New Performance Measures for Object Tracking under Complex Environments

Ajoy Mondal

Abstract Various performance measures based on the ground truth and without ground truth exist to evaluate the quality of a developed tracking algorithm. The existing popular measures - average center location error (ACLE) and average tracking accuracy (ATA) based on ground truth, may sometimes create confusion to quantify the quality of a developed algorithm for tracking an object under some complex environments (e.g., scaled or oriented or both scaled and oriented object). In this article, we propose three new auxiliary performance measures based on ground truth information to evaluate the quality of a developed tracking algorithm under such complex environments. Moreover, one performance measure is developed by combining both two existing measures (ACLE and ATA) and three new proposed measures for better quantifying the developed tracking algorithm under such complex conditions. Some examples and experimental results conclude that the proposed measure is better than existing measures to quantify one developed algorithm for tracking objects under such complex environments.

Keywords Tracking, evaluation, performance measure, scale, orientation and complex environment.

1 Introduction

Object tracking in a video sequence has been an active research area in computer vision during the last few decades due to its several real-life application [5,11,13,15,17]. Every year, large number of tracking algorithms have been developed and corresponding papers are published in journal and conferences. The quality of the developed algorithm is compared with state-of-the-art techniques based on some performance measures. When the developed algorithm is compared with state-of-the-art techniques, several questions may be arises. Are there standard sequences that can be used for evaluation purposes? Is there a standard evaluation...
protocol? What kind of performance measure should be used? Unfortunately, there are currently no definite answers to all these questions.

However, several performance measures are proposed to quantify the quality of a newly developed tracking algorithm as compared to state-of-the-art techniques [2][4][8][18]. Center location error (CLE) and area overlap (AOL) are basically considered to evaluate the performance of a tracking algorithm. These two measures are useful to compare performances of different tracking algorithms. Other measures like precision plot (PP) based on center location error and success plot (SP) based on area overlap are also considered to evaluate the performance of the tracking algorithm throughout the complete sequence. All these measures may sometimes be unable or create confusion to quantify the quality of an algorithm for tracking single object under complex environments (like object with shape deformation, scaled object, oriented object, both scaled and oriented object). In this article, we propose three new complementary measures: error in estimated height of optimal tracker ($E_{tH}$), error in estimated width of optimal tracker ($E_{tw}$) and error in estimated orientation angle of optimal tracker ($E_{tθ}$); and another composite measure based on three proposed and two existing measures to evaluate the performance of a tracking algorithm under such complex environments. Various examples and experimental results conclude that the proposed composite measure is better to quantify the quality of an algorithm than the existing measures.

2 Existing Performance Measures based on Ground Truth Information

Two performance measures namely center location error ($E_{tC}$) and area overlap ($E_{tA}$) are popularly considered to quantify a tracking algorithm. They can be calculated as

$$E_{tC} = \| C_{tg} - C_t \|;$$

$$E_{tA} = \frac{T_t \cap T_{tg}}{T_t \cup T_{tg}},$$

where $T_{tg}$ is a ground truth rectangular tracker with center $C_{tg}$ at $t^{th}$ frame and $T_t$ is a rectangular tracker with center $C_t$ at $t^{th}$ frame generated by a tracking algorithm. Similarly average center location error ($E_{avg}^C$) and average area overlap or average tracking accuracy ($E_{avg}^A$) over all the frames can be calculated as

$$E_{avg}^C = \frac{1}{N} \sum_{t=1}^{N} \| C_{tg} - C_t \|;$$

$$E_{avg}^A = \frac{1}{N} \sum_{t=1}^{N} \frac{T_t \cap T_{tg}}{T_t \cup T_{tg}},$$

if video contains $N$ number of frames. For a good tracking algorithm, $E_{tC}$ and $E_{avg}^C$ should be 0 whereas $E_{tA}$ and $E_{avg}^A$ should be 1. Some plots (e.g., precision plot and success plot) based on center location error and area overlap, respectively are considered to evaluate performance of the tracking algorithms.

**Precision Plot:** It is considered to measure the overall tracking performance [1][2][15]. It shows that the percentage of frames where the estimated location of the object is within the given threshold distance with respect to the ground truth.
Success Plot: To measure the performance of an algorithm on a sequence of frames, we count the number of successful frames whose overlap $E_t^A$ is larger than a given threshold value. The success plot shows the ratios of successful frames at different thresholds varied from 0 to 1. Success rate value at a specific threshold (e.g., 0.5) may not always be a good representative for evaluating the tracker’s performance [15]. Instead of that, the area under the curve (AUC) of each success plot is considered to evaluate the performance of the tracking algorithm [15]. Though these measures are popularly considered to evaluate the performance of a tracking algorithm, they may sometimes unable or create confusion to quantify the quality of a developed algorithm for tracking an object under some complex environments (e.g., scaled or oriented or both scaled and oriented object). The following section presents some examples where the existing performance measures fail or create confusion to evaluate the performance of a tracking algorithm under such complex environments.

3 Problems of Existing Evaluation Measures for Scaled Object Tracking

3.1 Example1:

Here, we consider a simple example (see Figure 1(a)). Suppose $T_{g}^1$ (green colored rectangle) with center $C_{g}^1$, height $h_{g}^1$ and width $w_{g}^1$ be the ground truth tracker. Again let $T_{1}^1$ (red colored rectangle) with center $C_{1}^1$, height $h_{1}^1$ and width $w_{1}^1$ and $T_{2}^1$ (pink colored rectangle) with center $C_{2}^1$, height $h_{2}^1$ and width $w_{2}^1$ be two trackers generated by two different (scaled object) tracking algorithms Alg1 and Alg2. Here, we also assume that the trackers $T_{g}^1$ and $T_{1}^1$ have same center locations but different heights and widths (see Figure 1(a)). Therefore, we have $C_{g}^1 = C_{1}^1$, $h_{g}^1 < h_{1}^1$ and $w_{g}^1 < w_{1}^1$. Now let $w_{1}^1 = w_{g}^1 + w'$ and $h_{1}^1 = h_{g}^1 + h'$. We also assume that the trackers $T_{2}^1$ and $T_{g}^1$ have same heights but different center locations and widths (see Figure 1(b)). Therefore, we have $C_{g}^1 < C_{2}^1$, $h_{g}^1 = h_{2}^1$ and $w_{g}^1 < w_{2}^1$. Now let $C_{2}^1 = C_{g}^1 + \alpha'$ and $w_{2}^1 = w_{g}^1 + w'$. 

