Image and Vision Computing 30 (2012) 630–640

Contents lists available at SciVerse ScienceDirect

Image and Vision Computing

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

Filling the gap in quality assessment of video object tracking ☆

Pedro Carvalho *, Jaime S. Cardoso [1], Luís Corte-Real

INESC Porto, Campus da FEUP, Rua Dr. Roberto Frias, n 378, 4200-465 Porto, Portugal

ARTICLE INFO

Article history:
Received 25 May 2011
Received in revised form 16 April 2012
Accepted 17 June 2012

Keywords:
Computer vision
Tracking
Algorithm assessment
Evaluation metrics
Information fusion

ABSTRACT

Current evaluation methods either rely heavily on reference information manually annotated or, by completely avoiding human input, provide only a rough evaluation of the performance of video object tracking algorithms. The main objective of this paper is to present a novel approach to the problem of evaluating video object tracking algorithms. It is proposed the use different types of reference information and the combination of heterogeneous metrics for the purpose of approximating the ideal error. This will enable a significant decrease of the required reference information, thus bridging the gap between metrics with different requirements concerning this type of data. As a result, evaluation frameworks can aggregate the benefits from individual approaches while overcoming their weaknesses, providing a flexible and powerful tool to assess and characterize the behavior of the tracking algorithms.

1. Introduction

Computer vision is moving away from fiction and making its way into everyday life. Over the last decades many steps have been taken toward the goal of automatically interpreting an image, or sequence of images, but this goal has not yet been reached. Instead, researchers continue struggling and proposing many algorithms. One of the most relevant topics within computer vision (CV) is human motion capture and tracking or, more generically, video object tracking (VOT), which is commonly used as part of a larger system, providing information for higher level analysis, as in the human motion capture system described by Moeslund and Granum [1]. Hence, the performance of a component tracking algorithm will have a great influence on all posterior processing. Also, the specificities of the application scenario may exert a strong conditioning on the choice of the algorithm. These factors, in conjunction with the large number of tracking proposals in the literature [2,3] tackling different application scenarios or scopes, augmented the need and importance of objectively comparing algorithms, characterizing their behavior and suitability for a given operational scenario.

Currently, many researchers define their own test sequences and evaluation methodologies, making it extremely difficult to replicate results and compare algorithms. Furthermore, the criteria used in the evaluation of the algorithms should be appropriate to the application scenario. This results in the use of different features (e.g., as objects' trajectory, silhouette or assigned identifier) and application of different metrics (e.g., trajectory root mean square error, detected and reference region overlap, identity consistency). Proposals for evaluation frameworks and metrics already exist, but they haven't been generally adopted by the research community. Some of these approaches focus on evaluation without comparing with a reference (commonly known as ground truth (GT)), but the results provided typically lack sufficient discriminative information. Consequently, evaluations based on comparisons with GT are commonly favored. In this case, the problem lies in the difficulty to generate such information; it is a cumbersome job, especially when detailed pixel-based references (reference silhouettes) are needed.

The gap between the two types of approaches to the VOT assessment problem described above is mainly due to the required information. The use of reference information is desirable because it enables more precise results, but it is not reasonable to expect the existence of such information for every frame to be used for testing or evaluation purposes. We propose to bridge these approaches, complementing and unifying existing metrics, based on the concept of making the very best use of the information to our disposal. The research line pursued binds information from GT-based and stand-alone (without GT) metrics with the objective of obtaining an error assessment of the tracking algorithm that is as precise as possible, but with a significant decrease of the effort associated with the generation of reference information. The proposed approach aims to combine the flexibility of metrics without GT with the greater precision of metrics that compare algorithms' results with reference information. This will provide more flexible frameworks and means to deal with a wider range of input information. Such frameworks can be used over different video datasets, making use of the GT available.

The reference information and metrics used, and their proposed combination, are intended for tracking algorithms whose results can

☆ This paper has been recommended for acceptance by Tele Tan.

* Corresponding author. Tel.: +351 22 209 4299; fax: +351 22 209 4250.
E-mail address: pedro.carvalho@inescporto.pt (P. Carvalho).
URL: http://www.inescporto.pt/~pmc (P. Carvalho).
[1] Tel.: +351 22 209 4299; fax: +351 22 209 4250.

0262-8856/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.imavis.2012.06.002

be mapped to an image region. Examples of such outputs are bounding boxes or labeled object silhouettes such as the ones illustrated in Fig. 1. To the best of our knowledge no such approach to combine information and metrics to obtain a measure for video object tracking evaluation has ever been attempted. Given that the application of metrics using a single type of information has been strongly documented, this paper will focus on the experiments with metrics combination.

The paper is structured as follows. We begin by presenting a list of abbreviations used in this paper to assist in reading. Section 2 describes different state-of-the-art approaches to the evaluation of tracking algorithms as well as efforts in the creation of data repositories and software tools. In Section 3 the concept of reference information and metrics combination is described. First the combination of metrics using reference silhouettes and bounding box segmentations is described, followed by the additional fusion of information of metrics without GT. Furthermore, we make a preliminary discussion and put forward exploratory ideas on the distribution of reference information. Section 4 describes the different metric combinations tested and the methodology used in the assessment. Results from the experiments are presented in Section 5. Conclusions and some of the possible research lines to be pursued in the future are expressed in Section 6.

Abbreviations

BB	bounding box
BBE	bounding box error
CV	computer vision
FIW	full interval weighting
GT	ground truth
HIW	half interval weighting
LI	linear interpolation
LT	linear transformation
MF	multiplication factor
NGT	non ground truth
NGTE	non ground truth error
NW	normal weighting
PD	partition-distance
PE	primary error
RS	reference silhouette
RSE	reference silhouette error
SE	secondary error
VOT	video object tracking

2. Related work

In the current context of VOT algorithm assessment, one can divide evaluation strategies into two major groups: those that use GT and those that do not, each with their strengths and weaknesses. The first type generates more precise results, but is strongly dependent on scarce information, while the second type is more easily applied, but produces noisier outputs.

The manual creation of ground truth data is typically a time consuming and tedious process, and is prone to errors. Ellis [4] has described possible types of reference data as well as the difficulties associated with the generation of such information. Black et al. [5] also identified problems associated with tracking assessment and the generation of the required ground truth, and suggested the use of pseudo-synthetic video for the evaluation. Tracking data was captured online, stored in a database and later used to generate reference video sequences with a controlled level of complexity. With this framework it was possible to generate a large variety of datasets representing different tracking scenarios with varied perceptual complexities. The problem was the bias toward the tracking algorithm used to capture the original data. In [6], semi and full-synthetic sequences were also used to overcome the problem of generating large amounts of GT. Moreover, the authors separated the evaluation of motion segmentation and tracking. Other semi-automatic tools [7,8] have been proposed to help overcome the difficulty in GT generation. Fig. 1 depicts three different types of reference information: on Fig. 1 a) silhouette reference segmentation; on Fig. 1 b) bounding box (BB) GT, shown represented over the image for illustrative purposes; on Fig. 1 c) bounding box GT in the form of a segmentation mask.

Regarding the video surveillance application area, significant steps have been recently given with the creation of the Video Surveillance Online Repository (ViSOR) [9], intended to aggregate video sequences and ground truth data to be used by the research community. An ontology was defined for annotation of the video sequences and modules for evaluation and automatic annotation can be integrated. Krinidis et al. [10] have presented an audio-visual database intended for person tracking algorithms evaluation which includes GT in the form of 3D position in time. In [11], Karasulu and Korukoglu described a software tool for comparing people detection and tracking algorithms which aggregates and implements metrics proposed in the literature.

