking within groups with RGB-D data. In: Proceedings of the international conference on intelligent robots and systems (IROS), Algarve (Portugal)

Munoz-salinas R, Miguel GS, Carnicer RM (2008) Adaptive multi-modal stereo people tracking without background modelling. J Vis Commun Image Represent 19:75–91

Murphy RR (1996) Biological and cognitive foundations of intelligent sensor fusion. IEEE Trans Syst Man Cybern A Syst Hum 26(1):42–51

Nghiem AT, Bremond F, Thonnat M, Valentin V, ETISEO (2007) Performance evaluation for video surveillance systems. In: IEEE international conference on advanced video and signal based surveillance (AVSS), London (UK)

Nickel K, Gehrig T, Ekenel H, Stiefelhagen R, McDonough J (2005) A joint particle filter for audio-visual speaker tracking. In: International conference on multimodal interfaces (ICMI05). Torento, Italy, pp 61–68

Nickel K, Stiefelhagen R (2008) Dynamic Integration of generalized cues for person tracking. ECCV 2008, Part IV, LNCS 5305:514–526

Patino L, Ferryman J (2014) PETS 2014: dataset and challenge. In: 11th IEEE international conference on advanced video and signal based surveillance (AVSS)

Perez P, Vermaak J, Blake A (2004) Data fusion for visual tracking with particles. In: Proceedings of the IEEE, vol 92(3)

PETS 2006. http://www.cvg.reading.ac.uk/PETS2006/data.html

PETS 2006 IEEE international workshop on performance evaluation of tracking and surveillance. http://www.pets2006.net/

PETS 2014 benchmark data. Multi camera sequences containing activity with different threat and difficulty levels. Pets2014.net

Pingali G, Segen J (1996) Performance evaluation of people tracking systems. In: Proceedings of IEEE workshop on applications of computer vision, pp 33–38

Polycom Worldwide [Online]. http://www.polycom.com/

Portmann J, Lynen S, Chli M, Siegwart R (2014) People detection and tracking from aerial thermal views. In: Proceedings of IEEE conference on robotics and automation

SanMiguel JC, Cavallaro A, Martinez JM (2012) Adaptive on-line performance evaluation of video trackers. IEEE Trans Image Process 21(5):1828–1837

Scheutz M, McRaven J, Cserey Gy (2004) Fast, reliable, adaptive bimodal people tracking for indoor environments. In: IEEE conference on robots system

Schlogl T, Beleznai C, Winter M, Bischof H (2004) Performance evaluation metrics for motion detection and tracking. In: Proceedings of the pattern recognition, 17th international conference on ICPR'04, 4: IEEE Computer Society, Washington, DC, USA, pp 519–522

Schulz D (2006) A probabilistic exemplar approach to combine laser and vision for person tracking. In: Robotics: science and systems (RSS), Philadelphia, USA

Senior A, Hampapur A, Ying-Li Tian, Brown L, Pankanti S, Bole R (2001) Appearance models for occlusion handling. In: IEEE international workshop on performance evaluation of tracking and surveillance, Kauai, HI

Shen C, Hengel AVD, Dick A (2003) Probabilistic multiple cue integration for particle filter based tracking. In: Sun C, Talbot H, Ourselin S, Adriansen T (eds) Proceedings of the VIIth digital image computing: techniques and applications

Smeulders AWM, Chu DM, Cucchiara R, Calderara S, Dehghan A, Shah M (2014) Visual tracking: an experimental survey. IEEE Trans Pattern Anal Mach Intell 36(7):1442–1468

Song X, Zhao H, Cui J, Shao X, Shibasaki R, Zha H (2013) An online system for multiple interacting targets tracking: fusion of laser and vision, tracking and learning. ACM Trans Intell Syst Technol 4(1):1–21