Fig. 1 Example 1 - area overlap between ground truth tracker and trackers obtained by (a) Alg1 and (b) Alg2.
Now area overlap ($E_{A_1}^t$) between ground truth tracker $T_g^t$ and tracker $T_1^t$ generated by Alg1 at $t^{th}$ frame, is calculated as

$$E_{A_1}^t = \frac{T_g^t \cap T_1^t}{T_g^t \cup T_1^t} = \frac{w_1^t h_1^t}{w_1^t h_1^t + w_1^t h_1^t + w_1^t h_1^t} = \frac{w_1^t h_1^t}{w_1^t h_1^t + w_1^t h_1^t} = \frac{w_1^t h_1^t}{w_{1,h}^t h_{1,w}^t}, \quad (3)$$

where $s^t = w_1^t h_1^t + w_1^t h_1^t$.

Again area overlap ($E_{A_2}^t$) between ground truth tracker $T_g^t$ and tracker $T_2^t$ generated by Alg2 at $t^{th}$ frame, is calculated as

$$E_{A_2}^t = \frac{T_g^t \cap T_2^t}{T_g^t \cup T_2^t} = \frac{w_2^t h_2^t}{w_2^t h_2^t + w_2^t h_2^t + w_2^t h_2^t} = \frac{w_2^t h_2^t}{w_{2,h}^t h_{2,w}^t}, \quad (4)$$

since $s^t = w_2^t h_2^t + w_2^t h_2^t$.

Now center location errors $E_{C_1}^t$ and $E_{C_2}^t$ of trackers $T_1^t$ and $T_2^t$ are calculated as

$$E_{C_1}^t = ||C_g^t - C_1^t|| = 0, \quad (5)$$

since $C_g^t = C_1^t$.

$$E_{C_2}^t = ||C_g^t - C_2^t|| = ||C_g^t - (C_g^t + \alpha t)|| = \alpha t. \quad (6)$$

Eqs. (3) and (4) conclude that $E_{A_1}^t > E_{A_2}^t$, i.e., tracker $T_2^t$ has more area overlap with ground truth tracker $T_g^t$, than tracker $T_1^t$ at $t^{th}$ frame. On the other hand, Eqs. (5) and (6) conclude that $E_{C_1}^t < E_{C_2}^t$, i.e., tracker $T_1^t$ has lesser center location error than tracker $T_2^t$ at $t^{th}$ frame. If we consider center location error as a performance measure then tracker $T_1^t$ is better than tracker $T_2^t$ at $t^{th}$ frame. If we consider area overlap as a performance measure then tracker $T_2^t$ is better than tracker $T_1^t$ at $t^{th}$ frame. If we consider both these measures no decision can be made. However, tracker $T_1^t$ has lesser center location error with lesser area overlap and tracker $T_2^t$ has higher center location error with higher area overlap at $t^{th}$ frame. Therefore, minimum center location error indicates maximum area overlap for the fixed sized tracker is not true for the scaled tracker and performance measures: center location error and area overlap individually can make ambiguous decision for the scaled tracker.

3.2 Example2:

Here, we consider another simple example (see Figure 2). Suppose $T_g^t$ (green colored rectangle) with center $C_g^t$, height $h_g^t$ and width $w_g^t$ be the ground truth tracker at $t^{th}$ frame. Again let $T_1^t$ (red colored rectangle) with center $C_1^t$, height $h_1^t$ and width $w_1^t$ and $T_2^t$ (pink colored rectangle) with center $C_2^t$, height $h_2^t$ and width $w_2^t$ be two trackers generated by two different (scaled object) tracking algorithms Alg1 and Alg2, respectively. Here, we also assume that the trackers $T_g^t$ and $T_1^t$ have same center location and width but different height (see Figure 2(a)). Therefore, we have $C_g^t = C_1^t$, $w_g^t = w_1^t$ and $h_g^t > h_1^t$. Now let $h_2^t = h_1^t + h^t$. We also assume that
the trackers $T^t_1$ and $T^t_2$ have same center location and height but different width (see Figure 2(b)). Therefore, we have $C^t_y = C^t_2$, $h^t_y = h^t_2$ and $w^t_1 > w^t_2$. Now let $w^t_2 = w^t_1 + w^t_2$.

Now area overlap ($E^t_{A_1}$) between ground truth tracker $T^t_g$ and tracker $T^t_1$ generated by Alg1 at $t^{th}$ frame, is calculated as

$$E^t_{A_1} = \frac{T^t_g \cap T^t_1}{T^t_g \cup T^t_1} = \frac{h^t_1 w^t_1}{h^t_y w^t_y} = \frac{(h^t_y - h^t_1) w^t_1}{h^t_y w^t_y} = 1 - \frac{h^t_1}{h^t_y}. \quad (7)$$

Again area overlap ($E^t_{A_2}$) between the ground truth tracker $T^t_g$ and tracker $T^t_2$ generated by Alg2 at $t^{th}$ frame, is calculated as

$$E^t_{A_2} = \frac{T^t_g \cap T^t_2}{T^t_g \cup T^t_2} = \frac{w^t_2 h^t_2}{w^t_y h^t_y} = \frac{(w^t_1 - w^t_2) h^t_2}{w^t_1 h^t_y} = 1 - \frac{w^t_2}{w^t_y}. \quad (8)$$

Now center location errors $E^t_{C_1}$ and $E^t_{C_2}$ of the trackers $T^t_1$ and $T^t_2$ at $t^{th}$ frame, respectively are calculated as

$$E^t_{C_1} = \| C^t_g - C^t_1 \| = 0, \quad (9)$$

since $C^t_0 = C^t_1$.

$$E^t_{C_2} = \| C^t_g - C^t_2 \| = 0, \quad (10)$$

since $C^t_0 = C^t_2$.

