mvHOTA: A multi-view higher order tracking accuracy metric to measure temporal and spatial associations in multi-point tracking

Lalith Sharan[a,b 000-0003-0835-042X], Halvar Kelm[a], Gabriele Romano[a], Matthias Karck[a], Raffaele De Simone[a], Sandy Engelhardt[a,b 0000-0001-8816-7654]

[a]Department of Cardiac Surgery, Heidelberg University Hospital, Heidelberg; [b]DZHK (German Centre for Cardiovascular Research), partner site Heidelberg/Mannheim

ARTICLE HISTORY
Compiled September 15, 2022

ABSTRACT
Multi-point tracking is a challenging task that involves detecting points in the scene and tracking them across a sequence of frames. Computing detection-based measures like the F-measure on a frame-by-frame basis is not sufficient to assess the overall performance, as it does not interpret performance in the temporal domain. The main evaluation metric available comes from Multi-object tracking (MOT) methods to benchmark performance on datasets such as KITTI with the recently proposed *higher order tracking accuracy* (HOTA) metric, which is capable of providing a better description of the performance over metrics such as MOTA, DetA, and IDF1. While the HOTA metric takes into account temporal associations, it does not provide a tailored means to analyse the spatial associations of a dataset in a multi-camera setup. Moreover, there are differences in evaluating the detection task for points when compared to objects (point distances vs. bounding box overlap). Therefore in this work, we propose a multi-view higher order tracking metric *mvHOTA* to determine the accuracy of multi-point (multi-instance and multi-class) tracking methods, while taking into account temporal and spatial associations. mvHOTA can be interpreted as the geometric mean of detection, temporal, and spatial associations, thereby providing equal weighting to each of the factors. We demonstrate the use of this metric to evaluate the tracking performance on an endoscopic point detection dataset from a previously organised surgical data science challenge. Furthermore, we compare with other adjusted MOT metrics for this use-case, discuss the properties of mvHOTA, and show how the proposed *multi-view Association* and the *Occlusion index (OI)* facilitate analysis of methods with respect to handling of occlusions. The code is available at `https://github.com/Cardio-AI/mvhota`.

KEYWORDS
Evaluation metrics; Point detection; Tracking; Multi-View

1. Introduction

Recent seminal works in our field by the MICCAI Special Interest Group Biomedical Image Analysis Challenges (SIG-BIAC) (Reinke et al. 2021; Maier-Hein et al. 2018) have emphasized the considerable impact of evaluation metrics in benchmarking and ranking the performance of different machine learning methods. The authors recommend usage of orthogonal performance criteria and careful interpretation of the results with respect to the task definition. However, little emphasis so far has been

CONTACT Lalith Sharan Email: lalithnag.sharangururaj@med.uni-heidelberg.de

given to metrics that consider a temporal correlation of data. Given that this is a key aspect in Surgical Data Science (SDS), where image modalities such as ultrasound, fluoroscopy or endoscopy are traditionally temporally resolved to provide real-time guidance, we found that it is interesting to our field to focus on this in more detail. Key use-cases in the domain of SDS require the tracking of objects over time, e.g., surgical instruments, catheters and other equipment, anatomical landmarks, pathological structures, etc. Specific applications are mitral valve leaflet tracking from echocardiography (Chandra et al. 2020), or real time tracking of aortic valve landmarks from fluoroscopy (Karar, Noack, Kempfert, Falk, and Burgert Karar et al.). Furthermore, many detection-based methods exist, such as endoscopic surgical action triplet detection (Zia et al. 2022), polyp detection (Brandao et al. 2018) and artefact detection (Yin et al. 2022). However, in these methods, frame-level metrics such as the F-score, *precision*, and *recall* are computed for the frames of the sequence. Besides, one can differentiate between single- and multi-view settings, where the same scene or object is captured from different angles. This is typically the case in stereo-endoscopy, where spatial associations need to be identified between multiple views. There is hitherto little emphasis given to the robustness of both temporal and spatial consistency of the detection-based results. In this work, we extend a Multi-object tracking (MOT) benchmark metric for the point tracking use-case to account for temporal and spatial associations, and analyse the results on an endoscopic dataset.