Correia and Pereira [12] described a set of metrics for the objective evaluation of video segmentation quality. The proposed metrics targeted stand-alone (without GT) and relative (with GT) scenarios and are to be applied differently depending on the type of content class: stable content; moving content. As expected, the error in the evaluation is larger for the stand-alone metrics, which the authors report more adequate for a qualitative evaluation (e.g., ranking) rather than a quantitative one. In [13], a perceptually driven metric for the evaluation of video segmentation was proposed. The authors used sequences with synthetic artifacts to conduct psychophysical experiments and defined a metric capable of predicting the quality perceived by the human viewers. Parameter optimization was conducted for a small set of application scenarios.

Erdem et al. [14] targeted video object segmentation and tracking evaluation without reference information (a non ground truth (NGT)

(a) Silhouettes as reference segmentation

(b) Bounding box GT represented over the image

(c) Bounding box as reference segmentation

Fig. 1. Illustration of some of the possible types of reference information.

approach) by performing intra- and inter-frame evaluation. Intra-frame evaluation was accomplished by measuring color differences along the boundary of an object while the inter-frame assessment consisted of comparing an object's color histogram in subsequent frames and measuring motion vector differences along the object's boundary.

Pingali and Segen [15] evaluated tracking algorithms by comparing the computed and reference trajectories. A framework for evaluating object detection algorithms through the use of GT information is described in [16]. Correspondences between detected and reference regions are established by measuring the overlap between them. A matrix of correspondences is defined based on the overlaps and used to measure correct detections, false detections, merges and splits and detection failures.

Recent research work by Bashir et al. [17] and McManus et al. [18] aimed to define a set of measures and frameworks for the evaluation of motion detection and tracking algorithms through the use of a set of parameters such as the rate of false alarms and misdetection, and accuracy. In [17], bounding boxes were used as GT and frame-based and object-based paradigms for evaluation were defined. In [18], the F-measure is proposed to assess the performance of motion detection algorithm in the precision-recall space.

In [19] the results of the ETISEO project are presented, which was oriented to the evaluation of video surveillance systems. It defined a methodology intended to address the related problems separately and study dependencies between the algorithms and video characteristics. Video sequences were collected for the purpose of illustrating specific problems. Denman et al. [20] proposed a set of metrics intended to dynamically assess tracking systems at various stages and enable real-time feedback about the system's performance. The proposed metrics were assessed by visual comparison and by using the ETISEO metrics.

Unlike most tracking evaluation methods which are based on features and criteria derived from processing the image, Roth et al. [21] proposed the detection of events as a basis for the evaluation. The process is based on a comparison with a ground truth consisting of a list of events.

Most proposals in the literature use bounding boxes as GT to compute a set of measures encompassing the entire sequence. Although some analysis about split/merge and fragmentation errors is sometimes made, typically there is no information about the temporal evolution of the error, causing a loss of temporal resolution and making it more difficult to identify failure points or events in the sequences. Moreover, as the dimension of the bounding box is typically greater than the enclosing object, usually there is also a loss of spatial resolution.

In [22] a novel framework for evaluating tracking algorithms was described. It uses reference silhouettes as GT and partition-distance (PD) metrics [23] to compute the error measures. The proposed framework captures relevant tracking errors and has proven capable of computing an overall error measure and exhibits its temporal evolution over the sequence. The drawback is the requirement for reference segmentations. While bounding box segmentations are easier to generate, the PD computed error is less accurate than with the use of silhouettes. In [24] an initial proposal to combine different types of GT and their use in the computation of PD metrics was made. Specifically, reference silhouettes (RS) and bounding box (BB) GT were considered. The intended objective was to obtain an error measure more accurate than with bounding boxes while minimizing the need for reference silhouettes.

3. Evaluation framework

As previously stated, most evaluation approaches compare the results of tracking algorithms with some form of reference information. Although pixel-based reference silhouettes can provide more exact and complete information (other types of GT can be derived from it), they are very cumbersome to generate. To minimize this effort, the use of bounding boxes to locate and provide a coarse dimension of the objects has been the preferred strategy. The approach proposed in this paper is based on the outcome and capabilities exhibited by the partition-distance (PD) metrics as used in [22] for the evaluation of video segmentation and tracking — the metrics of this framework using only reference silhouettes will be taken as the gold standard for the evaluation. Note that PD metrics can be used with both silhouette and bounding box segmentation GT.

At the core of the partition-distance is the intersection-graph between two segmentations (a reference segmentation and an automatically produced one), defined as the bipartite graph with one node for each region of the segmentations. Two nodes are connected by an undirected, weighted edge if and only if those two regions intersect each other. The intersection-graph associated with two image segmentations can now be used as a factory of indices of similarity between partitions. The partition-distance has been defined as the problem of finding a maximum weighted matching in the intersection-graph [23]. The sum of the weights of the unmatched edges on this matching process provides the distance between both segmentations.

For the remainder of this paper it will be assumed that the reader is familiar with the basic principles of the partition-distance framework and its application to video object tracking. For simplicity the reader may consider the output of the PD framework for each frame as the distance between the reference segmentation and the segmentation resulting from the tracking algorithm's output. Nevertheless, we strongly recommend the reading of [23,22].

In Section 3.1 the GT-hybrid approach to VOT assessment as initially proposed in [24] to minimize the drawback of the original PD framework is described. It makes use of pixel-based reference silhouettes to compute the metrics. Although the GT-hybrid fulfills its objective, it is still dependent on GT information. To overcome this obstacle, the concept of combining metrics with and without GT is described in Section 3.2.

3.1. Ground truth hybrid approach

The proposal of a GT-based hybrid approach [24] was intended to enable the approximation of the “ideal” error computed with the PD metrics [22] while minimizing the problem of generating a large number of frames with exact reference silhouettes. This results in significantly less effort in the data preparation. The objective was achieved through the combination of different types of GT and their use in the computation of PD metrics (see Fig. 2). Specifically, silhouettes and bounding box masks were considered.

The use of bounding boxes in the computation of PD metrics enables the capture of the most common types of problems in tracking. However, the obtained error is coarser due to the very nature of this type of GT (see Fig. 1). To solve this, frames with reference silhouettes, when available, can be used to correct the previous error. Such a combination builds on the following: GT information in the form of a bounding box is easier to generate and there are datasets available with this type of information; the availability of reference silhouettes is of added value particularly in challenging situations such as occlusions or group movement; reference silhouette information does not need to exist for the complete sequence and therefore can focus on the most relevant segments of it; there can be different frame intervals between consecutive frames with reference silhouettes.