Now if,

$$1 - \frac{h^t_1}{h^t_y} = 1 - \frac{w^t}{w^t_y} \Rightarrow \frac{h^t_1}{h^t_y} = \frac{w^t}{w^t_y}, \quad (11)$$

then $E^t_{C_1} = E^t_{C_2}$. Therefore, this example shows that both these algorithms Alg1 and Alg2 are equally good.
Fig. 3 Example 1 - area overlap between ground truth oriented tracker and fixed trackers obtained by (a) Alg1 (b) Alg2.

4 Problems of Existing Evaluation Measures for Oriented Object Tracking

4.1 Example1:

Here, we consider one example where the object is changing its orientation from one frame to another, but the trackers $T_1$ and $T_2$ generated by two different tracking algorithms Alg1 and Alg2, respectively are fixed. We assume that ground truth tracker ($T_g$) is not fixed i.e., it changes its orientation based on object orientation. Suppose, in $t^{th}$ frame, the ground truth tracker $T_g^t$ (green colored rectangle) with center $C_g^t$, height $h_g^t$ and width $w_g^t$, is look like as depicted in Figure 3(a). After few number of frames, say in $r^{th}$ ($t < r$) frame, the object changes its orientation (assume with 90°). And the corresponding ground truth tracker $T_g^r$ (green colored rectangle) with center $C_g^r$, height $h_g^r$ and width $w_g^r$ in the $r^{th}$ frame can be look like as in Figures 3(b) and (c). Two different tracking algorithms (Alg1 and Alg2) generate two fixed sized trackers $T_1^r$ (dotted red colored rectangle) with center $C_1^r$, height $h_1^r$ & width $w_1^r$ and $T_2^r$ (dotted pink colored rectangle) with center $C_2^r$, height $h_2^r$ & width $w_2^r$ in the $r^{th}$ frame are displayed in Figures 3(b) and (c), respectively. We also assume that $C_g^r = C_1^r$, $C_g^r < C_2^r$, $h_1^r = h_2^r = w_g^r$, $w_1^r = w_2^r = h_g^r$. 
and \(C_2^r = C_2^g + \alpha^r\). We also assume that \(w_1^r = w_2^g + \alpha^r\), \(h_2^g = h_2^r + \alpha^r\), \(w_2^r = w_2^g + \alpha^r\) and \(h_2^r = h_2^g + \alpha^r\).

Now area overlap \((E_{A_1}^r)\) between ground truth tracker \(T_g^r\) and tracker \(T_1^r\) generated by Alg1 at the \(r^{th}\) frame, is calculated as

\[
E_{A_1}^r = \frac{T_g^r \cap T_1^r}{T_g^r \cup T_1^r} = \frac{h_1^r w_1^g + h_1^g w_1^r + h_1^r w_1^r}{h_1^r w_1^g + h_1^g w_1^r + h_1^r w_1^r},
\]

(12)

Again area overlap \((E_{A_2}^r)\) between the ground truth tracker \(T_g^r\) and tracker \(T_2^r\) generated by Alg2 at the \(r^{th}\) frame, is calculated as

\[
E_{A_2}^r = \frac{T_g^r \cap T_2^r}{T_g^r \cup T_2^r} = \frac{h_2^r w_2^g + h_2^g w_2^r + h_2^r w_2^r}{h_2^r w_2^g + h_2^g w_2^r + h_2^r w_2^r},
\]

(13)

since \(h_1^r = h_2^r\) and \(w_1^r = w_2^r\).

Now center location errors \(E_{C_1}^r\) and \(E_{C_2}^r\) of the trackers \(T_1^r\) and \(T_2^r\) at the \(r^{th}\) frame, respectively are calculated as

\[
E_{C_1}^r = \|C_g^r - C_1^r\| = 0,
\]

(14)

since \(C_2^g = C_1^r\).

\[
E_{C_2}^r = \|C_g^r - C_2^r\| = \|C_g^r - (C_g^r + \alpha^r)\| = \alpha^r,
\]

(15)

since \(C_2^g = C_g^r + \alpha^r\).

Eqs. (12) and (13) conclude that \(E_{A_2}^r = E_{A_1}^r\), i.e., both trackers \(T_2^r\) and \(T_1^r\) at \(t^{th}\) frame are equally good. On the other hand, Eqs. (14) and (15) conclude that \(E_{C_1}^r < E_{C_2}^r\), i.e., the tracker \(T_1^r\) has lesser center location error than the tracker \(T_2^r\) at the \(r^{th}\) frame. If we consider center location error as a performance measure then the tracker \(T_1^r\) is better than the tracker \(T_2^r\) at the \(r^{th}\) frame. If we consider area overlap as a performance measure then both the trackers \(T_1^r\) and \(T_2^r\) are equally good at the \(t^{th}\) frame. If we consider both these measures no decision can be made. Therefore, minimum center location error indicates maximum area overlap for the fixed sized tracker is not true for the oriented object and performance measures individually can make ambiguous decision for the oriented object.

4.2 Example2:

Here, we consider another simple example (see Figure 4). Suppose \(T_g^r\) (green colored rectangle) with center \(C_g^r\), height \(h_g^r\), width \(w_g^r\) and \(\theta_g^r = 45^o\) be the ground truth tracker at the \(t^{th}\) frame is displayed in Figure 4(a). After few number of frames, the object changes its orientation and the corresponding ground truth
Fig. 4 Example 2 - area overlap between ground truth oriented tracker and oriented trackers obtained by (a) Alg1 (b) Alg2.