In particular, the use-case of point tracking forms a special case of a Multi-object tracking (MOT) task, where several points should be identified. They can either represent multiple instances of the same class (e.g., the same type of cell), or represent multiple classes (e.g., the tip of different surgical instruments). We refer to it as a *multi-point tracking* problem. In endoscopy, multi-point tracking is relevant for applications such as respiratory motion estimation (Silverstein and Snyder 2018), and surgical suture detection (Sharan et al. 2021), besides having more general applications in the computer vision domain spanning motion tracking in traffic scenes (Geiger et al. 2012), and pose estimation (Cao et al. 2019). In the context of mitral valve repair, a minimally invasive surgery of the heart valve, the detection and tracking of entry and exit points of sutures from an endoscopic image is a useful task for understanding a surgical scene. It enables a potential comparison between different skill levels, analysis of suture configurations, and can function as markers for augmented reality visualisations (Engelhardt et al. 2014). However, multi-point tracking is a challenging task due to objects moving in and out of the scene during a surgery, leading to occlusions in the temporal domain. Furthermore in a multi-view setting, the same point can be contained in all perspectives or might be occluded in some. In reality a point may be occluded from one frame to another, may move out of the scene and reappear, or be visible in only one of the views. This is crucial in applications that involve keypoint matching or point cloud generation. While multiple methods address one or more of these issues, the proper evaluation of these methods is difficult in more complex scenarios, for all involved subtasks of finding temporal and spatial associations, *and* detection of the points themselves. Hence, developing a single metric which treats all of these subtasks as equally important is crucial.

Although multi-point tracking is a closely related task to, and can be considered a special case of a multi-object tracking task (Dendorfer et al. 2019; Geiger et al. 2012), important differences exist in evaluating the intrinsic detection task. In MOT, bounding boxes are predicted for each object, where an *Intersection over Union* (IoU) is typically used to compute the overlap of a source and target bounding box (see Fig. 1(Ia)). The IoU has the value 0 when there is no overlap, and 1 when there is

complete overlap. Then, a matching algorithm is employed to match the ground-truth labels and predictions. The matches obtained through this algorithm are then filtered with a threshold α, where the label-prediction pairs with an overlap greater than α are considered a successful match. The algorithm maximises the global similarity and the number of true positive detections (Luiten et al. 2021). In contrast in point detection, the minimum distance between two points is 0, but the maximum distance is theoretically unbounded. Practically, the maximum distance between two points in an image is bounded by the image diagonal, e.g., as shown by Koehler et al. (2022). Moreover, unlike bounding boxes, comparison in the case of point detection is equivalent to considering a radius centered around the point (see Fig. 1(Ib)).

The previously proposed *Higher Order Tracking Accuracy* (HOTA) metric (Luiten et al. 2021) equally weights the detection and temporal association, and is used by the KITTI (Geiger et al. 2012) and MOT (Dendorfer et al. 2019) benchmarks to evaluate submitted methods. In this work, we extend the HOTA metric (Luiten et al. 2021) to evaluate the multi-view higher order tracking accuracy (*mvHOTA*) by incorporating both temporal and spatial associations (see Fig. 1(II)). This is realized by computing a *multi-view Association (mvAssc)* along with detection and temporal association. While previous work (Luiten et al. 2021) presents a trivial extension to a multi-view setup, here we propose a metric that can be used to analyse the handling of both temporal and spatial occlusions. We demonstrate how this method can be used to evaluate and analyse a point tracking method on a surgical stereo-endoscopic dataset of mitral valve repair (Engelhardt et al. 2021). In addition, we introduce the computation of a *Occlusion Index* for multi-camera setups to indicate the amount of occlusions in both the temporal and spatial domains of a dataset. Furthermore, we present an analysis of the metric properties, and a comparison to other MOT metrics for the point tracking use case.

2. Related Work

Currently, multiple metrics exist to evaluate a tracking task from multi-object tracking (MOT) applications (Bernardin and Stiefelhagen 2008; Dave et al. 2020; Ristani et al. 2016). The CLEAR-MOT (Bernardin and Stiefelhagen 2008) metrics were introduced as a standard to evaluate a range of tracking methods, and has been used for years as the standard to benchmark a range of MOT methods. Primarily, the *Multi Object*