First, the detected and reference bounding boxes are used as segmentation masks to compute the PD metrics. This error measure (referred to as bounding box error — BBE) is typically higher than what would be obtained through the use of reference silhouettes (reference silhouette error — RSE); also, the use of bounding boxes can mask errors only observable with silhouettes. Frames with reference

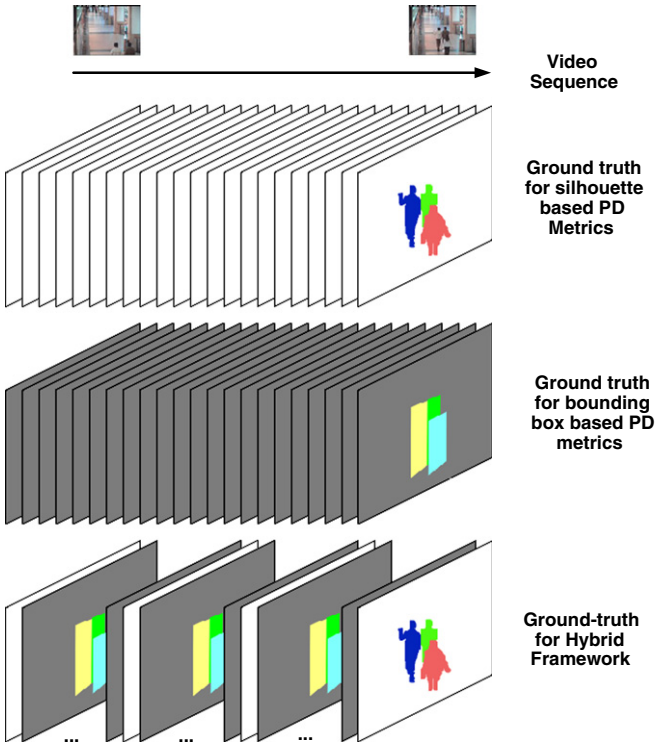

Fig. 2. Ground truth combination over a video sequence for the hybrid evaluation framework.
Reproduced from [24].

silhouettes and the corresponding error (RSE) can be used to correct the BBE in these excerpts of the sequence. An example of such a combination is illustrated in Fig. 3 and consists of the following: for two consecutive reference silhouette frames and the corresponding RSE (represented by the filled circles), a transformation (in this example a linear transformation) from BBE to RSE is computed; the transformation is then applied to the values of BBE in the interval being considered. The circles represent the RSE values: filled circles for the values used in the combination; empty circles for the values used for assessment. Note that the linear transformation is one of the possible forms of combining the information and is provided here as an example.

Through the experiments conducted in [24] and those summarized in this paper it is empirically demonstrated that the tracking error using a coarser type of GT, with a spatial error associated, can be approximated to the ideal error using a smaller number of frames containing pixel-wise silhouette GT, thus leading to a significant decrease of the effort associated with the generation of this type of information. Although corrected, information of the error evolution conveyed by the BBE is still preserved, as can be seen in Fig. 3.

3.2. A hybrid approach combining GT and NGT metrics

The use of different types of GT in the computation of the tracking error enables the minimization of the effort involved in the generation of detailed reference silhouettes. Nevertheless, it still shares to some degree the shortcoming of the original framework: it requires GT information of some form. We argue that metrics without GT (NGT) can be used in conjunction with the previously defined approach to obtain an overall error measure while avoiding the need of reference information for every frame.

In this paper we assess whether the overall error can be approximated with a significant reduction of every type of GT and when a considerable number of frames have no corresponding reference information. The overall error is expected to provide a less exact value than in the previous approach. However, we argue that even if the final measure has a slightly higher distance to the ideal error, the flexibility and less effort required can justify it. With the proposed combination we aim to enable scenarios where test and validation datasets do not require full reference information. Instead, GT (of different types) can be generated only for critical or more relevant segments of the video sequence and used in the computation of the PD metrics. Information from the NGT metrics, referred to as NGT error (NGTE), is used to fill the gaps between the intervals of frames with GT (see Fig. 4). As in the previous case, available GT is used to correct the errors associated with the NGT metrics. The process for this hybrid approach is similar to the GT-hybrid one, but now consists of two steps: first NGTE is combined with BBE; second, the result of the first operation is fused with RSE. It is out of the scope of this paper to determine the very best form of combining the information

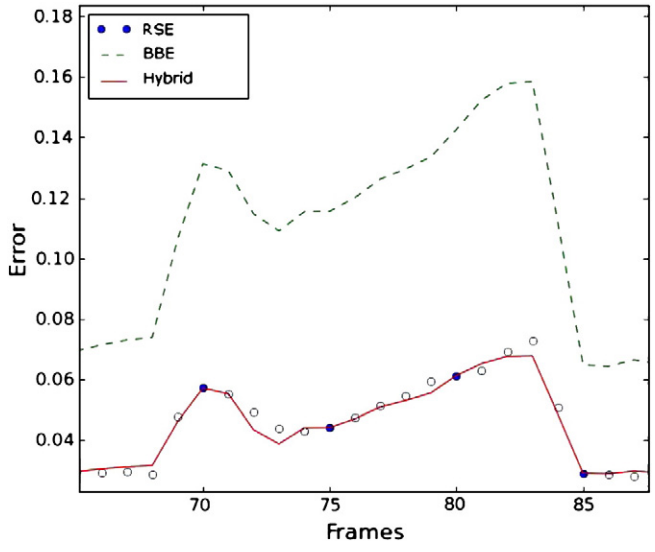

Fig. 3. Excerpt of a graphic depicting evaluation results and illustrating the error transformation effect.
Reproduced from [24].

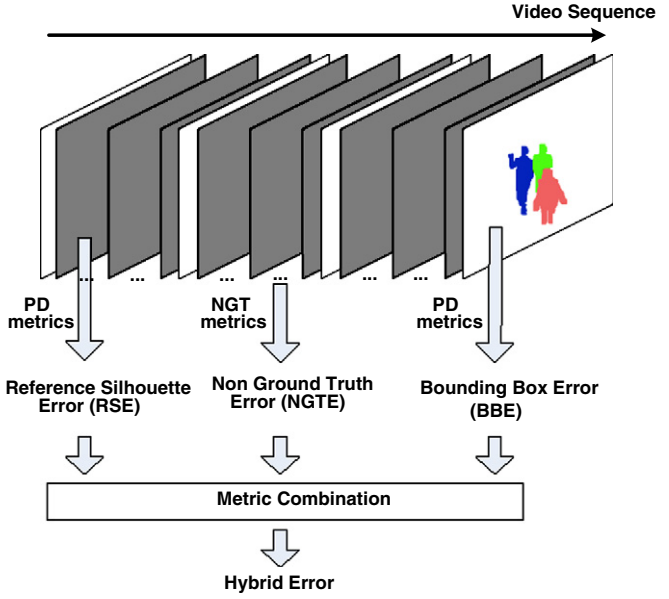

Fig. 4. Illustration of the concept of using different types of reference information and metrics.

for a given scenario or the distribution of the GT over a test/evaluation sequence.

In our experiments we have chosen the metrics proposed in [14] as they are well documented and the authors validated the results with GT information.

3.3. Identifying frames for reference information generation

The proposed framework approximates the ideal error with sparse GT. The error approximation improves with the number and precision of frames with reference information. From the practical point of view, a natural question is the selection of frames for detailed GT. For which frames is reference information critical? Which frames should receive more attention of the user in the manual annotation and for which frames a coarser reference suffices? Note that the framework proposed in this article is completely independent of this selection criteria and it is out of the scope of this work to address this question in detail. In fact, we envisage that different applications will require different options. Nevertheless, we put forward some exploratory ideas.

A default solution is to place the detailed references always equidistant, at a certain rate. Another baseline solution is to leave for the manual annotator the decision of which level of ground truth is appropriate based on the perceived complexity of the frame. Nevertheless, more automated solutions are also possible. The output of the proposed framework could be used to identify frames candidate for GT generation. For instance, one could start by annotating every frame at coarser level and evaluate the tracking quality of an algorithm under this reference. Under the assumption that the difficulty in tracking correlates with the difficulties of the coarser reference to capture the tracking quality, frames with high tracking errors are natural candidates for more precise references. These ideas, used individually or in combination, are just some of the options to decide the level of detail in the annotation.