The tracker is also oriented based on object orientation. Let $T^o_g$ (green colored rectangle) with center $C^g$, height $h^g$, width $w^g$ and $\theta^g = 0^\circ$ be the ground truth tracker at the $n^{th}$ frame. Again let $T^o_1$ (red colored rectangle) with center $C^1$, height $h^1$, width $w^1$ & orientation $\theta^1 = 135^\circ$ and $T^o_2$ (pink colored rectangle) with center $C^2$, height $h^2$, width $w^2$ & orientation $\theta^2 = 45^\circ$ be two trackers generated by two different (oriented object) tracking algorithms Alg1 and Alg2 and are displayed in Figures (b) and (c), respectively.

Here, we also assume that the trackers $T^o_g$ and $T^o_1$ have same center location, width, height but different orientation (see Figure 4(b)). Therefore, we have $C^g = C^1$, $w^g = w^1$, $h^g = h^1$ and $\theta^g \neq \theta^1$. Now let $\theta^1 = \theta^g + 135^\circ$. We also assume that the trackers $T^o_g$ and $T^o_2$ have same center location, height, width but different orientation (see Figure 4(c)). Therefore, we have $C^g = C^2$, $h^g = h^2$, $w^g = w^2$ and $\theta^g \neq \theta^2$. Now let $\theta^2 = \theta^g + 45^\circ$.

This example concludes that two different oriented trackers have the same values for area overlap and center location error. If we consider the existing performance measures: area overlap and center location error, then both these Alg1 and Alg2 are equally good. But for oriented object tracking, both these algorithms Alg1 and Alg2 provide different interpretations.
5 Problems of Existing Evaluation Measure for Scaled and Oriented Object Tracking

5.1 Example 1:

Here, we consider one example where object is changing its scale and orientation together from one frame to another and the trackers $T_1$ and $T_2$ generated by two different tracking algorithms Alg1 and Alg2, respectively are not fixed (i.e., both are changing their scale and orientation together based on object’s scale and orientation). We assume that ground truth tracker ($T_g$) is not fixed i.e., it changes its scale and orientation based on object’s scale and orientation. We assume that at $t^{th}$ frame, the ground truth tracker (green colored rectangle) $T_g^t$ with center $C_g^t$, height $h_g^t$, width $w_g^t$ and orientation $\theta_g^t = 0^o$ is presented as in Figure 5(a). In the $r^{th}$ ($r > t$) frame, the object changes its scale and orientation jointly and the corresponding ground truth tracker (green colored rectangle) $T_g^r$ with center $C_g^r$, height ($h_g^r$), width ($w_g^r$) and $\theta_g^r = 90^o$ is shown in both Figures 5(b) and (c). Here,
it is also assume that $h'_r > h'_g$ and $w'_g > w'_y$. Therefore, we have \( h'_g = h'_t + h' \), \( w'_g = w'_t + w' \) and \( \theta'_g = \theta'_t + 90^\circ \).

We assume that two different tracking algorithms generates two trackers \( T^r_1 \) (dotted red colored rectangle with center \( C^r_1 \), height \( h^r_1 \) & orientation \( \theta^r_1 = 0^\circ \)) and \( T^r_2 \) (dotted pink colored rectangle with center \( C^r_2 \), height \( h^r_2 \), width \( w^r_2 \) and orientation \( \theta^r_2 = 90^\circ \)) in Figure 5(b) and (c), respectively. Here, we also assume that \( w^r_1 = w^r_2, h^r_1 = h^r_2 - h', h^r_2 = h^r_g, w^r_2 = w^r_g - w', C^r_1 = C^r_g, C^r_2 = C^r_g + w' \) and \( h^r > w' \).

Area overlap for both the trackers \( T^r_1 \) and \( T^r_2 \) with ground truth tracker \( T^g \) are calculated as

\[
E_{A1} = \frac{T^r_1 \cap T^r_1}{T^r_1 \cup T^r_1} = \frac{w'_r(h'_r - h')}{w'_r h'_g} = 1 - \frac{h'}{h_g}
\]

and

\[
E_{A2} = \frac{T^r_2 \cap T^r_2}{T^r_2 \cup T^r_2} = \frac{w'_r(h'_r - w')}{w'_r w'_y} = 1 - \frac{w'}{w_y}.
\]

Now center location errors \( E^r_{C_1} \) and \( E^r_{C_2} \) for both the trackers \( T^r_1 \) and \( T^r_2 \), respectively are calculated as

\[
E^r_{C_1} = ||C^r_1 - C^r_g|| = 0
\]

and

\[
E^r_{C_2} = ||C^r_2 - C^r_g|| = ||C^r_g + w' - C^r_g|| = w'.
\]

Now if \( \frac{h^r_g}{h^r_y} = \frac{w^r}{w^r_y} \), then \( E^r_{A1} = E^r_{A2} \). This example highlights that two different trackers have different center location error but same amount of area overlap if \( \frac{h^r_g}{h^r_y} = \frac{w^r}{w^r_y} \). Therefore, in this example also, the existing performance measures fail to evaluate quality of scaled and oriented object tracking algorithms. In the next section, we have developed three new auxiliary measures to quantify tracking algorithms in such complex environments.

6 Proposed Performance Measures

Three new auxiliary evaluation measures namely error in estimated height of optimal tracker (\( E^t_h \)), error in estimated width of optimal tracker (\( E^t_w \)) and error in estimated orientation angle of optimal tracker (\( E^t_\theta \)), respectively have been proposed to validate the tracking algorithms. For the \( t^{th} \) frame, these are calculated as

\[
E^t_h = ||h^t_g - h^t||, \quad E^t_w = ||w^t_g - w^t||, \quad E^t_\theta = ||\theta^t_g - \theta^t||.
\]
where $h^t_g$, $w^t_g$ and $\theta^t_g$ are the actual height, width and orientation angle, respectively, and $h^t$, $w^t$ and $\theta^t$ denote the estimated height, width and orientation angle for the $t^{th}$ frame, respectively. $\|\|$ is the Euclidean norm. For a good tracking system, these three measures should be equal to zero.

