Evaluation Metrics and Ranking Methodology for SSA Dataset

1 Evaluating detection results for one sequence

Required inputs

- Distance thresholds τ and ϵ, both non-negative, and $0 \leq \epsilon < \tau$.
- Ground truth coordinates $\mathcal{Y} = \{\mathcal{Y}^f\}_{f=1}^5$ in 5 frames, where each $\mathcal{Y}^f = \{\boldsymbol{y}_j^f\}_{j=1}^N$ contains the coordinates of N objects in frame f. Note that N can be 0, in which case $\mathcal{Y}^f$ for all f, and hence $\mathcal{Y}$, are empty.
- Coordinates of detected objects $\mathcal{X} = \{\mathcal{X}^f\}_{f=1}^5$ in 5 frames, where each $\mathcal{X}^f = \{\boldsymbol{x}_i^f\}_{i=1}^{M_f}$ contains the coordinates of M_f objects in frame f. Note that M_f is allowed to vary across frames. Also, M_f can be 0, in which case $\mathcal{X}^f$ is empty.

Matching For a given frame f, a *one-to-one matching* between $\mathcal{X}^f$ and $\mathcal{Y}^f$ is first obtained. Assuming for now $M_f \leq N$, the matching is encapsulated in binary matrix

$$\mathbf{H}^f \in \{0,1\}^{M_f \times N} \tag{1}$$

with the following constraints to the rows and columns

$$\sum_{j=1}^{N} \mathbf{H}_{i,j}^f = 1, \quad \forall i, \quad \text{and} \quad \sum_{i=1}^{M_f} \mathbf{H}_{i,j}^f \leq 1, \quad \forall j. \tag{2}$$

In words, each point in $\mathcal{X}^f$ must be matched uniquely to a point in $\mathcal{Y}^f$; not all points in $\mathcal{Y}^f$ need to be matched to a point in $\mathcal{X}^f$, but those that are matched do so uniquely.

The matching is solved via the *minimum weighted unbalanced assignment problem*

$$\underset{\mathbf{H}^f}{\operatorname{argmin}} \quad \sum_{i=1}^{M_f} \sum_{j=1}^{N} \mathbf{H}_{i,j}^f \delta(\boldsymbol{x}_i^f, \boldsymbol{y}_j^f) \tag{3}$$

subject to the constraints (2), where function δ implements the truncated distance

$$\delta(\boldsymbol{x}_i^f, \boldsymbol{y}_j^f) = \begin{cases} \|\boldsymbol{x}_i^f - \boldsymbol{y}_j^f\|_2, & \text{if } \|\boldsymbol{x}_i^f - \boldsymbol{y}_j^f\|_2 \leq \tau, \\ \ell, & \text{otherwise.} \end{cases} \tag{4}$$

Here ℓ is a sufficiently large positive number, e.g., the diagonal length of the image. The problem can be solved efficiently via the Hungarian algorithm, maximum flow, linear programming, etc. Examples of the assignment problem and one-to-one matching solutions are given in Figure 1.

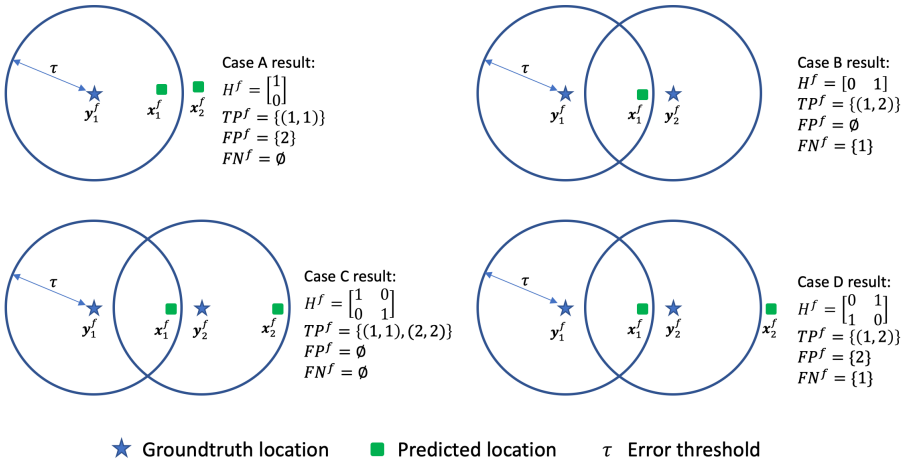

Figure 1: Examples of the assignment problem and one-to-one matching solutions.

If $M_f > N$, the roles of $\mathcal{X}^f$ and $\mathcal{Y}^f$ are swapped and the same problem is solved to perform the matching. This swap is transparent to most algorithms since only simple changes are needed (e.g., adding dummy points at infinity so that $\mathbf{H}^f$ is always square).

If $N = 0$ or $M_f = 0$, then $\mathbf{H}^f = \text{NULL}$.

The matching procedure above is conducted for all frames $f = 1, \ldots, 5$.