4. Experiments

To verify the validity of our hypothesis we followed an experimental approach similar to the one described in [22]. Synthetic sequences with over 2000 frames were generated, depicting several illustrative challenges for tracking algorithms. Group movements, occlusions, miss and over detections, similarity between objects and with the background, tracks fusion and switch were simulated. Moreover, different levels of noise for the boundary localization error were used as proposed by Jiang et al. [25]: these errors were obtained by randomly selecting a point p and finding the point q nearest to p which does not belong to the same region as p; then, q is switched to the region of p provided that this step will not produce additional regions; this basic operation is repeated for n% (noise level) of all points. The purpose of the boundary localization error is to simulate common segmentation errors and make the detection of tracking errors harder; still it is only one of the types of errors simulated. A real background was also used. For object simulation, an elliptic form was chosen since this shape has been recognized as a good approximation of the form of a human in the upright position [3,26]. The elliptic shape varies according to its position and image perspective.

Although there are datasets with standard test sequences such as the PETS [27] and CAVIAR [28] datasets, few possess bounding box ground truth and fewer still reference silhouettes, thus motivating the use of synthetic sequences. Synthetic sequences enable the automatic generation of different types of GT. Specifically reference segmentations and bounding boxes were produced. A subset of reference segmentations was manually generated for two well known real sequences of the CAVIAR project [28] which were also used in the assessment of the proposed approach. Illustrative frames of the dataset are depicted in Fig. 5.

4.1. GT- and NGT-based metrics correlation

Due to the differences between the GT- and NGT-based metrics, an analysis of their correlation was conducted. The NGT-based metrics proposed by Erdem et al. [14] extract three measures which are combined to produce the overall measure for each object in each frame. The individual components are intended to measure: color differences; motion difference; histogram difference. The first is measured along the border of the object while the others involve measures over two or more consecutive frames. For each object two possible combinations are proposed, a weighted average and a fuzzy combination. Nevertheless, the authors do not close this issue and other forms of combination can be pursued in future research. Concerning the frame measure, the authors suggest an average over the objects present or the use of the maximum error for an object, since this will tend to capture most of the attention.

The correlation of individual components and their combination, as suggested in [14], with the values obtained with the PD metrics using reference silhouettes – the gold standard – was analyzed to assess the feasibility of combining the two types of metrics. The results of this study are presented in Section 5.

4.2. Metrics combination

For the combination of metrics, several possibilities were considered and compared. These are not exhaustive and others can be experimented with in future research work. Before proceeding to the description of the combinations tested, some underlying concepts are first presented.

For the following description, the primary error (PE) is considered a more exact error and used to correct a noisier second measure, which will be referred to as secondary error (SE); PE_i and SE_i are the errors at the ith frame respectively. Moreover, M intervals are considered over the sequence. Each interval can be defined as $I_k = [a_k,\ b_k],\ k = 1,\ldots,\ M$. The extremities of the interval, a_k and b_k, represent two consecutive values of the PE. Note that the intervals do not need to cover the full sequence. These concepts are illustrated in Fig. 6.

The following combinations of metrics were considered: linear transformation; linear interpolation; multiplication factor; half-interval weighting; full interval weighting; normal weighting. These forms of combination are described with more detail below.

- *Linear transformation* (*LT*)
  For each interval, the values of the PE and SE at a_k and b_k are used to compute a linear transformation f; the overall error E value is obtained by applying the transformation to the values of SE.

$$E_i = f(\mathrm{SE}_i),\quad i \in [a_k,\ b_k]\,. \tag{1}$$

- *Linear interpolation* (*LI*)
  The error values E in each interval are determined through linear interpolation of the values of PE at a_k and b_k. No SE values are used.
- *Multiplication factor* (*MF*)
  Known values of PE are used to compute the factor K that minimizes

$$\sum_{\gamma} \left(\mathrm{PE}_\gamma - K \times \mathrm{SE}_\gamma\right)^2 \tag{2}$$

  where γ represents frames for which PE exists. The error value in each interval is then obtained as:

$$E_i = K \times \mathrm{SE}_i,\quad i \in [a_k,\ b_k]\,. \tag{3}$$

- *Half-interval weighting* (*HIW*)
  In each interval, the error value E is obtained as a weighted combination of SE and PE at the extremities.

$$E_i = \alpha_1 \times \mathrm{PE}_{a_k} + \alpha_2 \times \mathrm{PE}_{b_k} + \alpha_3 \times \mathrm{SE}_i,\ \ i \in [a_k,\ b_k]. \tag{4}$$

(a) Natural frame (b) Synthetic frame

Fig. 5. Sample frames of the dataset used for validation.

The values of α_1, α_2 and α_3 are obtained as follows:

$$\alpha_1 = \begin{cases} 1-\frac{2}{D}\times d & , \quad d \leq \frac{D}{2} \\ 0 & , \quad d > \frac{D}{2} \end{cases} \tag{5}$$

$$\alpha_2 = \begin{cases} \frac{2}{D}\times d-1 & , \quad d \geq \frac{D}{2} \\ 0 & , \quad d < \frac{D}{2} \end{cases} \tag{6}$$

$$\alpha_3 = 1-\alpha_1-\alpha_2 \tag{7}$$

where

$$D = b_k - a_k \tag{8}$$

$$d = i - a_k, \quad i \in [a_k, \ b_k]. \tag{9}$$

- *Full-interval weighting* (*FIW*)
  In each interval, the error value is obtained through Eq. (4), but with the weights α_1 and α_2 computed as indicated below; the weights are normalized before being applied.

$$\alpha_1 = 1-\frac{d}{D} \tag{10}$$

$$\alpha_2 = \frac{d}{D} \tag{11}$$

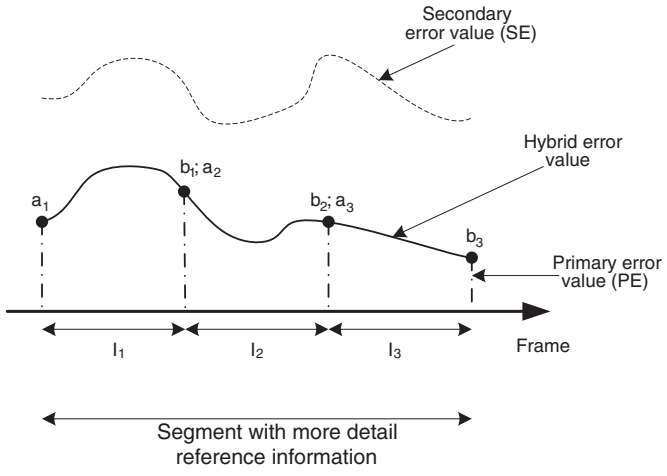

Fig. 6. Illustration of the primary error (PE) and secondary error (SE) combination, and concepts underlying the proposed combinations.

$$\alpha_3 = 1-\frac{2}{D}\times\left|d-\frac{D}{2}\right|. \tag{12}$$

The values of D and d are obtained according to Eqs. (8) and (9) respectively. The final error value is given by Eq. (4).

- *Normal weighting* (*NW*)
  In this combination, the contribution of SE is weighted by a normal function centered at the middle of the interval I_k. The values of α_1, α_2 and the final error value are computed as in the previous case.

In the experiments conducted, different interval widths were considered with each interval being delimited by consecutive frames with reference silhouette, and corresponding PE. We define step as: $\Delta = b_k - a_k$. Specifically, steps ranging from 3 to 200 frames were used. Following the procedures described in [22], the computed error was the symmetric distance (*SymDist*) of the PD metrics.