Now, we define average error in estimated height of optimal tracker ($E_{avg}^h$), average error in estimated width of optimal tracker ($E_{avg}^w$) and average error in estimated orientation angle of optimal tracker ($E_{avg}^\theta$), to measure an overall performances of tracking algorithms. These are calculated as

$$E_{avg}^h = \frac{1}{N_{frames}} \sum_{t=1}^{N_{frames}} E^h_t,$$

$$E_{avg}^w = \frac{1}{N_{frames}} \sum_{t=1}^{N_{frames}} E^w_t,$$

$$E_{avg}^\theta = \frac{1}{N_{frames}} \sum_{t=1}^{N_{frames}} E^\theta_t,$$

(21)

where $N_{frames}$ is the total number of frames of a video sequence. For a good tracking system, $E_{avg}^h$, $E_{avg}^w$ and $E_{avg}^\theta$ should be equal to zero.

Another composite measure called as matching score ($E_{ms}^t$) at the $t^{th}$ frame between ground truth tracker ($T^t_g$) and tracker ($T^t$) generated by any tracking algorithm is proposed to evaluate performance of tracking algorithm under such complex environment. Three new proposed measures and two existing measures are combined to calculate matching score as

$$E_{ms}^t = \frac{1}{5} \left[ f^t_c + f^t_h + f^t_w + f^t_\theta + f^t_A \right],$$

(22)

where

$$f^t_c = \frac{1}{1+E^t_C},$$

(23)

$$f^t_h = \frac{1}{1+E^t_h},$$

(24)

$$f^t_w = \frac{1}{1+E^t_w},$$

(25)

$$f^t_\theta = \frac{1}{1+E^t_\theta},$$

(26)

$$f^t_A = E^t_A$$

(27)

with $E^t_C$, $E^t_h$, $E^t_w$, $E^t_\theta$ and $E^t_A$ are center location, height, width, orientation errors and area overlap at $t^{th}$ frame, respectively.

Now we define average matching score ($E_{avg}^{ms}$) to measure overall performance of tracking algorithm, is calculated as

$$E_{avg}^{ms} = \frac{1}{N_{frames}} \sum_{t=1}^{N_{frames}} E_{ms}^t,$$

(28)

where $N_{frames}$ is the total number of frames of a video sequence. For a good tracking system, $E_{avg}^{ms}$ should be equal to 1.
| Sequence  | Area Under The Curve (AUC): Success Plot† | Area Under The Curve (AUC): Success Plot of Matching Score ↑ |
|----------|------------------------------------------|------------------------------------------------------|
| Bag      | 0.293 0.195 0.203 0.297 0.142 0.269 0.294 0.277 0.376 0.295 | 0.293 0.195 0.203 0.297 0.142 0.269 0.294 0.277 0.376 0.295 |
| Basketball | 0.269 0.315 0.178 0.329 0.308 0.329 0.261 0.307 0.374 0.327 | 0.269 0.315 0.178 0.329 0.308 0.329 0.261 0.307 0.374 0.327 |
| Blanket | 0.358 0.406 0.328 0.337 0.273 0.374 0.309 0.277 0.244 0.318 | 0.358 0.406 0.328 0.337 0.273 0.374 0.309 0.277 0.244 0.318 |
| Bmx | 0.071 0.138 0.176 0.100 0.125 0.173 0.154 0.151 0.110 0.144 | 0.071 0.138 0.176 0.100 0.125 0.173 0.154 0.151 0.110 0.144 |
| Book | 0.144 0.234 0.181 0.296 0.220 0.318 0.246 0.326 0.322 0.287 | 0.144 0.234 0.181 0.296 0.220 0.318 0.246 0.326 0.322 0.287 |
| Butterfly | 0.124 0.241 0.172 0.193 0.148 0.220 0.188 0.209 0.241 0.209 | 0.124 0.241 0.172 0.193 0.148 0.220 0.188 0.209 0.241 0.209 |
| Crabs1 | 0.208 0.231 0.146 0.213 0.211 0.247 0.244 0.258 0.259 0.282 | 0.208 0.231 0.146 0.213 0.211 0.247 0.244 0.258 0.259 0.282 |
| Dinosaur | 0.254 0.173 0.182 0.190 0.183 0.217 0.292 0.239 0.261 0.275 | 0.254 0.173 0.182 0.190 0.183 0.217 0.292 0.239 0.261 0.275 |
| Fernando | 0.177 0.156 0.149 0.166 0.193 0.184 0.171 0.169 0.219 0.162 | 0.177 0.156 0.149 0.166 0.193 0.184 0.171 0.169 0.219 0.162 |
| Fish2 | 0.206 0.243 0.195 0.236 0.236 0.239 0.233 0.299 0.299 0.210 | 0.206 0.243 0.195 0.236 0.236 0.239 0.233 0.299 0.299 0.210 |
| Girl | 0.211 0.223 0.136 0.222 0.214 0.247 0.244 0.258 0.258 0.214 | 0.211 0.223 0.136 0.222 0.214 0.247 0.244 0.258 0.258 0.214 |
| gymnastics2 | 0.160 0.236 0.302 0.222 0.231 0.286 0.244 0.258 0.259 0.282 | 0.160 0.236 0.302 0.222 0.231 0.286 0.244 0.258 0.259 0.282 |
| hand | 0.251 0.275 0.276 0.199 0.212 0.250 0.288 0.259 0.338 0.279 | 0.251 0.275 0.276 0.199 0.212 0.250 0.288 0.259 0.338 0.279 |
| helicopter | 0.328 0.302 0.254 0.222 0.197 0.277 0.389 0.235 0.369 0.319 | 0.328 0.302 0.254 0.222 0.197 0.277 0.389 0.235 0.369 0.319 |
| iceskater2 | 0.166 0.221 0.176 0.217 0.236 0.247 0.211 0.242 0.274 0.260 | 0.166 0.221 0.176 0.217 0.236 0.247 0.211 0.242 0.274 0.260 |
| Matrix | 0.161 0.251 0.254 0.285 0.251 0.353 0.234 0.295 0.268 0.209 | 0.161 0.251 0.254 0.285 0.251 0.353 0.234 0.295 0.268 0.209 |
| Motocross2 | 0.193 0.188 0.101 0.146 0.091 0.189 0.203 0.157 0.116 0.163 | 0.193 0.188 0.101 0.146 0.091 0.189 0.203 0.157 0.116 0.163 |
| Shaking | 0.096 0.301 0.374 0.357 0.220 0.321 0.364 0.293 0.417 0.360 | 0.096 0.301 0.374 0.357 0.220 0.321 0.364 0.293 0.417 0.360 |
| Soccer1 | 0.186 0.229 0.349 0.308 0.236 0.369 0.209 0.265 0.314 0.340 | 0.186 0.229 0.349 0.308 0.236 0.369 0.209 0.265 0.314 0.340 |
| Traffic | 0.392 0.296 0.508 0.389 0.434 0.471 0.469 0.467 0.519 0.504 | 0.392 0.296 0.508 0.389 0.434 0.471 0.469 0.467 0.519 0.504 |

