

# ONLINE EVALUATION OF TRACKING ALGORITHM PERFORMANCE

Duc Phu CHAU, François BREMOND and Monique THONNAT

Pulsar team, INRIA, 2004 route des Lucioles, 06560 Valbonne, France  
{Duc-Phu.Chau, Francois.Bremond, Monique.Thonnat}@sophia.inria.fr

<http://www-sop.inria.fr/pulsar>

**Keywords:** Cognitive vision, object tracking, surveillance video, online evaluation.

## Abstract

This paper presents a method to evaluate online the performance of tracking algorithms in surveillance videos. We use a set of features to compute the confidence of trajectories and also the precision of tracking results. A global score is computed online based on these features and is used to estimate the performance of tracking algorithms. The method has been tested with two real video sequences and two tracking algorithms. The similar variations between the results obtained by the proposed method and the output of a supervised evaluation tool using ground truth data have showed the performance of our global score. The advantages of our approach over the existing state of the art approaches are: (i) few a priori knowledge information is required, (ii) the method can be applied in complex scenes containing several mobile objects and (iii) we can simultaneously compare the performance of different tracking algorithms.

## 1 Introduction

Many video understanding systems have recently been developed in the computer vision community. One of most important issues of these systems is mobile object tracking. Different techniques are used to track moving objects such as particle filtering [1] and mean shift [3] based techniques.

However the quality of tracking algorithms are always a big problem due to illumination changes, motion complexity and object contrast level. A tracking algorithm can provide satisfying tracking results in particular scenes and poor results in other real world scenes. A measure of performance evaluation of these algorithms is necessary to quantify how reliable a tracking algorithm is in a particular scene. Many types of metrics have been proposed [4, 6] and defined to address this issue but most of them are dependent on ground truth data in order to compare tracking results. In [4], the authors define three different metrics. The metric referred to as “tracking time” measures the percentage of time during which a reference data is detected and tracked. The metric referred to as “object ID per-

sistence” computes over time how many tracked objects are associated to one reference object. The last metric computes the number of reference object Ids per detected object. In [6], the performance of a tracking algorithm is based on the distance of tracked points and their corresponding true target positions. The authors also compute the area between the ground truth and detected trajectories. The greater the area between the two trajectories, the poorer the performance of the tracking algorithm. The works listed above have obtained satisfactory results but they have a common limitation: they all need ground truth data to evaluate their algorithms. However the generation of ground truth data is time consuming since human interaction is necessary and the ground truth data can only be used for the corresponding annotated video sequences. Furthermore, these methods cannot detect online the errors of tracking algorithms.

In order to solve these limitations, some approaches [5, 8] have tried to compute the performance of tracking algorithm without ground truth data. The metric proposed in [5] is based on color and motion differences along the boundary of the estimated object. For each object, the authors compute and compare the histogram of detected object pixels (inside and outside the object contour) in different time instants to know whether the tracking is good or not. However this method can only be applied for contour tracking and the color histogram of background has to be different from the color histogram of the mobile objects. The method in [8] evaluates the coherence of some moving object features (e.g. shape, color) over time. It then combines each evaluation result to form a total score of tracking quality. However with the proposed features, there are still some tracking errors that the system cannot recognize. For example the tracker can lose an object track even if this object remains in the scene. It is a popular error of most tracking algorithms and no feature has been proposed to characterize this error. Besides that, the authors of these studies have only tested their algorithms on simple videos with few people in the scene.

In this paper, we propose a new online evaluation method that is able to solve these limitations. Here we want to compute the coherence of the obtained trajectories based on a set of features. We are specially interested in the four features proposed by [8] such as motion smoothness, scale, shape and color similarity. However, for each object, the authors of this paper have

compared the feature values at current frame with the corresponding values at first frame when it appears. This calculation is not correct when the mobile object approaches or goes away from the camera. In this case, the shape, size (or even the color in the case where camera is set up on the top of the scene) will have a large change compared to the initial state. Therefore, we will reuse these features with a more adapted evaluation procedure and we also use new features for our system to be more robust.