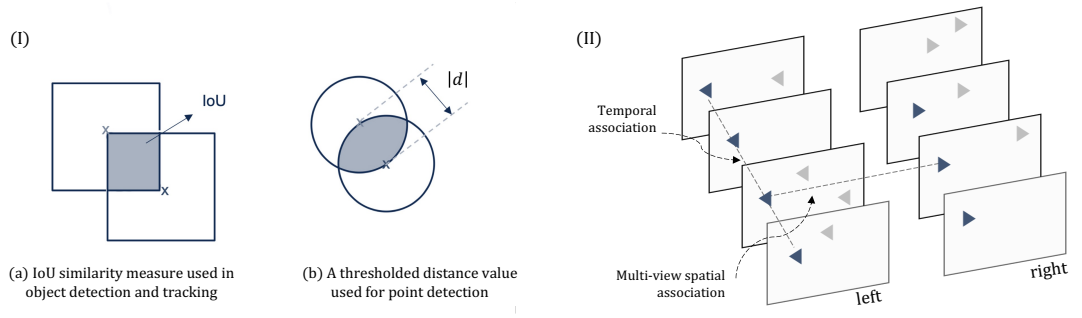

Figure 1.: (Ia) The Intersection over Union (spatial *Jaccard* formulation) used in object detection vs. (Ib) thresholded radius used in point detection. (II) A sample case that shows the detection, temporal and spatial associations.

Tracking Accuracy (MOTA) and *Multi Object Tracking Precision (MOTP)* are used to track the association of detection over time. The $IDF1$ proposed by Ristani et al. (2016) is another metric initially proposed for use in multi-target multi-camera tracking systems and later also adopted for MOT evaluation. Track-mAP is commonly used in benchmarks such as Image-Net (Russakovsky et al. 2015), and TAO (Dave et al. 2020), which requires the use of confidence scores along with the tracker predictions. The recently proposed *higher Order Tracking Accuracy* (HOTA) introduced by Luiten et al. (2021) has replaced many of the multi-object tracking benchmarks, for example in KITTI MOT (Geiger et al. 2012) and the MOTS challenge (Dendorfer et al. 2019). Besides proposing a single unifying metric to measure detection and association, the work analyses the drawbacks of the previously used MOTA and IDF1 metrics in terms of monotonicity and error type differentiability, as explained by Leichter and Krupka (2012).

3. Methods

The goal is to evaluate point detection methods not only on a per-frame basis but across temporal sequences and a multi-camera setup, taking into account both temporal and spatial associations. This means, it is not enough to quantify the detection performance on a frame level (Section 3.2), but also define the associated detection in the temporal domain, which we call temporal associations (Section 3.3), and the detection across multiple views, which we call spatial associations (Section 3.4). In other words, the final metric incorporates how well a point is detected, and additionally tracked over time and space. While the previously proposed *Higher order tracking accuracy (HOTA)* (Luiten et al. 2021) formulates the detection and temporal association for a MOT task, here we propose a metric to also include the multi-view spatial association. The use of a higher-order metric enables us to benchmark and compare methods with an aggregate, while at the same time decompose the metric into different source of tracking errors (Leichter and Krupka 2012) that can be attributed to per-frame detection, temporal and spatial associations.

3.1. Pre-requisites

The computation of this metric requires that we are able to match a predicted point (p_{pred}) to a ground-truth point (p_{gt}) in a frame, and additionally track this particular point over time and the different camera views. This means, a p_{gt} needs a unique ID with which it can be identified over different time steps and across the different camera views. These labels are typically assigned during the annotation process. Similarly, a predicted point p_{pred} requires a consistent ID across a temporal sequence and camera views. This can be directly obtained from the predictions of a multi-view tracking model, or alternatively the predictions can be matched using a temporal matching algorithm (cf. Section 4.2). The spatial associations across the different camera views can be inferred by matching a p_{pred} with a p_{gt} in a view, since we have the spatial matches for the p_{gt} points between the views. With this preparation, we are now ready to formulate the detection, temporal association, and spatial association, and combine it into the *multi-view Higher order tracking accuracy (mvHOTA)*.

3.2. Detection

We first define *True Positive (TP)*, *False Positive (FP)* and *False Negative (FN)* detections for an image. The points that lie within the detection threshold radius of α are classified as TP. The $\{p_{gt}\}$ without a matched prediction are the FN, and the set of predictions $\{p_{pred}\}$ without a matched ground-truth point are the FPs.