True positives, false negatives and false positives If $\mathbf{H}^f$ is not NULL for frame f, the set of correct detections for the frame is

$$\mathcal{TP}^f = \left\{(i,j) \in \{1,\ldots,M_f\} \times \{1,\ldots,N\} \;\middle|\; \mathbf{H}^f_{i,j} = 1, \delta(\boldsymbol{x}^f_i, \boldsymbol{y}^f_j) \leq \tau\right\}. \quad (5)$$

The set of missed detections for the frame is

$$\mathcal{FN}^f = \left\{ j \in \{1,\ldots,N\} \;\middle|\; \left(\sum_{i=1}^{M_f} \mathbf{H}^f_{i,j} = 0\right) \vee \left(\sum_{i=1}^{M_f} \mathbf{H}^f_{i,j}\delta(\boldsymbol{x}^f_i, \boldsymbol{y}^f_j)\right) > \tau \right\}. \quad (6)$$

The set of false alarms for the frame is

$$\mathcal{FP}^f = \left\{ i \in \{1,...,M_f\} \;\middle|\; \left(\sum_{j=1}^{N} \mathbf{H}^f_{i,j} = 0\right) \vee \left(\sum_{j=1}^{N} \mathbf{H}^f_{i,j}\delta(\boldsymbol{x}^f_i, \boldsymbol{y}^f_j)\right) > \tau \right\}. \quad (7)$$

If $\mathbf{H}^f = \text{NULL}$, then

$$\mathcal{TP}^f = \emptyset, \quad \mathcal{FN}^f = \{j\}_{j=1}^N, \quad \text{and} \quad \mathcal{FP}^f = \{i\}_{i=1}^{M_f}. \quad (8)$$

For example, if $M_f = 0$ and $N = 0$ (i.e., no true objects and no detections made), then $\mathcal{TP}^f = \emptyset$, $\mathcal{FN}^f = \emptyset$ and $\mathcal{FP}^f = \emptyset$. As another example, if $M_f > 0$ but $N = 0$ (i.e., no true objects but detections were made), then $\mathcal{TP}^f = \emptyset$, $\mathcal{FN}^f = \emptyset$, $\mathcal{FP}^f = \{i\}_{i=1}^{M_f}$.

The accounting procedure above is conducted for all frames $f = 1, \ldots, 5$. The true positive, false negative and false positive values for the sequence are

$$TP = \sum_{f=1}^{5} |\mathcal{TP}^f|, \quad FN = \sum_{f=1}^{5} |\mathcal{FN}^f|, \quad \text{and} \quad FP = \sum_{f=1}^{5} |\mathcal{FP}^f|. \tag{9}$$

Regression error Another measure of detection accuracy is regression error; specifically the mean squared error (MSE) of localising the objects. Given the results of the procedures above, if not all $\mathcal{TP}^f$, $\mathcal{FN}^f$ and $\mathcal{FP}^f$ are empty for frame f, the sum of squared error (SSE) for frame f is

$$SSE^f = \sum_{(i,j)\in\mathcal{TP}^f} \pi(\boldsymbol{x}_i^f, \boldsymbol{y}_j^f) + \sum_{j\in\mathcal{FN}^f} \tau^2 + \sum_{i\in\mathcal{TP}^f} \tau^2, \tag{10}$$

where

$$\pi(\boldsymbol{x}_i^f, \boldsymbol{y}_j^f) = \begin{cases} 0, & \text{if } \|\boldsymbol{x}_i^f - \boldsymbol{y}_j^f\|_2 \leq \epsilon, \\ \|\boldsymbol{x}_i^f - \boldsymbol{y}_j^f\|_2^2, & \text{otherwise.} \end{cases}. \tag{11}$$

In words, SSE^f accumulates the squared distance (upper bounded by τ^2) between predicted object locations and ground truth object locations in frame f, with a tolerance of ϵ to account for inaccuracies in manual labelling. Further, each missed detection and false alarm respectively contribute a constant squared error of τ^2 to SSE^f.

If $\mathcal{TP}^f$, $\mathcal{FN}^f$ and $\mathcal{FP}^f$ are all empty, then $SSE^f = 0$.

The SSE for the sequence is thus

$$SSE = \sum_{f=1}^{5} SSE^f, \tag{12}$$

and the MSE for the sequence is

$$MSE = \frac{SSE}{TP + FN + FP}. \tag{13}$$

(If $SSE = 0$ we define $MSE = 0$.)

2 Evaluating detection results for whole dataset

Let there be K sequences of 5 frames each in the whole dataset.

Denote by TP_k the true positive value for the k-th sequence computed according to (9) (similarly for FN_k and FP_k). The overall precision is

$$P = \frac{\sum_{k=1}^{K} TP_k}{\sum_{k=1}^{K} TP_k + FP_k}; \tag{14}$$

the overall recall R is

$$R = \frac{\sum_{k=1}^{K} TP_k}{\sum_{k=1}^{K} TP_k + FN_k}; \tag{15}$$

and the F_1 score is

$$F_1 = 2\frac{PR}{P+R}. \tag{16}$$

Denote by SSE_k the SSE for the k-th sequence computed according to (12). The overall regression MSE is thus

$$MSE = \frac{\sum_{k=1}^{K} SSE_k}{\sum_{k=1}^{K} TP_k + FN_k + FP_k}. \tag{17}$$

3 Ranking methodology

Methods will first be ranked using F_1 (16). In the event of ties (i.e., a number of methods having the same F_1), the methods will be further ranked using regression MSE (17).