The rest of the paper is organized as follows. Our approach is described in details in the next section. In section 3, we present different experimentations in order to validate the proposed approach. A conclusion is given in the last section as well as future work.

## 2 Online evaluation of tracking algorithm

### 2.1 Features for evaluating the tracking algorithm performance

In this paper, we aim at extracting trajectory features which best discriminate good trajectories from erroneous trajectories. For each feature, we define a local score in the interval  $[0, 1]$  to determine whether the mobile object is correctly tracked or not. The quality of a trajectory is estimated by the association of local scores computed from the extracted features. The score decreases when the system detects a tracking error and increases otherwise.

There are in total seven features, denoted from (1) to (7). Based on the frequency of each feature for the mobile objects, these features are divided into two groups: “one time features” and “every time features”. While “one time features” are the features that can be computed once (one unique time) for a mobile object, “every time features” can be computed for each frames during its tracked time. The two feature groups have the same mechanism for decreasing their local scores (in function to detected errors). However the increasing mechanisms are different because for “one time feature”, we can compute easily good tracks but for “every time features” we cannot.

A global score is calculated by combining these local scores to evaluate the quality of the tracking algorithm.

#### \* One time features

There are two features in this group: temporal length and exit zone (denoted (1) and (2)). The local score of feature  $i$  in this group is initialised with a value of 0.5 (the average score) and is calculated at time instant  $t$  ( $t > 0$ ) as follows:

$$S_t^i = \begin{cases} 0 & \text{if } S_{t-1}^i + \gamma^i \cdot a_t^i - \gamma^i \cdot b_t^i < 0 \\ S_{t-1}^i + \gamma^i \cdot a_t^i - \gamma^i \cdot b_t^i & \text{if } 0 < S_{t-1}^i + \gamma^i \cdot a_t^i - \gamma^i \cdot b_t^i < 1 \\ 1 & \text{if } S_{t-1}^i + \gamma^i \cdot a_t^i - \gamma^i \cdot b_t^i > 1 \end{cases} \quad (1)$$

where  $a_t^i$  and  $b_t^i$  are respectively the number of mobile objects that are correctly and wrongly tracked according to the feature  $i$  at time instant  $t$ ;  $\gamma^i$  is the cooling factor for increasing and decreasing the local score of feature  $i$ .

Each feature has a proper computation for the two values  $a_t^i$  and  $b_t^i$ . We can know how to determine these values in each feature section. The first and third line of the formula above

force the local score values to be in the interval  $[0,1]$ .

**1. Temporal length:** The temporal length is defined as the number of frames when a mobile object exists in the scene. If an object appears only in a scene for a very short period of time, this object is likely to be induced by noise (e.g due to segmentation errors).

The local score at time instant  $t$  of this feature is calculated as the formula (1) (with  $i = 1$ ) where  $a_t^1$  is the number of mobile objects satisfying the two conditions (correct length):

- they are not tracked any more at time instant  $t$ ,
- their temporal lengths are greater than a predefined threshold  $T_1$ ;

and  $b_t^1$  is the number of mobile objects satisfying the two conditions (wrong length):

- they are not tracked any more at time instant  $t$ ,
- their temporal lengths are lower than a predefined threshold  $T_1$ .

**2. Exit zone:** An exit zone is defined as the zone where the mobile objects are supposed to leave the scene. For each scene, we determine manually the exit zones or we can learn them given previously correctly tracked objects [7]. When an object trajectory does not end in an exit zone, we consider this object tracking to be of poor quality.

The local score  $S_t^2$  of exit zone feature at time instant  $t$  is calculated as the formula (1) (with  $i = 2$ ) where  $a_t^2$  is the number of mobile objects disappearing in an exit zone at instant  $t$  (correct exit zone); and  $b_t^2$  is the number of mobile objects disappearing in a zone different from the exit zones at instant  $t$  (wrong exit zone).