For the combination of metrics using different types of GT, the error obtained through the use of reference silhouettes (RSE) is taken as PE, while the error obtained through the use of bounding boxes (BBE) assumes the role of SE. When combining GT and NGT-based metrics the process consists of two phases: first the error without GT (NGTE) is taken as SE and BBE as PE, which are fused through one of the previously described combinations; the result is an error measure SE′ which is combined with the RSE (as PE) in the second phase, resulting in the final error value.

In the hybrid approach the interval widths described above were considered for the second phase, i.e., these intervals are delimited by consecutive frames with reference silhouette. Again, steps from 3 to 200 were considered for these intervals. For the combination of the BBE with the NGTE, sub-divisions were performed over the interval. Specifically, divisions by 2, 4, 8 and 16 were considered. When dividing the interval, a minimum width of 3 for the sub-intervals was guaranteed. For the real sequences only a division of the interval by 2 was considered due to a smaller number of frames with manually generated reference segmentation and to assure a considerable distance between consecutive frames with reference bounding boxes.

4.3. Identifying GT frame candidates

To explore the use of the proposed hybrid framework in the identification of key frames candidate for GT generation we put forward two ideas; these are compared with uniform distributions (GT frames placed equidistant) over the sequences. Both of the proposed approaches can be iterative procedures. The first idea consists of interval refinement. The error output is analyzed and intervals for which the error measure is above a given value are candidates for refinement; in our experiments we considered values above the average error and with at least 5 consecutive frames. In these candidate intervals, frames with reference information are uniformly distributed with smaller steps. The process can then be repeated. Interval steps of 25,

10 and 5 were considered. In the second idea putted forward, the error output is analyzed. Reference information is generated for frames with error values that are local maximum and the process is repeated.

5. Results

This section summarizes the outcome of the experiments conducted. First, results from combining two different types of GT in the computation of the error with PD metrics are presented, conveying the information initially presented in [24] with more detail through new experiments. Next, the correlation analysis of PD and the chosen NGT metric are presented, followed by the results regarding the combination of the metrics.

For each experiment on GT and metrics combination, the root mean square error (RMSE) of each approach was computed, relative to the PD measure using only reference silhouettes — RSE.

5.1. PD metrics with GT combination

The graphics in Fig. 7 depict the RMSE variation for the combination of metrics using reference bounding box segmentations and silhouettes in the computation of the PD metric considering different levels of noise in the boundary localization error (in addition to the common tracking and detection simulated errors). For clarity, only the results for some of the combinations are shown. It was observed that the values for full and half interval weighting were very similar, with the same happening for the values for linear interpolation and normal weighting. Hence, we chose to only present the values for half interval and normal weighting (HIW and NW respectively). In the graphics of Fig. 7 it can be observed that, up to a given interval size, the combination of the two types of GT produces better results than using only BBs, with the benefit becoming clearer with the increase of the boundary error noise level; this is mainly due to the increased error of using only BBs.

It can be observed that, although achieving better results in some of the experiments, the behavior of the linear transformation is more erratic. This is mainly caused by the magnification of the *SE* when the *PE* values delimiting the interval are very similar. The best results are obtained through the combination with the multiplication factor. This form of combination always outperforms the use of only BBs.

The results for the real sequences are depicted in Fig. 8. As in the synthetic case, the combination of error information obtained with different GT produces better results, with the exception of the linear transformation for the same reason stated above. For the real sequence, the combination MF is outperformed by other forms of fusion, which is not unexpected; when generating a synthetic sequence, the reference bounding boxes are derived from the corresponding silhouette. As the simulated form is an ellipse, there is a close relation between these two types of masks and their areas, thus favoring the combination through a multiplication factor.

5.2. GT and NGT-base metrics correlations

As described in Section 4.1, in the NGT metrics three measures are computed for each object and then combined to obtain an error measure. Two forms of combination were suggested by the authors: average and fuzzy. Two strategies for combining the error of the objects and obtaining a frame level error were also proposed: average of the errors; maximum error. Table 1 summarizes the correlation results through the computation of the normalized correlation coefficient of each individual component and their combination with regard to the value obtained with partition-distance metrics using reference silhouettes. Moreover, the two frame-level combination strategies are also considered. In each frame strategy, the best results

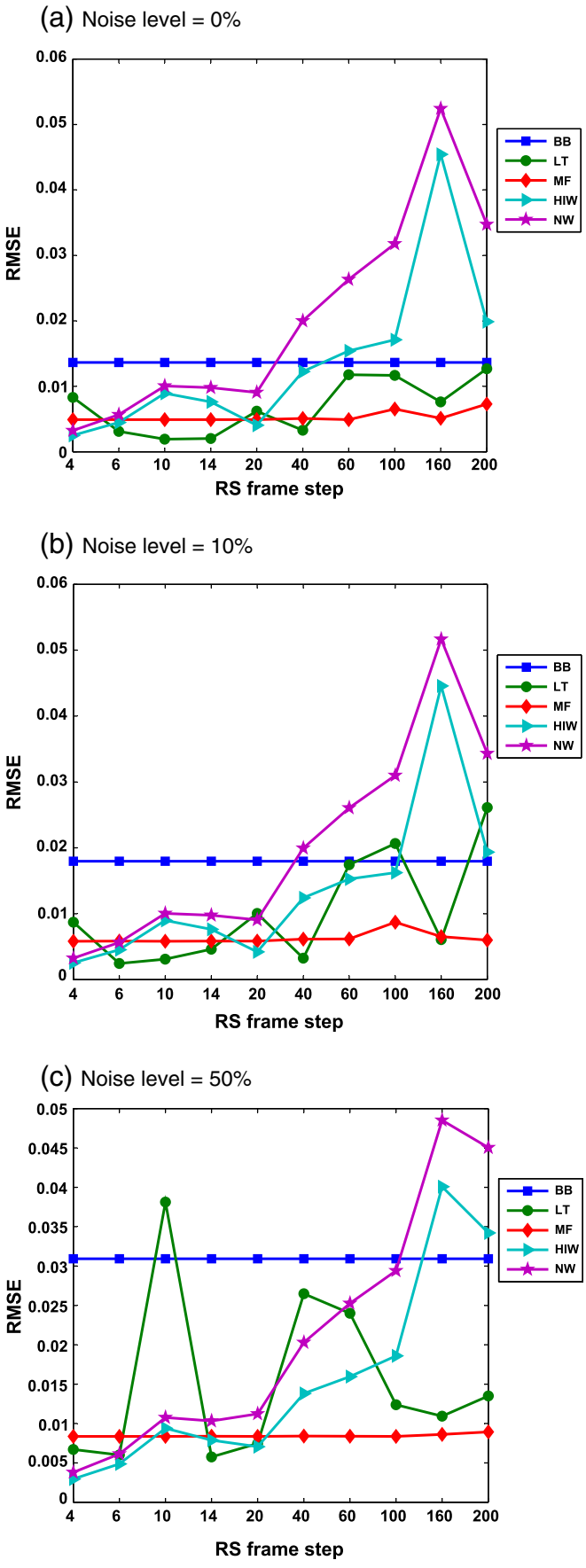

Fig. 7. Error comparisons for the GT hybrid approach over a synthetic sequence, considering different levels of generated noise and distance between consecutive RS frames.