Table 1 Shows AUC of the success plots of the existing measures. The bold value indicates the best tracker.

Table 2 Shows AUC of the success plots of new measure - matching score. The bold value indicates the best tracker.
To measure the performance of an algorithm on a sequence of frames, we count the number of successful frames whose matching score ($E_{ms}$) is larger than a given threshold value. The matching score plot shows the ratios of successful frames at the thresholds varied from 0 to 1. The matching rate value at a specific threshold (e.g., 0.5) may not always be a good representative for evaluating the trackers’ performance. Instead of that, the area under the curve (auc) of each matching score plot is considered to evaluate the performance of the tracking algorithms.

7 Experiments

7.1 Dataset

We use entire vot-2018 dataset for our experiments. VOT-2018 dataset contains 60 video sequences. The target in the sequences is annotated by a rotated bounding box and all sequences are per-frame annotated by the following visual attributes: (i) occlusion, (ii) illumination change, (iii) motion change, (iv) size change, and (v) camera motion. Frames that do not correspond to any of the five attributes are denoted as (vi) unassigned. All these complex factors make object tracking difficult on this dataset. If a tracker fails to track object in a frame, we reset the tracker to 0 (as oppose to vot protocol). We only consider one-pass evaluation (ope) to draw the plots to compare performances of all the trackers with respect to both existing and the proposed measures.

7.2 Baselines

We use various state-of-the-art trackers on vot-2018 dataset such as ant\textsuperscript{1}[9], kcf\textsuperscript{2}[6], asm\textsuperscript{3}[14], updt\textsuperscript{3}[3], ladcf\textsuperscript{4}[16], meem\textsuperscript{5}[19], rco\textsuperscript{1}[9], sa-siam-r\textsuperscript{1}[9], sa-siam-p\textsuperscript{1}[9], and siamrpn\textsuperscript{10} for our experiments. Among them, sa-siam-r, sa-siam-p, and siamrpn apply siamese networks, ladcf, rco, kcf, and updt are developed based on discriminative correlation filters, asm applies mean shift, ant uses optical flow, and meem is based on support vector machine.

7.3 Quantitative Results Analysis

Tables\textsuperscript{1} and\textsuperscript{2} show auc of success plots of the existing measure and the proposed measure, respectively. The bold value indicates the best tracker among the consid- ered state-of-the-art trackers. Tables\textsuperscript{3} and\textsuperscript{4} present precision value at threshold 20 pixels, success rate at threshold 0.5, and new success rate at threshold 0.5. Bold values in the tables indicate the best trackers. According to auc of success plots of existing measures, SiamRPM performs the best on 10 sequences among randomly selected 20 sequences. In case of average of 60 sequences, SiamRPM also performs
Table 3: Shows precision value at threshold (20 pixels). The bold value indicates the best tracker.

| Sequence | ANT | ASMS | KCF | LAD | DCF | MEEM | RCO | OSASiam-P | Siam-P | Siam-R | SiamRPN | CPDT |
|----------|-----|------|-----|-----|-----|------|-----|-----------|-------|--------|---------|------|
| Bag      | 0.907 | 0.914 | 0.538 | 0.989 | 0.804 | 0.992 | 0.885 | 0.989 | 0.934 | 0.990 |
| Basketball | 0.933 | 0.941 | 0.987 | 0.996 | 0.987 | 0.987 | 0.991 | 0.987 | 0.973 | 0.889 |
| Blanket  | 0.263 | 0.237 | 0.316 | 0.237 | 0.263 | 0.263 | 0.145 | 0.211 | 0.197 | 0.158 |
| Book     | 0.360 | 0.600 | 0.383 | 0.909 | 0.611 | 0.909 | 0.646 | 0.783 | 0.817 | 0.806 |
| Butterfly| 0.291 | 0.907 | 0.318 | 0.323 | 0.325 | 0.649 | 0.358 | 0.589 | 0.709 | 0.609 |
| Crabs1   | 0.381 | 0.713 | 0.250 | 0.363 | 0.350 | 0.456 | 0.588 | 0.594 | 0.719 | 0.806 |
| Dinosaur | 0.715 | 0.248 | 0.267 | 0.684 | 0.340 | 0.494 | 0.809 | 0.638 | 0.813 | 0.985 |
| Fernando | 0.507 | 0.223 | 0.562 | 0.627 | 0.647 | 0.678 | 0.619 | 0.363 | 0.634 | 0.606 |
| Fish2    | 0.571 | 0.845 | 0.259 | 0.761 | 0.371 | 0.803 | 0.800 | 0.874 | 0.819 | 0.855 |
| Girl     | 0.916 | 0.921 | 0.313 | 0.729 | 0.849 | 0.378 | 0.281 | 0.335 | 0.595 | 0.176 |
| Gymnastics2 | 0.359 | 0.685 | 0.738 | 0.688 | 0.729 | 0.738 | 0.692 | 0.796 | 0.671 | 0.725 |
| Hand     | 0.978 | 0.719 | 0.697 | 0.873 | 0.655 | 0.742 | 0.828 | 0.757 | 0.865 | 0.933 |
| Helicopter | 0.637 | 0.863 | 0.595 | 0.472 | 0.301 | 0.623 | 0.643 | 0.836 | 0.879 | 0.675 |
| Iceskater2 | 0.402 | 0.454 | 0.266 | 0.521 | 0.646 | 0.897 | 0.636 | 0.761 | 0.714 | 0.829 |
| Matrix   | 0.480 | 0.770 | 0.650 | 0.850 | 0.850 | 0.990 | 0.730 | 0.830 | 0.760 | 0.910 |
| Motocross2 | 0.098 | 0.148 | 0.000 | 0.066 | 0.033 | 0.098 | 0.098 | 0.098 | 0.213 | 0.197 |
| Shaking  | 0.660 | 0.488 | 0.858 | 0.939 | 0.923 | 0.997 | 0.995 | 0.997 | 0.997 | 0.997 |
| Soccer1  | 0.513 | 0.515 | 0.865 | 0.918 | 0.837 | 0.954 | 0.546 | 0.819 | 0.859 | 0.959 |
| Traffic  | 0.953 | 0.613 | 0.993 | 0.995 | 0.995 | 0.995 | 0.995 | 0.995 | 0.995 | 0.995 |
| Average over 60 sequences | 0.576 | 0.617 | 0.519 | 0.707 | 0.578 | 0.730 | 0.664 | 0.707 | 0.758 | 0.758 |