#### \* Every time features

There are five features in this group (denoted from (3) to (7)). The local score of feature  $i$  in this group is initialised with a value of 0.5 (the average score) and is calculated at time instant  $t$  ( $t > 0$ ) as follows:

$$S_t^i = \begin{cases} 0 & \text{if } S_{t-1}^i + \gamma_1^i \cdot \delta_{0b_t^i} - \gamma_2^i \cdot b_t^i < 0 \\ S_{t-1}^i + \gamma_1^i \cdot \delta_{0b_t^i} - \gamma_2^i \cdot b_t^i & \text{if } 0 < S_{t-1}^i + \gamma_1^i \cdot \delta_{0b_t^i} - \gamma_2^i \cdot b_t^i < 1 \\ 1 & \text{if } S_{t-1}^i + \gamma_1^i \cdot \delta_{0b_t^i} - \gamma_2^i \cdot b_t^i > 1 \end{cases} \quad (2)$$

where  $b_t^i$  is the number of mobile objects that are wrongly tracked according to the feature  $i$  at time instant  $t$ ;  $\gamma_1^i$  and  $\gamma_2^i$  are respectively the cooling factor for increasing and decreasing the local score of feature  $i$ ; and  $\delta_{0b_t^i}$  is the Kronecker delta of 0 and  $b_t^i$  ( $\delta_{0b_t^i} = 0$  if  $b_t^i \neq 0$ ;  $\delta_{0b_t^i} = 1$  otherwise).

Similar to previous features, in this group, each feature also has a proper computation for the value  $b_t^i$ . We can find how to determine this value in each feature section.

**3. Shape ratio:** The shape ratio of a mobile object at time instant  $t$  is calculated by  $W_t/H_t$  where  $W_t$  and  $H_t$  are its 3D width and height. If the shape of a mobile object undergoes large variations, the tracking quality becomes poor. The local score  $S_t^3$  of shape feature at time instant  $t$  is calculated as the formula (2) (with  $i = 3$ ) where  $b_t^3$  is the number of mobile objects satisfying the condition:

$$\left| \frac{W_t}{H_t} - \frac{W_{t-1}}{H_{t-1}} \right| \geq T_3 \quad (3)$$

$T_3$  is a predefined threshold.

**4. Area:** The area of a mobile object at time instant  $t$  is calculated by  $W_t H_t$  where  $W_t$  and  $H_t$  are the 3D width and height of object at time  $t$  respectively. If there is a large variation of area, the quality of tracking algorithm is not good. The local score  $S_t^4$  of area feature at time instant  $t$  is calculated as the formula (2) (with  $i = 4$ ) where  $b_t^4$  is the number of mobile objects satisfying the condition:

$$\begin{cases} \frac{W_{t-1}H_{t-1}}{W_t H_t} < T_4 \text{ if } W_t H_t \geq W_{t-1} H_{t-1} \\ \frac{W_t H_t}{W_{t-1} H_{t-1}} < T_4 \text{ if } W_t H_t < W_{t-1} H_{t-1} \end{cases} \quad (4)$$

$T_4$  is a predefined threshold.

**5. Speed:** The speed of a mobile object at time instant  $t$  is defined as the displacement of the object from the previous frame  $t - 1$  to the current one  $t$ . When a mobile object is wrongly identified, this value could increase abnormally. In order to detect that, we associate a local score  $S_t^5$  to the speed feature at time instant  $t$  as the formula (2) (with  $i = 5$ ) where  $b_t^5$  is number of mobile objects having a speed greater than a threshold  $T_5$ . The determination of this threshold value is based on the possible maximum speed of mobile objects and the frame rate of the analysed video. However, because the segmentation phase is done in 2D image, a small error in this phase (due to object shadow for example) can cause a big variation in the corresponding 3D coordinate system. Therefore, the speed threshold value has to take into account this error.

**6. Color:** In this work, the color of a mobile object is characterised by a histogram of number of pixels inside the bounding box. Other color features (e.g. MSER) can be used but this one has given good enough results. We define the histogram distance of two objects 1 and 2 as follows:

$$H^1 - H^2 = \frac{\sum_{i=1}^n |H^1(i) - H^2(i)|}{\sum_{j=1}^n H^1(j)} \quad (5)$$

where,  $n$  is the number of histogram bins;  $H^1(i)$  and  $H^2(i)$  are successive the number of pixels of object 1, 2 at bin  $i$ .