For every frame, the set of ground-truth points $\{p_{gt}\}$, and the set of predicted points $\{p_{pred}\}$ are matched using the Hungarian matching algorithm (Kuhn 1955), which maximises the similarity or in this case minimises the distance cost. For the point detection case, the Hungarian algorithm is thresholded with a radius of α, so the algorithm is optimised for number of *TPs* in addition to the computed distance. Typically, a balanced F_1 score (e.g., in Sharan et al. (2021)) is used to combine the TP, FP, and FN points in a single metric. However, the F_1 score is non-monotonic with respect to detections and therefore, we use a *Jaccard* formulation (see Eq.1) to compute the detection accuracy, similar to (Luiten et al. 2021),

$$detAcc = \frac{|TP|}{|TP| + |FP| + |FN|}. \tag{1}$$

It is to be noted that this *Jaccard* score is computed on the set of TPs, FPs, and FNs that are obtained after thresholding. The *Intersection over Union* that is computed as a measure of similarity in an object detection use-case is itself a *Jaccard* metric (Fig. 1(Ia)) where this is referred to as a *spatial Jaccard* index (Luiten et al. 2021).

3.3. Temporal Association

Besides evaluating the detections in every frame, we additionally compute the temporal association as defined for the HOTA metric (Luiten et al. 2021) for each *True Positive* (TP) point. For each TP, the *True Positive Association (TPA)* is the number of points in the sequence having the same ground-truth and prediction IDs (Luiten et al. 2021). The points with the same ground-truth ID but a different prediction ID are counted as FNA, and a same prediction ID with a different ground-truth ID is counted as a FPA. The temporal association (*tempAssc*) is then computed as,

$$tempAssc = \frac{1}{|TP|} \sum_{c \in \{TP\}} \frac{|TPA(c)|}{|TPA(c)| + |FPA(c)| + |FNA(c)|}. \tag{2}$$

The formulation of HOTA weights each *TP* by its *tempAssc* value, to combine the detection accuracy and temporal association in a balanced metric, as detailed in Luiten et al. (2021).

3.4. Multi-view association

As an extension to the previously proposed HOTA metric (Luiten et al. 2021), we introduce the concept of a matched spatial association between multiple views.

The multi-view spatial association is defined for a TP in at least one of the views, where it is classified as a True Positive Correspondence (TPC) if the points containing the same ground-truth ID in each of the views are matched with the same prediction

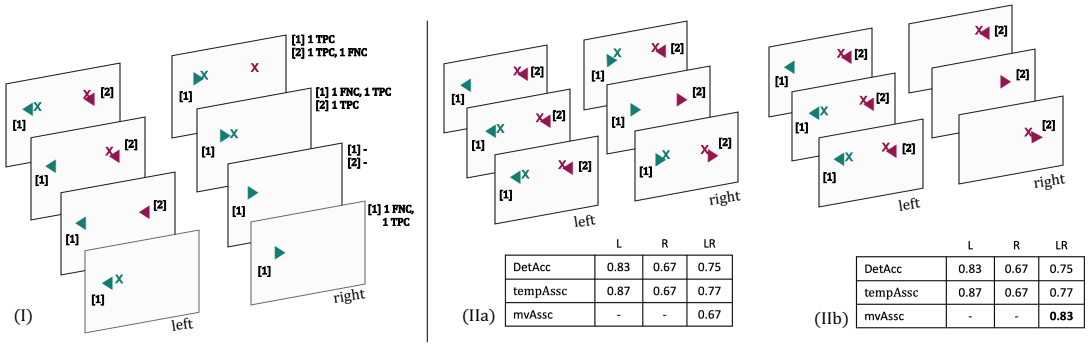

Figure 2.: An extension to the HOTA metric that incorporates a multi-view setup. (I) An illustration of the different cases of multi-view association when ground-truth spatial association exists (Pt. [1]), and does not exist (Pt. [2]). (IIa,b) An example that illustrates a change in the *mvAssc* without a change in detection (*detAcc*) and temporal association (*tempAssc*) (ground-truth from the ◀: left camera view, ▶: right camera view, **x**: predictions)

.

ID. If the same ground-truth ID in a corresponding view is matched with a different prediction ID or if there is no prediction found, it is classified as a False Negative Correspondence (FNC). A False Positive Correspondence (FPC) occurs when the same prediction ID in a corresponding view is either matched with another ground-truth ID or if there is no ground-truth found.