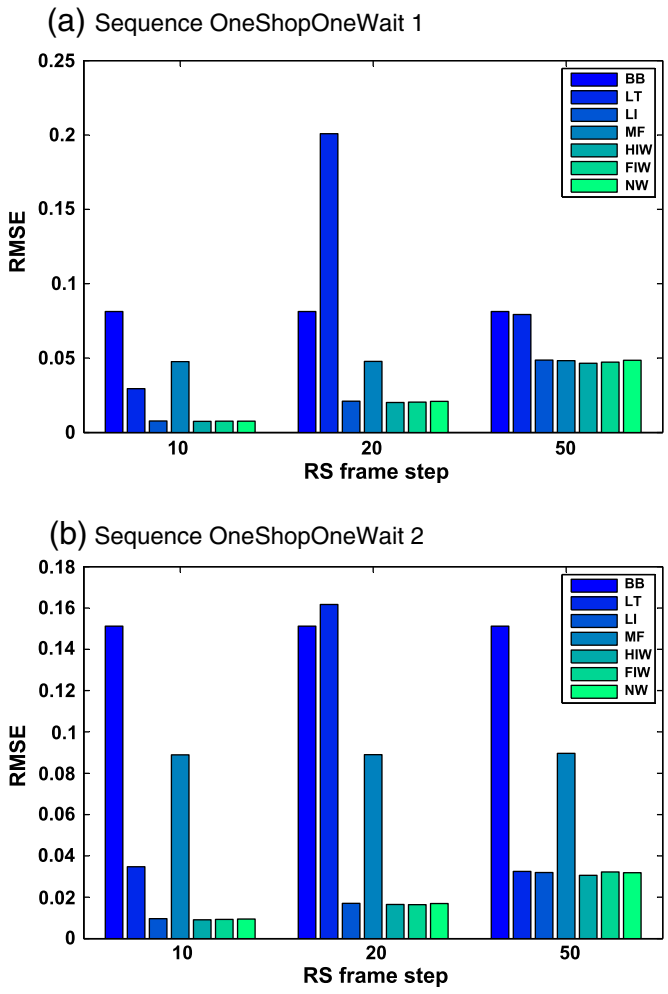

Fig. 8. Error comparisons for the GT hybrid approach over CAVIAR sequences OneShopOneWait1 and OneShopOneWait2 considering different distances between consecutive RS frames.

for each individual component and form of combination are highlighted.

From this analysis, one can observe that the correlations are weak for the synthetic case with better results being obtained for the real sequences. This is mainly due to the difficulty level associated with the synthetic sequence, specifically concerning similarity of movements and appearance of the objects which augment the error associated with the NGT measures. When analyzing the results for the synthetic sequence, the best correlation at the frame level is obtained when the maximum error per object is considered; an expected behavior since in PD metrics errors in objects with higher dimensions tend to have a higher contribution. This is consistent with the argument that errors in larger objects tend to capture most of the attention [14]. For the real sequences the correlation values are significantly higher which reveals that, even though the NGT measure is noisier than the PD metrics with reference silhouettes, their behaviors are related (i.e., they tend to capture the same errors).

The average of the maximum errors was taken for the metric fusion. Regarding the results of Table 1, note that only the suggestions of the authors regarding the combination of the components of the NGT metrics were implemented. Other forms of combination of measured components can be researched and may produce higher correlations.

5.3. GT and NGT metrics combination

This section summarizes the results on GT and NGT-based metric combination over the dataset. From the several combinations tested, two have stand out: full interval and normal weighting (FIW and NW respectively). For clarity, only the RMSE, relative to the error obtained using PD metrics and only reference silhouettes, for the measurements resulting from these forms of combination are presented. Moreover, the RMSE values for BBE and for RSE with BBE interpolation are also presented for comparison purposes. The first uses bounding boxes as reference segmentations while the second results from the combination of the reference silhouette and bounding box associated errors (RSE and BBE respectively), with the values for frames between consecutive reference BBs given by linear interpolation. Thus, one can compare each fusion derived error with the: gold standard (RSE); use of only bounding boxes; combination of reference silhouettes and bounding boxes for a sparse existence of reference information.

Fig. 9 depicts the results over a synthetic sequence. For clarity, we chose to represent only the results for a boundary localization error of 50% as this corresponds to the worst case scenario. One can observe that, with the exception of high intervals (above 100 frames), the hybrid approach error is closer to the gold standard than by simply using BBs. Even though the use of RSE with BBE interpolation presents, in some instances, an error distance similar or smaller than the full combination, its behavior is more erratic. This is due to the non-detection of errors occurring between the consecutive reference BBs. Up to 100 frame intervals, the full combination of metrics enables a smaller and/or more stable error distance.

For the real sequences, the advantage of using NGT metric information is more obvious, as shown by the experimental results summarized in Fig. 10. With one exception, for sequence OneShopOneWait2 and interval between consecutive RS frames of 50, the full fusion derived error is significantly better than the two measures being used for comparison purposes.

The better results obtained through the FIW and NW combinations are justified by the contribution of the three components (the two measures at the extremities of the interval and the noisier measure in between) for the complete interval. In HIW the weight of the extremities' values decreases toward the middle of the interval were only the noisier measure is considered.

For a better understanding of the benefits of this approach consider the following: for the real sequence OneShopOneWait1, the error obtained with GT-hybrid and hybrid approaches is approximately the same (≈0.05), but with significantly less reference information;

Table 1
Correlation between PD using reference segmentations and NGT-based metrics.

		Correlation factor with regard to *PDSym*									
		Objects average					Objects maximum				
Sequence	Noise (%)	Color diff.	Hist diff.	Motion diff.	Average	Fuzzy	Color diff.	Hist diff.	Motion diff.	Average	Fuzzy
Synthetic	0	0.05	0.06	**0.06**	**0.06**	0.05	0.09	0.09	**0.16**	**0.13**	0.12
	10	0.07	−0.02	**0.07**	**0.07**	0.06	0.10	0.04	**0.17**	**0.14**	0.14
	50	**0.32**	−0.00	0.26	**0.29**	0.29	0.35	0.14	**0.37**	**0.37**	0.36
OneShopOneWait1	–	0.35	**0.44**	−0.45	0.79	**0.96**	0.02	**0.45**	−0.20	**0.94**	0.51
OneShopOneWait2	–	0.04	0.18	**0.92**	0.65	**0.98**	**0.97**	0.52	0.95	0.69	**0.98**