The histogram distance of two objects can be used to evaluate their color similarity. The greater this distance, the lower the color similarity. An object color usually remains similar. Therefore, for each detected mobile object, we compute the histogram distance between two consecutive frames. The tracking quality of an object can be poor if its color similarity varies too much throughout the time.

Let  $H_t^i$  be the histogram of object  $i$  at instant  $t$  and  $b_t^6$  be number of objects  $i$  satisfying the condition:

$$|H_t^i - H_{t-1}^i| \geq T_6 \quad (6)$$

$T_6$  is a predefined threshold. The local score  $S_t^6$  of color feature at instant  $t$  is calculated by the formula (2) (with  $i = 6$ ).

**7. Direction:** In general, the moving direction of a mobile object does not change or changes smoothly. In other words, it does not change suddenly during the movement of the object. So this is an important feature to detect errors of tracking algorithms. In our work, the coordinate system of ground plane

is used to calculate the mobile direction. The angle value of the movement direction is quite sensitive because the trajectory is never a complete line. In order to estimate correctly the direction of movement, the system needs to observe the mobile object in a long enough temporal interval  $\Delta t$  (more than 4 frames). Therefore, in this paper the direction of a mobile object is represented by the angle  $\alpha$  of the 2D vector formed by the object location at instant  $t$  and  $t - \Delta t$ . The value of this angle is in the interval  $[0, 2\pi]$ . In order to know whether a mobile object changes from its direction, after each  $\Delta t$  frames the system calculates  $\Delta\alpha_t$  as follows:

$$\Delta\alpha_t = \begin{cases} 0 & \text{if } t < \Delta t \\ |\alpha_t - \alpha_{t-\Delta t}| & \text{if } t \geq \Delta t \end{cases} \quad (7)$$

The direction of an object is assumed ‘‘changed’’ if a mobile object  $i$  satisfies the condition (8):

$$(T_{71} \leq \Delta\alpha_t \leq T_{72}) \quad \wedge \quad (D_{t-\Delta t, t} \geq T_{73}) \quad (8)$$

where  $T_{71}$ ,  $T_{72}$  and  $T_{73}$  are the predefined thresholds;  $D_{t-\Delta t, t}$  is the distance with which object  $i$  moves during interval  $[t - \Delta t, t]$ .

Let  $b_t^7$  be the number of mobile objects satisfying the condition (8). A local score  $S_t^7$  of the direction feature at time instant  $t$  is calculated as the formula (2) (with  $i = 7$ ).

## 2.2 The global score of tracking performance

Using the seven features we have described above, a global score is defined to evaluate online the quality of a tracking algorithm at each frame. A formula is proposed to calculate the global score at frame  $t$  as follows:

$$GS_t = \frac{\sum_{i=1}^7 w^i S_t^i}{\sum_{i=1}^7 w^i} \quad (9)$$

where  $GS_t$  is the online global score at instant  $t$ ;  $w^i$  is the weight (importance) of feature  $i$  and  $S_t^i$  is the local score of feature  $i$  at instant  $t$ . In the experimentation, we suppose that all features have the same weight.

The value of the global score is always in the interval  $[0, 1]$  because the local score values also varie in this interval. When the global score is greater than 0.5, we can say that the tracking algorithm performance is rather good. Whereas if the value of the global score is lower than 0.5, that means the tracker generally fails to track accurately the detected objects.

## 3 Experimentation and validation

In order to validate the performance of the proposed global score, we compare our results with the results of a supervised evaluation tool using ground truth data. The tool computes a set of metrics [4] to evaluate the quality of a video interpretation system. We are interested in the precision ( $TP/OD$ ) and sensitivity ( $TP/GT$ ) metric. Here  $TP$  is the True Positive,  $OD$  is the total number of detected tracks (i.e.  $TruePositive + FalsePositive$ ), and  $GT$  is the number of ground truth data

(i.e.  $TruePositive + FalseNegative$ ). Thanks to these two metrics, we compute the F-Score interpreted as a harmonic mean of the precision and sensitivity as follows:

$$F\text{-Score} = \frac{2PrecisionSensitivity}{Precision + Sensitivity} \quad (10)$$