Table 4: Shows success rate at threshold (0.5) in the success plot of the existing measure. The bold value indicates the best tracker.
Table 5 Shows the success rate at threshold (0.5) in the success plots of the proposed measure — matching score. The bold value indicates the best tracker.

7.4 Qualitative Results Analysis

Figures 6, 7, 8, and 9 show the tracking results (selected few frames) of vot-2018 dataset obtained using the considered baselines — ANT, ASMS, KCF, LADCF, MEEM, RCO, SA-Siam-P, SA-Siam-R, SiamRPN, and UPDT. Object in Bag, Bmx, Book, Butterfly, Dinosaur, Fernando, Fish2, Gymnastics2, Hand, Iceskater2, and Motocross2 changes scale and orientation. Basketball, Fish2, Girl, and Soccer1 sequences contain occlusions. Matrix and Shaking sequences undergo illumination variation. Object changes its scale in Blanket, Crabs1, and Helicopter sequences. For scaled and oriented object in the sequences such as Butterfly, Dinosaur, Fernando, Fish2, Gymnastics2, Hand, Iceskater2, and Motocross2 changes scale and orientation, Basketball, Fish2, Girl, and Soccer1 sequences contain occlusions. Matrix and Shaking sequences undergo illumination variation. Object changes its scale in Blanket, Crabs1, and Helicopter sequences. For scaled and oriented object in the sequences such as Butterfly, Dinosaur, Fernando, Fish2, Gymnastics2, Hand, Iceskater2, and Motocross2, the trackers — SA-Siam-P, SA-Siam-R, SiamRPN, and UPDT performs better than other trackers. SiamRPN, ANT, RCO, and SA-Siam-R obtain robust tracking results on Basketball, Girl, Soccer1, and Fish2 sequences, respectively. 

7.5 Analysis of Precision and Success Plots

The success plot shows the ratio of successful frames at threshold varied from 0 to 1. Sometimes, the success plots and the precision plots make a different conclusion
Fig. 6 Illustrates the tracking results of ten state-of-the-art trackers on Bag, Basketball, Blanket, Bmx, Book, and Butterfly sequences. Each colored rectangle corresponds to each tracker’s output.
Fig. 7 Illustrates the tracking results of ten state-of-the-art trackers on Crabs1, Dinosaur, Fernando, Fish2, Girl, and Gymnastics2 sequences. Each colored rectangle corresponds to each tracker’s output.
Fig. 8 Illustrates the tracking results of ten state-of-the-art trackers on Hand, Helicopter, Iceskater2, Matrix, Motocross2, and Shaking sequences. Each colored rectangle corresponds to each tracker’s output.
Fig. 9 Illustrates the tracking results of ten state-of-the-art trackers on Soccer1, and Traffic sequences. Each colored rectangle corresponds to each tracker’s output.

because they measure the tracker’s other characteristics based on various metrics. However, auc score of the success plot interprets the overall performance, which is more accurate than the center location error at the precision threshold value plot [15]. We analyze the tracker performance using auc score of success plot, precision plot, and new success plot. The tracker has a maximum auc score of the success plot indicates the best tracker. Since vot-2018 dataset contains 60 video sequences, it is challenging to analyze the state-of-the-art trackers’ performances on every sequence. For this, we randomly select few (20) video sequences to analyze the existing trackers’ performances based on auc of success, precision, and new success plots. It is also challenging to analyze the trackers’ average performances over vot-2018 dataset, using three plots to correspond to each video sequence of this dataset. For this reason, we use average success, precision, and new success plots over all sequences of this dataset. We analyze the performances of the trackers based on average auc corresponds to these three plots.

Figures 10, 11, 12, and 13 present precision, success, and new success plots correspond to randomly selected video sequences — Bag, Basketball, Blanket, Bmx, Book, Butterfly, Crabs1, Dinosaur, Fish2, Fernando, Girl, Gymnastics2, Hand, Helicopter, Iceskater2, Matrix, Motocross2, Shaking, Soccer1, and Traffic of vot-2018 dataset. For these plots, we use auc score to summarize the performances of the trackers. Tables 1 and 2 highlight auc of the success plots of the existing measure and the proposed measure — matching score. These tables also highlight the auc of average success plots over 60 sequences of vot-2018 dataset. Tables 3, 4, and 5 highlight the precision at threshold 20 pixels, success rate at threshold value 0.5. The precision and success plots make different conclusions for most of the se-
Fig. 10 Illustrates the precision and success plots popularly used for evaluating a tracking algorithm and the success plot based on the proposed new measure - matching score of ten different methods on Bag, Basketball, Blanket, Bmx, Book, and Butterfly sequences. Each plot corresponds to one-pass evaluation (OPE).
Fig. 11 Illustrates the precision and success plots popularly used for evaluating a tracking algorithm and the success plot based on the proposed new measure - matching score of ten different methods on Crabs1, Dinosaur, Fernando, Fish2, Girl, and Gymnastics2 sequences. Each plot corresponds to one-pass evaluation (one).
Fig. 12 Illustrates the precision and success plots popularly used for evaluating a tracking algorithm and the success plot based on the proposed new measure - matching score of ten different methods on Hand, Helicopter, Iceskater2, Matrix, Motocross2, and Shaking sequences. Each plot corresponds to one-pass evaluation (one).