The correspondence accuracy can then be defined as the *Jaccard* score averaged for all TP points for each frame in the sequence, computed as

$$mvAssc = \frac{1}{|TP|} \sum_{c \in \{TP\}} \frac{|TPC(c)|}{|TPC(c)| + |FPC(c)| + |FNC(c)|}. \tag{3}$$

Fig. 2 (I) illustrates the possible scenarios for a stereo setup. For example, point [1] in Fig. 2 (I) is visible in both views. Here, a TPC is defined when a TP is found in both views. An FNC is defined when there are no TP detections found for this p_{gt} in one or both the views. Point [2] in Fig. 2 (I), is present in only one of the views. This is a typical scenario due to occlusions or differences in field of view. Here, a TPC is defined only when a TP detection exists for the p_{gt} in its respective view, and an FPC occurs when the p_{pred} is found in both frames. The formulation of the *mvAssc* similar to the *tempAssc* helps analyse the performance of a model in the context of how well the spatial occlusions are handled. Additionally, separating the spatial from temporal associations, helps measure both the accuracies in a manner that is decomposable into the two different aspects of model performance.

3.5. Multi-view HOTA

We now combine the above mentioned concepts of detection, temporal and spatial associations to formulate the multi-view higher order tracking accuracy:

$$mvHOTA = \sqrt[3]{detAcc \cdot tempAssc \cdot mvAssc} \tag{4}$$

Each of *tempAssc* and *mvAssc* are averaged over the TP detections, and therefore can be seen as augmenting each TP with the *tempAssc* and *mvAssc* instead of double counting the errors. Furthermore, *mvHOTA* can be interpreted as the geometric mean of the detection, temporal association, and spatial association, thereby providing equal weighting to each of the factors. Additionally, the matching can be performed to optimise the number of True Positive predictions that contain True Positive Correspondences, in addition to minimising the distance cost and maximising the temporal associations.

3.6. Occlusion Index (OI)

Besides *mvHOTA*, we propose an index to quantify the extent of occlusions in both the temporal and spatial domains. The *OI* is computed as,

$$OI_v = 1 - \frac{1}{N}\sum_{f=1}^{N} p_f^v . c_f \quad \text{where} \quad c_f = \frac{1}{M}\sum_{v=1}^{M} p_v^f \tag{5}$$

where p_v^f is used while summing all the views in which a point is found, for a particular frame f, where $p_v^f = 1$ if the point is found in view v and 0 otherwise. Similarly, p_f^v is used while summing all the frames in a particular view v. Essentially, to compute the OI_v for a view v, a weighted average is computed for all the frames f, where the weighting factor for each point is the fraction of views in which this point exists (c_f). This index helps in analysing the extent of spatio-temporal occlusions for different objects in the scene. The temporal occlusion ($tempOI$) and multi-view occlusion ($mvOI$) can be easily obtained from Eqn. 5 by setting $p_v^f = 1$ for views v, and by setting $p_f^v = 1$ for frames f respectively. An example of OI computed for the endoscopic dataset used in this work is shown in Fig. 4.

3.7. Toy example

To demonstrate the need for *mvHOTA* in contrast to a trivial multi-view extension of MOT metrics, we construct a toy example to show how *mvHOTA* can accommodate the spatial association from different views while related MOT metrics remain unchanged. Fig. 2 (II) represents an example case in point detection with two ground-truth IDs. Fig. 2 (IIa) contains two ground-truth points and their respective predictions. In Fig. 2 (IIb), we now remove one of the ground-truth IDs and the associated predictions. This changes the spatial association without affecting the detection and temporal association. Here, *mvHOTA* is able to accommodate this change, while computing the MOT metrics for each view (see Tab. 1) does not account for the spatial association. A further example of *mvHOTA* applied to more than two views is shown in Figure 3.

Table 1.: Comparison of MOT metrics on the two cases of the toy example as shown in Fig. 2 (II). Change in *mvHOTA* is highlighted in bold.