The higher this value, the better the tracking quality. However it is important to note that the values of the F-Score and of the global score of our algorithm do not behave similarly: while the F-Score at instant  $t$  is only a score computed between instant  $t$  and  $t - 1$ , our global score at instant  $t$  takes into account the tracking quality computed on several frames. However, in theory the variation of global score has to be proportional to the F-Score value. In other word, when the F-Score increases (respectively decreases), the global score increases (respectively decreases) too. To verify that, at each instant  $t$  ( $t \geq \Delta T_{EA}$ ) we calculate the evaluation assessment value  $EA_t$  as follows:

$$EA_t = \begin{cases} 0 & \text{if } (GS_t - GS_{t-\Delta T_{EA}})(FS_t - FS_{t-\Delta T_{EA}}) < 0 \\ 1 - |GS_t - GS_{t-\Delta T_{EA}}| & \text{if } FS_t = FS_{t-\Delta T_{EA}} \\ 1 - |FS_t - FS_{t-\Delta T_{EA}}| & \text{if } GS_t = GS_{t-\Delta T_{EA}} \\ 1 & \text{if } (GS_t - GS_{t-\Delta T_{EA}})(FS_t - FS_{t-\Delta T_{EA}}) > 0 \end{cases} \quad (11)$$

where  $GS_t$  and  $FS_t$  are respectively the global score and the F-Score at  $t$ ;  $\Delta T_{EA}$  is a predefined threshold, set to 8 frames in our experimentation.

The evaluation assessment can be considered as an evaluation of global score at each time instant. The value of this metric is in the interval  $[0, 1]$ . The higher this value, the higher the performance of global score.

In our experimentation, because the frame number of a sequence is quite great, in this paper we only extract some interesting video chunks to analyse the similarity between the evaluation results of our algorithm and the output of the evaluation tool mentioned above. A global result can be found in table 3. Our online evaluation algorithm and the supervised evaluation algorithm have been tested and compared with the tracking results obtained by two tracking algorithms (denoted tracking algorithm 1 and 2). The first tracking algorithm uses long term tracking technique described in [2] and the second one is developed by the Keeneo company ([http://www.keeneo.com/uk\\_index.php](http://www.keeneo.com/uk_index.php)). The experimentation phase has been performed with two different real video sequences (see figure 1 and 5). The threshold (denoted Thrld) values used in the experimentation can be found in table 1.

Thrld	$T_1$	$T_3$	$T_4$	$T_5$	$T_6$	$T_{71}$	$T_{72}$	$T_{73}$
Value	10	0.2	0.7	2	0.45	$\frac{\pi}{4}$	$\frac{7\pi}{4}$	1.5m

Table 1. Values of thresholds used in the experimentation

The first video sequence has been provided by the Gerhome project (<http://gerhome.cstb.fr/en/home/introduction.html>). The objective of this project is to develop and try out technical solutions supporting the assistance services for enhancing autonomy of the elderly at home, by using intelligent technologies for house automation. The Gerhome video contains

one person in the scene (see figure 1). The frame rate of this sequence is 8 frames per second and the length is 5 minutes.

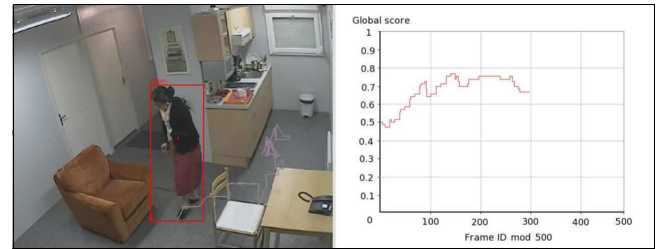


Figure 1. Online tracking evaluation of Gerhome sequence with global score computing. In the left image, the person movement is tracked by the tracking algorithm 1. The right image shows the online evaluation score of tracking algorithm 1.