Song X, Cui J, Zhao H, Zha H (2008) Bayesian fusion of laser and vision for multiple people detection and tracking. In: Proceedings of IEEE international conference on instrumentation, control and information technology, pp 3014–3019

Spampinato C, Palazzo S, Giordano D (2012) Evaluation of tracking algorithms performance without ground truth data. In: Proceedings of 19th IEEE international conference on image processing (ICIP), Orlando, FL, pp 1345–1348

Spengler M, Schiele B (2003) Towards robust multicue integration for visual tracking. Mach Vis Appl 14(1):50–58

Spinello L, Triebel R, Siegwart R (2009) A trained system for Multimodal perception in urban environment. In: Proceedings of the IEEE ICRA, workshop on people detection and tracking

Stolkin R, Rees D, Talha M, Florescu I (2012) Bayesian fusion of thermal and visible spectra camera data for region based tracking with rapid background adaptation. In: Proceedings of IEEE international conference on multisensor fusion information and integration, pp 192–199

Strigel E, Meissner D, Seeliger F, Wilking B, Dietmayer K (2014) The Ko-PER intersection laser scanner and video dataset. In: IEEE 17th international conference on intelligent transportation systems (ITSC), pp 1900–1901

Strobel N, Spors S, Rabenstein R (2001) Joint audio-video object localization and tracking. IEEE Signal Process Mag 18:22–33

Sun Y, Bentabet L (2010) A particle filtering and DSmT based approach for conflict resolving in case of target tracking with multiple cues. J Math Imaging Vis 36:159–167

Susperregi L, Martinez-Otzeta JM, Ansuategui A, Aitorlbarguren Sierra B (2013) RGB-D, laser and thermal sensor fusion for people following in a mobile robot. Int J Adv Robot Syst 100:1–9

Talantzis F, Pnevmatikakis A, Constantinides AG (2008) Audio-visual active speaker tracking in cluttered indoors environments. IEEE Trans Syst Man Cybern B Cybern 39(3):799–807

Talha M, Stolkin R (2012) Adaptive fusion of infra-red and visible spectra camera data for particle filter tracking of moving targets. In: Proceedings of IEEE sensors conference, pp 1–4

Talha M, Stolkin R (2014) Particle filter tracking of camouflaged targets by adaptive fusion of thermal and visible spectra camera data. IEEE Sens J 14(1):151–166

Thomaidis G, Manolis Tsgas, Lytrivis P, Karaseitanidis G, Amditis A (2013) Multiple hypothesis tracking for data association in vehicular networks. Inf Fusion 14:374–383

Torabi A, Masse G, Bilodeau GA (2012) An iterative integrated framework for thermal-visible image registration, sensor fusion and people tracking for video surveillance applications. Comput Vis Image Underst 116:210–221

Treptow A, Cielniak G, Duckett T (2005) Active people recognition using thermal and grey images on a mobile security robot. In: Proceedings of the IEEE/RSJ international conference on intelligent robots and systems (IROS), Edmonton, Alberta

Triesch J, Malsburg CVD (2001) Democratic integration: self-organized integration of adaptive cues. Neural Comput 13:2049–2074

Utasi A, Benedek C (2012) A multi-view annotation tool for people detection evaluation. In: Workshop on visual interfaces for ground truth collection in computer vision applications, Capri, Italy

ViPER-GT, the ground truth authoring tool. http://vipertoolkit.sourceforge.net/docs/gt/

VIVID database. http://vision.cse.psu.edu/data/vividEval/datasets/datasets.html

Vondrick C, Patterson D, Ramanan D (2012) Efficiently scaling up crowd sourced video annotation. Int J Comput Vis (IJCV) 101:184–204

Walia GS, Rajiv K (2014) Human detection in video and images—a state-of-the-art survey. Int J Pattern Recognit Artif Intell 28(3):1–25

Walia GS, Rajiv K (2015) Robust object tracking based upon multi-cue integration for video surveillance. Multimed Tools Appl. doi:10.1007/s11042-015-2890-0