Case	MOTA	IDF1	F1	HOTA	mvHOTA
(II)a.	0.67	0.76	0.85	0.76	0.73
(II)b.	0.67	0.76	0.85	0.76	**0.78**

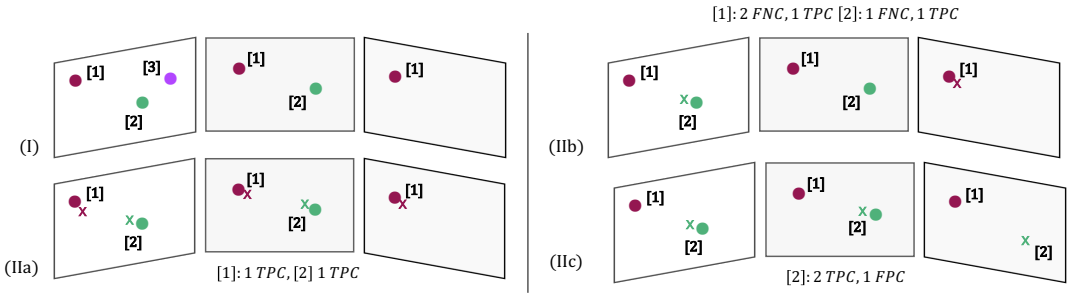

Figure 3.: An example illustration of a 3-view setup with ground-truth (•) and predicted(x) points. (I) shows how the multi-view spatial association is calculated for each point in different scenarios. (IIa), (IIb), and (IIc) show the calculation of TPC, FPC, and FNC in each case, which is then used to compute the respective *mvAssc*.

3.8. Properties

For a single-view, single-instance tracking scenario, *mvHOTA* simplifies into $|TP|/(|TP|+|FP|+|FN|)$. Luiten et al. (2021) showed that HOTA is both monotonic and differentiable into the different error types, as proposed by Leichter and Krupka (2012). *mvHOTA* follows a similar formulation while including the spatial associations between the multiple views. Furthermore, the interpretation as the geometric mean ensures the metric is balanced between the different aspects of detection, temporal and multi-view spatial associations. The ability to differentiate the metric into different error types (Leichter and Krupka 2012) and analyse their aspects is especially crucial for benchmarking, ranking, and interpreting the performance of the different methods. Evaluating a metric with respect to different aspects of performance not only helps in comparing methods, but also in aligning this comparison towards an application that favours specific aspects of performance. In this regard, Leichter and Krupka (2012) define 5 basic error types for MOT tasks, namely *False Negatives*, *False Positives*, *Fragmentation*, *Mergers*, and *Deviations*. Luiten et al. (2021) showed how for the tracking task HOTA decomposes into the five error types with an equivalent computation of detection recall, detection precision, association recall, association precision, and localisation. *mvHOTA* that is built upon this work can be similarly analysed with respect to the basic error types.

4. Experiments and Results

In this section we demonstrate the application of the proposed metric to evaluate the tracking of points for a stereo-endoscopic use-case in mitral valve repair (Engelhardt et al. 2021).

4.1. Dataset

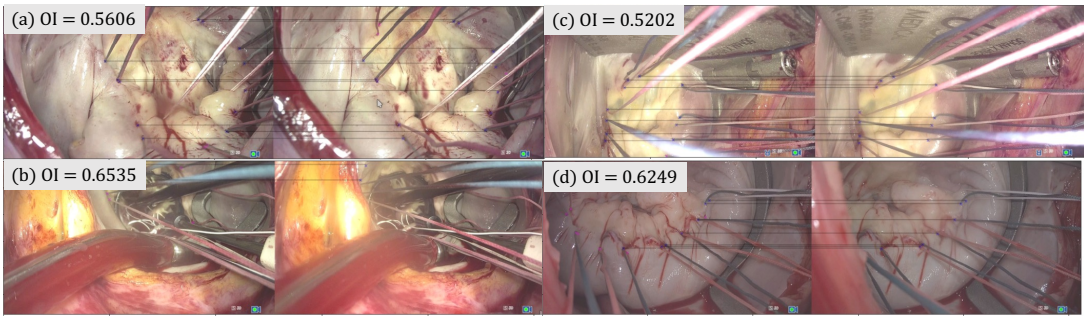

Figure 4.: Sample images from Folds $1-4$ of the mitral valve training dataset with the corresponding left and right views of the stereo-endoscopic setup are shown in (a-d) respectively, with the spatial associations drawn for the labelled suture points. A point can be spatially occluded due to the presence of sutures, instruments, or occluding tissue in the scene. The respective *Occlusion Indices (OIs)* of the surgeries are additionally shown, that indicate the average extent of occlusions in both the temporal and spatial domains.