In order to understand the role of the cooling factor  $\gamma$ , we compute the global score with different value sets of  $\gamma$ . For the last five features, the values of  $\gamma_1^i$  are empirically set much lower than  $\gamma_2^i$  because they are not symmetric. For example, a mobile object whose only 30% time is wrongly tracked, can be induced as an incorrect tracked object. The figure 2 represents the global score values of tracking algorithm 1 for Gerhome sequence with three different  $\gamma$  value sets (see table 2). We can see that the variation of global score values are very similar in all three cases. They increase or decrease simultaneously. However the higher the value of  $\gamma$ , the faster the variation of the global score. In other word, the reaction of global score is proportional to the values of  $\gamma$ . We choose the  $\gamma$  values of second test to continue the remain experimentations.

Parameter	$\gamma^i$ ( $i = 1, 2$ )	$\gamma_1^i$ ( $i = 3..7$ )	$\gamma_2^i$ ( $i = 3..7$ )
Test 1	0.3	0.015	0.3
Test 2	0.1	0.005	0.1
Test 3	0.05	0.0025	0.05

Table 2. The different values of  $\gamma$  used for testing (value of  $i$  denote the corresponding feature)

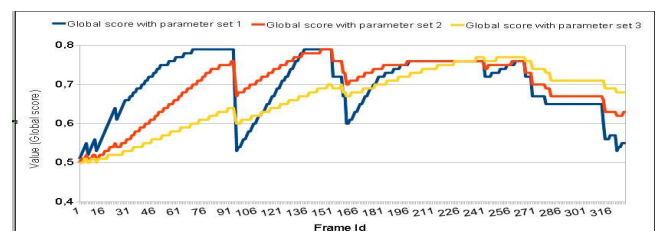


Figure 2. Global scores with different values of  $\gamma$ . The blue line corresponds to the values of  $\gamma$  in test 1, the red line corresponding to the values of  $\gamma$  in test 2, the yellow line corresponding to the values of  $\gamma$  in test 3.

The figure 3a represents the global score (blue line) and F-Score (red line) of tracking algorithm 1 from frame 1 to frame

324. We can remark that when the F-Score decreases (from frame 151 to 163 and from frame 265 to 283 for example), the graph of corresponding global score goes down too. Also when the value of F-Score is high, the global score generally increases too (for example from frame 115 to frame 145 or from frame 193 to frame 240).

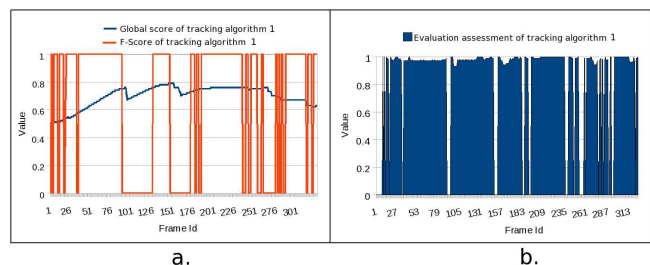


Figure 3. a) Global score (blue line) and F-Score (red line) of tracking algorithm 1 for the Gerhome sequence. b) Evaluation assessment of tracking algorithm 1 for the Gerhome sequence.

The figure 3b represents the evaluation assessment values of tracking algorithm 1 with Gerhome sequence. We can see that most of the time, this value is very high.

The figure 4 is similar to the figure 3a, but here we add the F-Score graph (the yellow one) and the global score graph (the green one) of tracking algorithm 2. We can see that the F-Score value of tracking algorithm 1 is mostly higher than the tracking algorithm 2. The global score of these two algorithms also indicates this order. The blue graph is always above the green line.

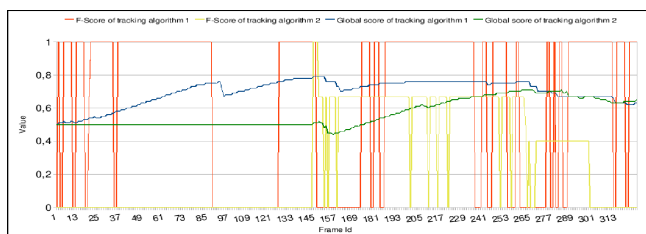


Figure 4. Global score and F-Score of tracking algorithms 1 and 2 for the Gerhome sequence. The red line and yellow lines are corresponding to the F-Score of tracking algorithm 1 and 2, the blue line and green lines corresponding to the global score of tracking algorithm 1 and 2.

