# Hybrid framework for evaluating video object tracking algorithms

P. Carvalho, J.S. Cardoso and L. Corte-Real

A simple and efficient hybrid framework for evaluating algorithms for tracking objects in video sequences is presented. The framework unifies state-of-the-art evaluation metrics with diverse requirements in terms of reference information, thus overcoming weaknesses of individual approaches. With foundations on already demonstrated and well known metrics, this framework assumes the role of a flexible and powerful tool for the research community to assess and compare algorithms.

*Introduction:* The large number of proposed algorithms for video object tracking has led to the need for a flexible framework that enables the comparison and characterisation of such algorithms. Many researchers define their own test sequences, processes and metrics for evaluation, making it extremely difficult to replicate results and compare algorithms. Proposals for evaluation frameworks and metrics [1–3] already exist. Although some proposals focus on an evaluation without ground-truth (GT) [4, 5], the results typically provide insufficient discriminative information, particularly when the noise/errors in the background/foreground segmentations increase. As a result, evaluations based on metrics using GT are commonly preferred. The generation of this type of information is cumbersome, especially when detailed pixel-based references (reference segmentations (RS)) are needed. To minimise the GT generation effort, the use of bounding boxes (BB) to locate the object and provide a coarse dimension of the object has been the preferred strategy. Proposals such as [3, 6] consist of using this type of GT and 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, making it more difficult to identify failure points or events in the sequences. Moreover, as the size of the BB is typically greater than the enclosing object, there may also be a loss of spatial resolution. In [1] a novel framework was described for evaluating video segmentation and tracking algorithms. It uses RS as GT and partition-distance (PD) metrics [7], and is capable of computing an overall error measure as well as exhibit its temporal evolution over the sequence. The drawback is the requirement for GT in the form of RS, which are harder to generate than the BB GT. This Letter proposes a framework, based on the PD metrics, capable of providing more information than that typically obtained through the use of BB, while minimising the need for RS (Fig. 1), thus making it more flexible and easier to use by the research community. Given that the application of metrics using a single type of information has been strongly documented, this Letter focuses on the experiments with the combination of GT information.

*Evaluation framework:* The proposal described in this Letter is a hybrid framework that minimises the problem of generating a large number of frames with exact pixel-based GT, while still enabling their use for a more detailed analysis of the algorithm's behaviour in challenging situations. The objective of the framework is achieved through the combination of different types of GT and their use in the computation of PD metrics. Specifically, RS and BB are considered. The use of BB as reference in the computation of PD metrics enables the capture of the most common types of problems in tracking, but the error is coarser owing to the very nature of this type of GT. RS frames, when available, can be used to correct the previous error. Such a combination of information builds on the following: GT information in the form of BB is easier to generate and there are datasets available with this type of information; the availability of RS is of added value particularly in challenging situations such as occlusions or group movement; RSs do not need to cover the entire sequence and therefore can focus on the most relevant segments of it; there can be different frame intervals between consecutive RS frames.

In the hybrid framework the BBs from the GT and from the output of the algorithms are used to compute the PD metrics. This already provides valuable information, but the error measure (bounding box error, BBE) is typically higher than that obtained through the use of reference segmentations (RSE). Also, the use of BBs can mask errors only observable with RS. The framework can then use RS frames and corresponding RSE to correct the BBE in the extracts of the sequence for which RS exist. This is illustrated in Fig. 2 and consists of the following: for two consecutive RS frames and the corresponding PD error (represented by the filled circles), a linear transformation from BBE to RSE is computed; the transformation is then applied to the values of BBE between the reference frames.