Walia Gurjit Singh, Kapoor Rajiv (2014) intelligent video target tracking using an evolutionary particle filter based upon improved cuckoo search. Expert Syst Appl 41:6315–6326

Wang JT, Chen DB, Chen HY, Yang JY (2012) On pedestrian detection and tracking in infrared videos. Pattern Recognit Lett 33(6):775–785

Wang Q, Fang J, Yuan Y (2014) Multicue based tracking. Neurocomputing 131:227–236

Wang H, David S (2006) Efficient visual tracking by probabilistic fusion of multiple cues. In: Proceedings of the 18th international conference on pattern recognition, pp 892–895

Wu H, Sankaranarayanan A, Chellappa R (2010) Online empirical evaluation of tracking algorithms. IEEE Trans Pattern Anal Mach Intell 32(8):1443–1458

Wu H, Zheng Q (2004) Self-evaluation for video tracking systems. Technical report, Department of Eletrical and Computer Engineering, Maryland University, College Park

Xu Y, Ye M, Zunhua Zhang X (2014) Locally adaptive combining color and depth for human body contour tracking using level set method. IET Comput Vis 8(4):316–328

Yang H, Shao L, Zheng F, Wang L, Song Z (2011) Recent advances and trends in visual tracking: a review. Neurocomputing 74(18):3823–3831

Yang X, Wang M, Tao D (2015) Robust visual tracking via multi-graph ranking. Neurocomputing 159:35–43

Yilmaz A, Javed O, Shah M (2006) Object tracking: a survey. ACM Comput Surv 38(4):1–45

Yin M, Zhang J, Sun H, Gu W (2011) Multicue based camshaft guided particle filter tracking. Expert Syst Appl 38(5):6313–6318

Ying L, Zhang T, Changsheng Xu (2014) Multi-object tracking via MHT with multiple information fusion in surveillance video. Multimed Syst 21:313–326

Yin F, Makris D, Velastin SA (2007) Performance evaluation of object tracking algorithms. In: 10th IEEE international workshop on performance evaluation of tracking and surveillance, PETS 2007

Zeng F, Xuan Liu, Huang Z, Ji Y (2013) Kernel based multiple cue adaptive appearance model for robust real-time visual tracking. IEEE Signal Process Lett 20(11):1094–1097

Zhang M, Ming X, Yang J (2013a) Adaptive multicue based particle swarm optimization guided particle filter tracking in infrared videos. Neurocomputing 121:163–171

Zhang X, Hong Liu, Xiaohong Sun (2013b) Object tracking with an evolutionary particle filter based on self-adaptive multi-features fusion. Int J Adv Robot Syst 10:1–11

Zhang S, Yao H, Sun X, Lu X (2013c) Sparse coding based visual tracking: review and experiment comparison. Pattern Recognit 46(7):1772–1788

Zhang X, Li Wei, Xiuzi Ye, Maybank S (2015) Robust hand tracking via novel multicue integration. Neurocomputing 157:296–305

Zhao J, Sen-ching Cheung S (2014) Human segmentation by geometrically fusing visible-light and thermal imageries. Multimed Tools Appl 73:61–89

Zheng W, Fuller N, Theriault D, Beltke M (2014) A thermal infrared video benchmark for visual analysis. In: IEEE conference on computer vision and pattern recognition workshops

Zheng Y, Meng Y (2008) Swarming Particles with multi-feature model for free-selected object tracking. In: IEEE international conference on Intelligent robots and systems, France

Zhou H, Fei M, Sadka A, Zhang Y, Li X (2014) Adaptive fusion of particle filtering and spatio-temporal motion energy for human tracking. Pattern Recognit 47:3552–3567

Zoidi O, Nikolaidis N, Tefas A, Pitas I (2014) Stereo object tracking with fusion of texture, color and disparity information. Signal Process Image Commun 29:573–589