The second tested video concerns the movements of people in a subway station. This sequence is extracted from the videos captured for the Caretaker project (<http://sceptre.king.ac.uk/caretaker>) (see figure 5). This project focuses on the extraction of a structured knowledge from large multimedia collections recorded over networks of cameras and microphones deployed in real sites. In this video sequence, the people number in the scene is much greater than the previous sequence. Although most people in subways follow rather simple trajectories (e.g. from entry zone straight to validating ticket zone), their trajectories are hard to be accurately detected. This is due to segmentation errors, poor video quality

(highly compressed data), numerous static, dynamic occlusions and 3D error measurement of far away object locations from the camera. The frame rate of this sequence is 5 frames per second and the length is 10 minutes.

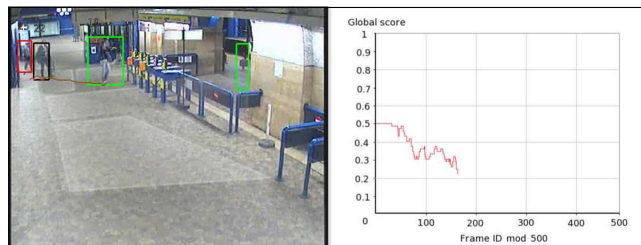


Figure 5. Online tracking evaluation of Caretaker sequence with global score computing.

The figure 6a represents the global score (blue line) and the F-Score (red line) of tracking algorithm 2 from frame 1 to frame 525. We can remark that when the F-Score decreases (from frame 46 to 222 for example), the corresponding global score goes down too. Moreover when the F-Score value increases, the global score generally increases too (for example from frame 310 to frame 486).

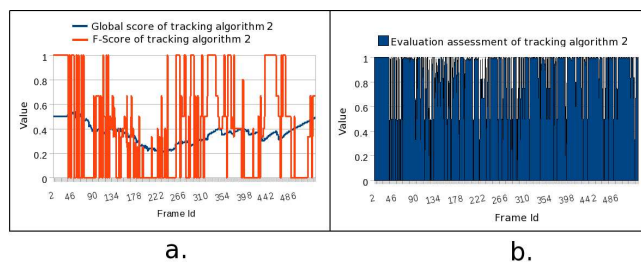


Figure 6. a) Global score (blue line) and F-Score (red line) for tracking algorithm 2 with Caretaker sequence. b) Evaluation assessment of tracking algorithm 2 with Caretaker sequence.

The figure 6b represents the evaluation assessment values of tracking algorithm 2 with Caretaker sequence. We can see that in most time, this value is very high.

In figure 7, the red and blue graphs represent respectively the F-Score values of tracking algorithm 1 and 2 from frame 1 to frame 525. We can find that in most of the time, the F-Score of algorithm 1 is lower than the value of algorithm 2.

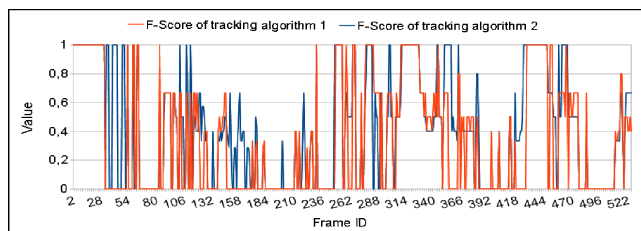


Figure 7. F-Score for the two tracking algorithms 1 (red line) and 2 (blue line) in Caretaker sequence.

The figure 8 represents the global score value of tracking algorithm 1 (the blue line) and 2 (the red line) from frame 1 to frame 525. The global score of these two tracking algorithms also shows that the tracking quality of algorithm 1 is lower than the quality given by algorithm 2.