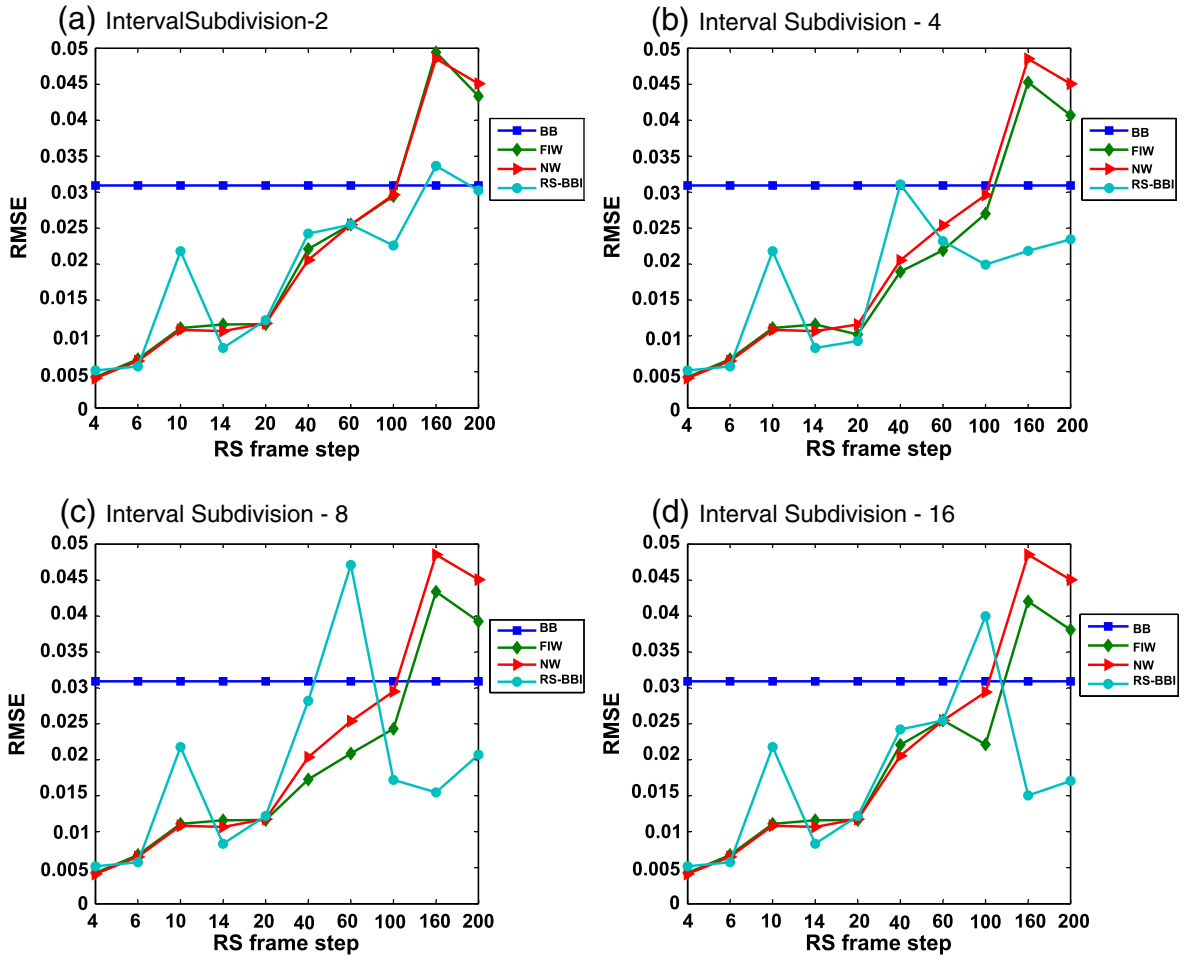

Fig. 9. Error comparisons for the hybrid approach over the synthetic sequence with 50% noise level.

for a sequence of 2000 frames with 50 frames between RS frames and interval sub-division by 2, the hybrid approach requires at least 93% less reference information than the GT-hybrid one.

5.4. GT key frame identification

Table 2 summarizes the results of the different approaches for identifying frames candidate for reference information. As previously, for each experiment it was computed the root mean square error (RMSE) relatively to the "gold standard". An initial regularization was made by uniformly distributing reference information over the sequence with steps of 50 frames. This was taken as basis for subsequent refinements. In Table 2 it is also indicated the number of frames with reference information used in each experiment.

It can be observed that using the error output to identify intervals candidate for more refined GT can produce better results than uniform spacing and, in general, with less reference frames. An exception is the uniform distribution with steps of 5 frames; however, the number of reference frames used is significantly greater. The highlight goes to the identification of key GT frames (frames with local maximum errors); it enabled better results than the other approaches and with less GT frames.

6. Conclusion

In this paper a novel approach for the evaluation of video object tracking algorithms based on the combination of reference information and metrics to provide a rich description of the algorithm's behavior was proposed. It is not intended to replace frameworks and metrics previously proposed. Rather, it should complement them by unifying the use of reference information and benefiting from the use of different metrics to overcome individual weaknesses. This proposal is intended for the assessment of visual tracking algorithms with outputs that can be mapped to image regions as is the case of bounding boxes or labeled silhouettes.

It was demonstrated that the combination of different types of ground truth can be used to decrease the effort in the generation of detailed reference silhouettes while still approximating an “ideal” error measure obtained using only this type of information. The composite measure still maintains information about the temporal evolution of the tracking error. Despite a valuable effort minimization, reference information is still required, in one form or another, for every frame in the sequence.

To overcome the shortcoming of the approach combining different types of GT the fusion of different types of GT was proposed as well as the combination of information of metrics without GT. Based on this concept, it was successfully demonstrated that it is possible to significantly decrease the GT required and eliminate its need for every frame. An error measure is obtained from the NGT metric and corrected by the available GT information. With such a solution the final error approximates, up to a given interval between consecutive frames with detailed reference silhouettes, the ideal error better than, for example, the solution using only bounding box references for every frame. This solution can contribute to an even more flexible video object tracking algorithm assessment framework.

The results from the fusion of GT and NGT metrics are expected to improve with the augmented correlation between them, resulting

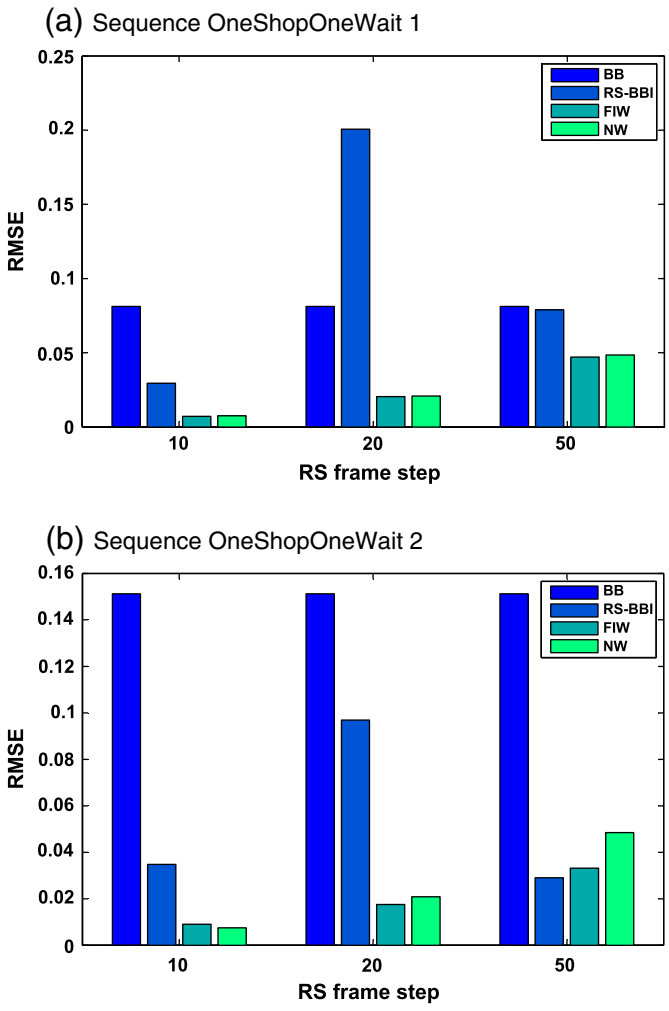

Fig. 10. Error comparisons for the hybrid approach over the real sequences.

from different combinations of the individual components of the NGT metric, as suggested by the better results obtained for the real sequences.