![Graphs showing precision, success, and new success plots for various sequences.](image-url)
Fig. 13 Illustrates the precision and success plots popularly used for evaluating a tracking algorithm and the success plot based on the proposed new measure - matching score of ten different methods on Soccer1 and Traffic sequences. Each plot corresponds to one-pass evaluation (ope).

quences of vot-2018 dataset. We make unique conclusion using new success plots of the proposed measure - matching score.

8 Conclusions

In this paper, we propose a new measure - matching score to better quantify a tracker for tracking scaled and oriented objects. The proposed measure - matching score is a combination of two existing measures and three new measures. For some sequences containing scaled and oriented objects, two existing measures make two different remarks. It is very difficult to make a conclusion from these remarks. On the other hand, the proposed measure makes a concrete remark which highlights the best tracker. The sufficient examples and results conclude that the proposed measure is more accurate than the existing one to quantify a tracker for tracking scaled and oriented objects.
9 Compliance with Ethical Standards

9.1 Conflict of Interest

All authors declare that they have no conflicts of interest.

9.2 Ethical Approval

This article does no contain any studies with human participants or animals performed by any of the authors.

References

1. Babenko, B., Yang, M., Belongie, S.: Robust object tracking with online multiple instance learning. IEEE Transactions on Pattern Analysis and Machine Intelligence 33(8), 1618–1632 (2011)

2. Baumann, A., Boltz, M., Elbing, J., Koenig, M., Locs, H.S., Merkel, M., Niem, W., Warzelhan, J.K., Yu, J.: A review and comparison of measures for automatic video surveillance systems. EURASIP Journal on Image and Video Processing 2008(824726), 1–30 (2008)

3. Bhat, G., Johnander, J., Danelljan, M., Shahbaz Khan, F., Felsberg, M.: Unveiling the power of deep tracking. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 483–498 (2018)

4. Čehovin, L., Leonardis, A., Kristan, M.: Visual object tracking performance measures revisited. IEEE Transactions on Image Processing 25(3), 1261–1274 (2016)

5. Flaz, M., Mahmood, A., Javed, S., Jung, S.K.: Handcrafted and deep trackers: Recent visual object tracking approaches and trends. ACM Computing Surveys (CSUR) 52(2), 43 (2019)

6. Henriques, J.F., Caseiro, R., Martins, P., Batista, J.: High-speed tracking with kernelized correlation filters. IEEE transactions on pattern analysis and machine intelligence 37(3), 583–596 (2014)

7. Henriques, J.F., Caseiro, R., Martins, P., Batista, J.: High-speed tracking with kernelized correlation filters. IEEE Transactions on Pattern Analysis and Machine Intelligence 37(3), 583–596 (2015)

8. Kasturi, R., Goldgof, D., Soundararajan, P., Manohar, V., Garofolo, J., Bowers, R., Boonstra, M., Kozhova, V., Zhang, J.: Framework for performance evaluation of face, text, and vehicle detection and tracking in video: Data, metrics and protocol. IEEE Transactions on Pattern Analysis and Machine Intelligence 31(2), 319–336 (2009)

9. Kristan, M., Leonardis, A., Matas, J., Felsberg, M., Pfugfelder, R., Čehovin Zajc, L., Vojir, T., Bhat, G., Lukezic, A., Eldesokey, A., et al.: The sixth visual object tracking vot2018 challenge results. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 0–0. Springer (2018)

10. Li, B., Yan, J., Wu, W., Zhu, Z., Hu, X.: High performance visual tracking with siamese region proposal network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 8971–8980 (2018)

11. Li, P., Wang, D., Wang, L., Lu, H.: Deep visual tracking: Review and experimental comparison. Pattern Recognition 76, 323–338 (2018)

12. Li, X., Hu, W., Shen, C., Zhang, Z., Dick, A., Hengel, A.V.D.: A survey of appearance models in visual object tracking. ACM Transactions on Intelligent Systems and Technology 4(4), 1–58 (2013)

13. Smeluders, A.W., Chu, D.M., Cucchiara, R., Calderara, S., Dehghan, A., Shah, M.: Visual tracking: An experimental survey. IEEE Transactions on Pattern Analysis and Machine Intelligence 36(7), 1442–1468 (2014)

14. Vojir, T., Noskova, J., Matas, J.: Robust scale-adaptive mean-shift for tracking. Pattern Recognition Letters 49, 250–258 (2014)
15. Wu, Y., Lim, J., Yang, M.H.: Online object tracking: A benchmark. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2411–2418. IEEE (2013)

16. Xu, T., Feng, Z.H., Wu, X.J., Kittler, J.: Learning adaptive discriminative correlation filters via temporal consistency preserving spatial feature selection for robust visual object tracking. IEEE Transactions on Image Processing 28(11), 5596–5609 (2019)

17. Yilmaz, A., Javed, O., Shah, M.: Object tracking: a survey. ACM Computing Surveys 38(4), 1264–1291 (2006)

18. Yin, F., Makris, D., Velastin, S.A.: Performance evaluation of object tracking algorithms. In: IEEE International Workshop on Performance Evaluation of Tracking and Surveillance, Rio De Janeiro, Brazil, pp. 1–25. Citeseer (2007)

19. Zhang, J., Ma, S., Sclaroff, S.: MEEM: robust tracking via multiple experts using entropy minimization. In: European conference on computer vision, pp. 188–203. Springer (2014)