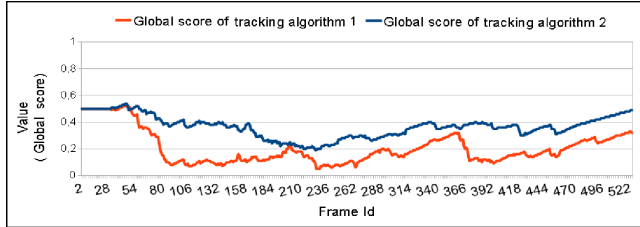


Figure 8. Global score for the two tracking algorithms (1) (red line) and (2) (blue line) in Caretaker sequence

A result summary of the experimentation phase can be found in table 3. In this table, we compute the mean value of Global score, F-Score and Evaluation assessment (denoted Eval. asses.) value of each tracking algorithm for whole sequence. Some conclusions are given:

- If the F-Score of tracking algorithm 1 is greater (respectively lower) than the F-Score of tracking algorithm 2, the global score of tracking algorithm 1 is also greater (respectively lower).

- If the F-Score of a tracking algorithm is high (e.g. greater than 0.5) (respectively low) then the global score of this tracking algorithm is also high (respectively low).

- The value of evaluation assessment is very high in all cases.

		Tracking algorithm 1	Tracking algorithm 2
Gerhome video	Global score	0.67	0.63
	F-Score	0.84	0.51
	Eval. asses.	0.93	0.94
Caretaker video	Global score	0.22	0.39
	F-Score	0.26	0.35
	Eval. asses.	0.77	0.78

Table 3. Summary of evaluation results

In conclusion, the results of the proposed online evaluation method are compatible with the output of the offline evaluation tool using ground truth data. These experimentations validate the performance of our evaluation algorithm.

## 4 Conclusion and Future work

This paper has presented a method to evaluate online the quality of a tracking algorithm. It is a difficult task and there are not many published studies concerning this problem so far. Most of the existing systems are offline and fully dependent on ground truth data. Compared to other methods, our evaluation framework is more reliable because we have tried to detect all possible cases that can cause errors during the tracking process. The

framework can be applied to different scenes. Our method has been tested and validated on real sequences. However some drawbacks still exist in this approach: thresholds are empirically set. We propose as future work an automatic learning phase to resolve this limitation. We also aim at constructing a task that is able to control the tracking process: when the system detects tracking errors, it will tune the necessary parameters so that the tracking algorithm can adapt to the current scene to get better performances.

## Acknowledgement

This work is supported by CIU Santé project (DGCIS), PACA region and General Council of Alpes Maritimes province, France.

## References

- [1] A. Almeida, J. Almeida, and R. Araujo. Real-time tracking of multiple moving objects using particle filters and probabilistic data association. *Automatika*, vol. 46, no. 1-2, pp 39-48, 2005.
- [2] A. Avanzi, F. Bremond, C. Tornieri and M. Thonnat. Design and Assessment of an Intelligent Activity Monitoring Platform. In *EURASIP Journal on Applied Signal Processing, special issue in "Advances in Intelligent Vision Systems: Methods and Applications"*, vol. 14, pp.2359-2374, 2005.
- [3] A. P. Leung and S. Gong. Mean-shift tracking with random sampling. In *British Machine Vision Conference (BMVC)*, Edinburgh (UK), September 4-7 2006.
- [4] A.T. Nghiem, F. Bremond, M. Thonnat and V. Valentin. ETISEO, Performance Evaluation for Video Surveillance Systems. In *IEEE International Conference on Advanced Video and Signal based Surveillance (AVSS)*, London (UK), September 5-7 2007.
- [5] C. Erdem, A. Tekalp and B. Sankur. Metrics for performance evaluation of video object segmentation and tracking without ground-truth. In *In Proceedings of IEEE International Conference on Image Processing (ICIP)*, 2, pp.69-72, Greece, October 7-10, 2004.
- [6] C. J. Needham, R. D. Boyle. Performance Evaluation Metrics and Statistics for Positional Tracker Evaluation. In *The Computer Vision Systems Third International Conference (ICVS)*, pp. 278-289 Springer-Verlag, New York (USA), 2003.
- [7] D.P. Chau, F. Bremond, E. Corvee, and M. Thonnat. Re-painging people trajectories based on point clustering. In *The International Conference on Computer Vision Theory and Applications (VISAPP)*, Lisboa (Portugal), February 5-8 2009.
- [8] H. Wu and Q. Zheng. Self-evaluation for video tracking systems. In *Proceedings of the 24th Army Science Conference*, USA, November 2004.