The experiments conducted were not exhaustive, covering every possible form of GT and metrics, since it was not the objective of this paper to identify the very best combination for every scenario. Rather, the aim was to demonstrate that based on widely used types of GT and state-of-the-art metrics the ideal error can be approximated with sparse ground truth. As the number of frames with reference information, and in particular reference silhouettes, increases the error measure can be made more accurate, but at the expense of greater effort; a trade-off must be decided according to a given assessment objective. As part of future work, it would be interesting to research, for a set of test/evaluation sequences, the impact of chosen start and end points of the intervals with reference information in order to avoid loss of information, minimizing the error uncertainty without a GT increase.

Even though the distribution of reference information and the corresponding selection criteria is out of the scope of this work, we putted forward some exploratory ideas. In particular, the use of an iterative procedure with candidate frames identified through local maximums enabled better approximations of the ideal error using significantly less GT. Nevertheless, these were just some possible ideas. Other reference frame selection criteria can be envisaged.

Other types of ground truth and metrics may be analyzed in future research work. Experiments can also be conducted to determine different forms of combination of the metrics and GT.

Acknowledgments

The authors would like to thank the Fundação para a Ciência e a Tecnologia (FCT), Portugal and the European Commission, for financing this work through the grant SFRH/BD/31259/2006 and Fundo Social Europeu (FSE).

References

[1] T.B. Moeslund, E. Granum, A survey of computer vision-based human motion capture, Comput. Vis. Image Underst. 81 (2001) 231–268.
[2] T.B. Moeslund, A. Hilton, V. Krüger, A survey of advances in vision-based human motion capture and analysis, Comput. Vis. Image Underst. 104 (2006) 90–126.
[3] A. Yilmaz, O. Javed, M. Shah, Object tracking: a survey, ACM Comput. Surv. (CSUR) 38 (2006) 13.
[4] T. Ellis, Performance metrics and methods for tracking in surveillance, In: 3rd IEEE International Workshop on Performance Evaluation of Tracking and Surveillance PETS'2002, Copenhagen, Denmark, 2002.
[5] J. Black, T. Ellis, P. Rosin, 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), 2003, pp. 125–132.
[6] T. Schlogl, C. Beleznai, M. Winter, H. Bischof, Performance evaluation metrics for motion detection and tracking, In: ICPR'04: Proceedings of the Pattern Recognition, 17th International Conference on (ICPR'04), vol. 4, IEEE Computer Society, Washington, DC, USA, 2004, pp. 519–522.
[7] D. Doennann, D. Mihalcik, Tools and techniques for video performance evaluation, In: Proceedings of the International Conference on Pattern Recognition (ICPR'00), 2000, pp. 4167–4170.
[8] C. Jaynes, S. Webb, R.M. Steele, Q. Xiong, An open development environment for evaluation of video surveillance systems, In: Proceedings of the Third International Workshop on Performance Evaluation of Tracking and Surveillance (PETS'2002), 2002, pp. 32–39.
[9] R. Vezzani, R. Cucchiara, Video surveillance online repository (ViSOR): an integrated framework, Multimed. Tools Appl. 50 (2010) 359–380.
[10] M. Krinidis, G. Stamou, H. Teutsch, S. Spors, N. Nikolaidis, R. Rabenstein, I. Pitas, An audio-visual database for evaluating person tracking algorithms, In: in Proc. IEEE ICASSP, 2005, pp. 237–240.
[11] B. Karasulu, S. Korukoglu, A software for performance evaluation and comparison of people detection and tracking methods in video processing, Multimed. Tools Appl. (2010) 1–47.
[12] P.L. Correia, F.M. Pereira, Objective evaluation of video segmentation quality, IEEE Trans. Image Process. 12 (2003) 186–200.
[13] E.D. Gelasca, T. Ebrahimi, On evaluating video object segmentation quality: a perceptually driven objective metric, IEEE J. Sel. Top. Signal Process. 3 (2009) 319–335.
[14] Çiğdem Eroğlu Erdem, B. Sankur, A.M. Tekalp, Performance measures for video object segmentation and tracking, IEEE Trans. Image Process. 13 (2004) 937–951.
[15] S.G. Pingali, J. Segen, Performance evaluation of people tracking systems, In: WACV '96: Proceedings of the 3rd IEEE Workshop on Applications of Computer Vision (WACV '96), IEEE Computer Society, Washington, DC, USA, 1996, p. 33.
[16] J.C.M.P. do Nascimento, J.S. Marques, Performance evaluation of object detection algorithms for video surveillance, IEEE Trans. Multimed. 8 (2006) 761–774.

Table 2
Error approximation using different methods for GT frame identification.

		NGT	Uniform step 50	Uniform distribution (steps)			Interval refinement (steps)			Key GT frames	
				25	10	5	25	10	5	Ite #1	Ite #2
OSOW1	RMSE	0.6064	0.0364	0.0339	0.0338	0.0281	0.0337	0.0333	0.0293	0.0318	0.0244
	# GT frames		29	56	139	276	47	100	173	67	113
OSOW2	RMSE	0.6242	0.0357	0.0268	0.0236	0.0218	0.0220	0.0222	0.0220	0.0257	0.0185
	# GT frames		31	60	147	293	64	78	127	73	119

[17] F. Bashir, F. Porikli, Performance evaluation of object detection and tracking systems, In: Proceedings of IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS), PETS, 2006.
[18] N. Lazarevic-McManus, J.R. Renno, D. Makris, G.A. Jones, An object-based comparative methodology for motion detection based on the F-measure, Comput. Vis. Image Underst. 111 (2008) 74–85.
[19] A.T. Nghiem, F. Bremond, M. Thonnat, V. Valentin, ETISEO, performance evaluation for video surveillance systems, In: Proceedings of the 2007 IEEE Conference on Advanced Video and Signal Based Surveillance, IEEE Computer Society, Washington, DC, USA, 2007, pp. 476–481.
[20] S. Denman, C. Fookes, S. Sridharan, R. Lakemond, Dynamic performance measures for object tracking systems, In: Proceedings of the 2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance, AVSS '09, IEEE Computer Society, Washington, DC, USA, 2009, pp. 541–546.
[21] D. Roth, E. Koller-Meier, D. Rowe, T.B. Moeslund, L.V. Gool, Event-based tracking evaluation metric, In: Proceedings of IEEE Workshop on Motion and Video Computing, WMVC, 2008, pp. 1–8.
[22] J.S. Cardoso, P. Carvalho, L.F. Teixeira, L. Corte-Real, Partition-distance methods for assessing spatial segmentations of images and videos, Comput. Vis. Image Underst. 113 (2009) 811–823.
[23] J.S. Cardoso, L. Corte-Real, Toward a generic evaluation of image segmentation, IEEE Trans. Image Process. 14 (2005) 1773–1782.
[24] P. Carvalho, J.S. Cardoso, L. Corte-Real, Hybrid framework for evaluating video object tracking algorithms, Electron. Lett. 46 (2010) 411–412.
[25] X. Jiang, C. Marti, C. Irniger, H. Bunke, Distance measures for image segmentation evaluation, EURASIP J. Appl. Signal Process. 2006 (2006) 1–10.
[26] T. Zhao, R. Nevatia, Tracking multiple humans in complex situations, IEEE Trans. Pattern Anal. Mach. Intell. 26 (2004) 1208–1211.
[27] PETS, In: IEEE International Workshop on Performance Evaluation of Tracking and Surveillance 2006, 2006.
[28] Caviar, EC-funded-CAVIAR-project, i. 2001-37540, , 2004..