The mitral valve dataset (Engelhardt et al. 2021; Sharan et al. 2021) comprises images from a stereo-endoscope captured during minimally invasive mitral valve repair surgery. The dataset contains surgeries captured under varying camera angles and illumination. Moreover, objects moving in an out of the scene (surgical instruments, suction pump, ring sizer, etc.) occlude points of interest (entry and exit points of sutures) in both the temporal and spatial domains. A dataset of 4 surgeries was used for training. A sample from each surgery of the training dataset with the left and right views of the multi-view setup is shown in Fig. 4 with the spatial associations highlighted. The *Occlusion index (OI)* is provided for the respective datasets, which indicate the average of extent of occlusion in both the temporal and spatial domains. Furthermore in Fig. 5, the sample trajectories of entry and exit suture points across a temporal sequence is illustrated for Surgery 1 of the training dataset, together with a closer look at one sample of a suture point. Additionally, Fig. 5 also shows the variance of the projected suture clusters for the whole sequence along with the disparity shifts between the corresponding views for the respective suture points.

With this training dataset, the models were trained with leave-one out cross-validation, yielding 4 different models for analysis. The trained models were tested on an external test dataset comprising of 5 challenging surgeries that contain occlusions in both the spatial and temporal domains. Fig. 6(a, b) illustrate the intensity distributions of the training and test dataset respectively. Fig. 6(c) provides an overview of the data split used in the experiments in this work.

The multi-point detection task is to detect the entry and exit points of sutures, which are stitched. The suture points do not occur at anatomically unique locations and the number of suture points vary in each image temporally and between the views. The endoscopic frames are publicly available as part of the AdaptOR challenge (Engelhardt et al. 2021). Other works on this data set have primarily focused on the detection task and treated the stereo-information as two mono instances. The authors have formulated it as a multi-instance heatmap regression problem (Sharan et al. 2021; Stern et al. 2021). Evaluation was performed by computing a balanced F_1 score. A threshold of $6px$ was set as the similarity threshold, as it roughly corresponds to the

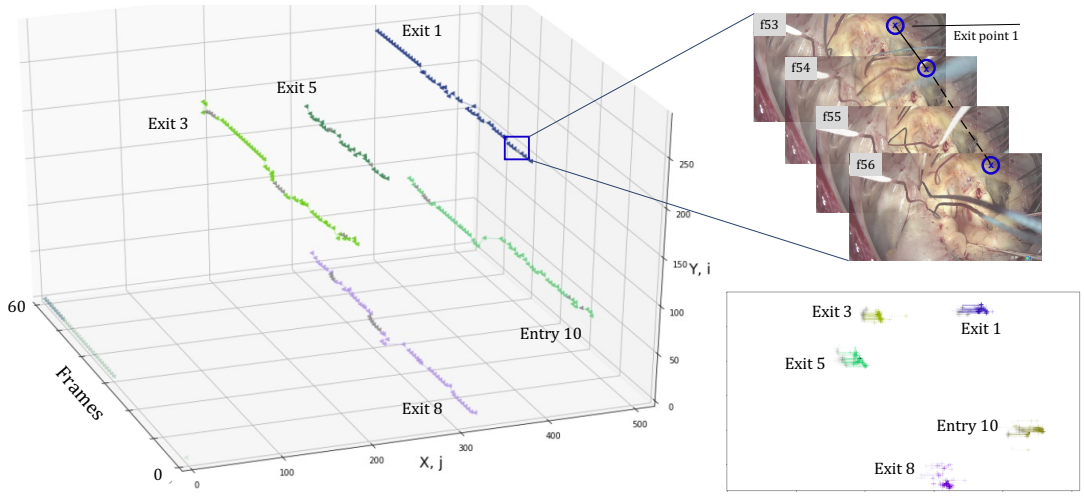

Figure 5.: Sample trajectories of sample entry and exit suture points from Surgery 1 of the training dataset, for the first 60 frames of the temporal sequence. The points plotted in gray represent temporal occlusions, where an interpolated value is plotted in the place of the true suture point location. Additionally, a closeup of Exit point 1 is shown for frames $53 - 56$ where a temporal occlusion due to the presence of white sutures can be seen in frame 55. Additionally, a cluster projection on the x-y plane of the whole sequence of the suture points are presented on the lower right, to illustrate the variance in the movement across the sequence. Furthermore, the disparities with the corresponding view of the stereo-setup are drawn for each of the points.

thickness of the suture in this image resolution. However, a frame level metric such as the F_1 score does not indicate the performance of the method in handling temporal and spatial occlusions. Therefore it is important to compute the respective associations in the temporal and spatial domains to better ascertain the performance of the method.