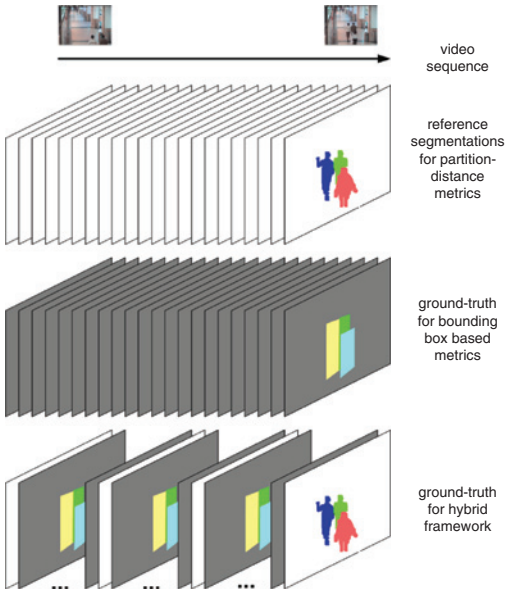


**Fig. 1** *Ground-truth combination in hybrid evaluation framework*

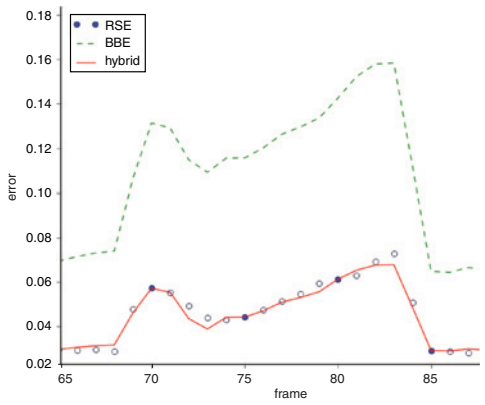


**Fig. 2** *Extract of graphic showing evaluation results and illustrating error transformation effect*

- Error value with RS (RSE) – used by hybrid framework
◦ Error value with RS (RSE) – used for validation
--- Error value using GT bounding boxes
— Error values obtained with hybrid framework

*Test of framework:* To verify the validity of our hypothesis we followed the approach described in [1]. Synthetic sequences depicting illustrative challenges and with different levels of noise were used. Such sequences enable the generation of PD errors using the reference and the bounding box segmentation. Other approaches to the combination of information were considered and are presented in this Letter for comparative and validation purposes:

- Only BB: the BBE is taken as the error.
- Factor scaling: each BBE value is multiplied by a factor K that better approximates the BBE to the RSE.
- RSE interpolation: the error for frames in the interval defined by two consecutive RS is determined by linear interpolation of the RSE of the RS frames.
- Weighted average: two consecutive RSE values and BBE in this interval are combined using weights that vary linearly with the position of the frame in the interval defined by the samples.

Experiments were conducted over the synthetic sequences using different intervals between consecutive RS frames (3, 5, 10, 15 and

20 frames). Also, two different distributions of the RS frames were considered: a uniform distribution over the sequence, and three disjoint sets with RS uniformly distributed over each set. Following the procedures described in [1], the computed error was the symmetric distance (SymDist) [7] of the PD metrics. The error obtained through the use of reference segmentations is considered the ideal error.

*Results:* Results of the experiments conducted over the synthetic sequences, with and without noise [1], are given in Table 1. For each experiment, the root mean square error (RMSE) of each approach was computed and the corresponding error percentage increase of comparative approaches relatively to the RMSE of the hybrid framework was calculated. From the obtained results it is observable that with the hybrid framework it is possible to obtain an error closer to the ideal error (using significantly less reference information). Moreover, it is possible to observe in the results shown in Fig. 2 that the hybrid approach not only approximates the BB-PD error to the RS-PD error, but it maintains to some extent the information conveyed by the BB-PD error shape, which is particularly relevant when the distance between RS frames increases.

**Table 1:** Error measure comparison between hybrid framework error and other approaches

| | Without noise | | | | | With noise | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Hybrid RMSE | Error percentage increase relatively to hybrid RMSE | | | | Hybrid RMSE | Error percentage increase relatively to hybrid RMSE | | | |
| Experiment | | Bounding box | Factor scaling | RSE interp. | Weighted avg. | | Bounding box | Factor scaling | RSE interp. | Weighted avg. |
| full seq. - step 3 | **0.0012** | +1929 | +370 | +556 | +366 | **0.0047** | +974 | +159 | +41 | +2 |
| full seq - step 5 | **0.0010** | +2405 | +481 | +1146 | +868 | **0.0033** | +1424 | +268 | +194 | +131 |
| full seq - step 10 | **0.0009** | +2737 | +558 | +2226 | +1864 | **0.0034** | +1385 | +259 | +405 | +336 |
| full seq - step 15 | **0.0054** | +364 | +8 | +342 | +264 | 0.0163 | +210 | **−25** | +4 | **−22** |
| full seq - step 20 | 0.0130 | +94 | **−55** | +176 | +139 | **0.0075** | +580 | +64 | +288 | +235 |
| three sets - step 3 | **0.0015** | +1924 | +370 | +304 | +176 | 0.0055 | +923 | +139 | **−22** | **−47** |
| three sets - step 5 | **0.0012** | +2399 | +480 | +775 | +582 | **0.0039** | +1370 | +243 | +101 | +56 |
| three sets - step 10 | **0.0011** | +2731 | +557 | +1312 | +1060 | **0.0038** | +1382 | +246 | +203 | +153 |
| three sets - step 15 | **0.0012** | +2359 | +471 | +1217 | +649 | **0.0040** | +1315 | +230 | +219 | +105 |
| three sets - step 20 | **0.0047** | +534 | +47 | +433 | +355 | **0.0052** | +995 | +155 | +212 | +163 |

*Conclusions:* The proposed framework is a flexible tool for evaluating video segmentation and tracking algorithms. It unifies previous proposals by enabling their straightforward application, but can provide richer information by applying the PD metrics to BB masks and by combining this information with the error obtained through the use of RS. Other types of ground-truth and metrics may be incorporated in the future. Moreover, experiments can be conducted to determine the impact of the chosen start and end points of the intervals with reference information in order to avoid loss of information.

*Acknowledgment:* This work has been supported by Fundação para a Ciência e a Tecnologia (FCT) – Portugal through the grant SFRH/BD/31259/2006.


*10 November 2009*
doi: 10.1049/el.2010.3165
One or more of the Figures in this Letter are available in colour online.

P. Carvalho, J.S. Cardoso and L. Corte-Real (*INESC Porto, Faculdade de Engenharia, Universidade do Porto, Campus da FEUP, Rua Dr. Roberto Frias, no. 378 4200-465 Porto, Portugal*)

E-mail: pmc@inescporto.pt

## References

1 Cardoso, J.S., Carvalho, P., Teixeira, L.F., and Corte-Real, L.: 'Partition-distance methods for assessing spatial segmentations of images and videos', *Comput. Vis. Image Underst.*, 2009, **113**, (7), pp. 811–823
2 Cavallaro, A., and Ziliani, F.: 'Characterisation of tracking performance'. Proc. Workshop on Image Analysis For Multimedia Interactive Services (WIAMIS), Montreux, Switzerland, 2005
3 Black, J., Ellis, T., and Rosin, P.: 'A novel method for video tracking performance evaluation'. Joint IEEE Int. Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance (VSPETS), Nice, France, 2003, pp. 125–132
4 Correia, P.L., and Pereira, F.M.: 'Objective evaluation of video segmentation quality', *IEEE Trans. Image Process.*, 2003, **12**, (2), pp. 186–200
5 Erdem, C.E., Sankur, B., and Tekalp, A.M.: 'Performance measures for video object segmentation and tracking', *IEEE Trans. Image Process.*, 2004, **13**, (7), pp. 937–951
6 Bashir, F., and Porikli, F.: 'Performance evaluation of object detection and tracking systems'. Proc. IEEE Int. Workshop on Performance Evaluation of Tracking and Surveillance (PETS), New York, USA, 2006, pp. 7–14
7 Cardoso, J.S., and Corte-Real, L.: 'Toward a generic evaluation of image segmentation', *IEEE Trans. Image Process.*, 2005, **14**, (11), pp. 1773–1782