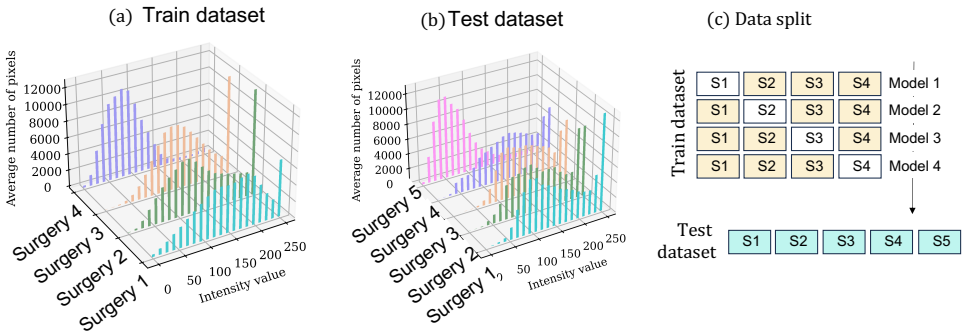

Figure 6.: (a, b) show the intensity distribution of the different surgeries in the train and test datasets respectively. (c) shows the data split used for the experiments.

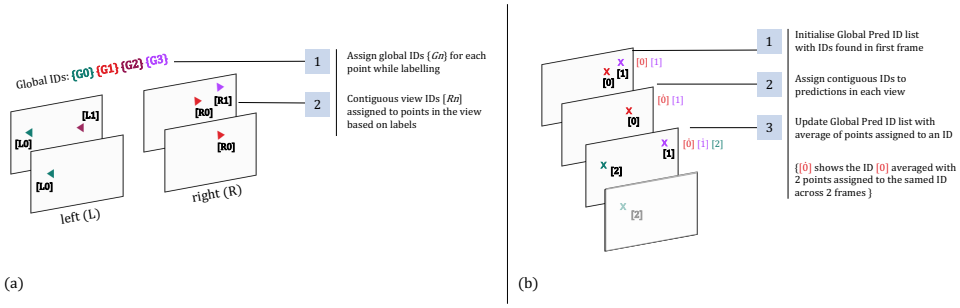

Figure 7.: The use of mvHOTA requires temporally matched IDs assigned to both the ground-truth (◀: left camera view, ▶: right camera view), and predicted (**x**) points. This figure shows how (a) global ground-truth IDs are created based on labels, but are mapped to locally contiguous IDs assigned to each view. (b) Predicted points are temporally matched to a global list for each view, which is averaged after every frame is matched.

4.2. Data preparation

In order to compute the multi-view higher order tracking accuracy for the suture detection method on the mitral valve dataset, we show how we can prepare the dataset. Firstly, the computation of temporal associations requires contiguous ID assignment in each view. This is achieved by mapping the IDs of each view to a global ID list that is common to both views (see Fig. 7 (a)). Secondly, we require unique contiguous IDs for a sequence of images for the predicted points. This is achieved by temporally matching the predicted points with the Hungarian matching algorithm (see Fig. 7 (b)).

4.3. Results

Table 2.: Comparison of MOT metrics for each fold (surgery) of the mitral valve dataset. The MOTA, IDF1, F1, and HOTA metrics are averaged for each view.

Metric	MOTA	IDF1	F1	detAcc	tempAssc	HOTA	mvAssc	mvHOTA
Model 1	-0.0256	0.1559	0.3892	0.2629	0.4109	0.3241	0.7363	0.4172
Model 2	0.0798	0.1815	0.3902	0.2619	0.4040	0.3141	0.7283	0.4104
Model 3	0.0661	0.1687	0.3815	0.2498	0.4686	0.3398	0.7241	0.4339
Model 4	0.0786	0.1572	0.3425	0.2264	0.4098	0.3034	0.7193	0.3868
Mean	0.0497	0.1658	0.3759	0.2503	0.4233	0.3204	0.7270	0.4121

Table 2 presents a comparison of various MOT metrics computed for each surgery of the mitral valve test dataset for the suture tracking task. The various models are trained on different splits of training dataset using leave-one out cross-validation and tested on all the surgeries of the test dataset (c.f. Fig. 6(c)). Besides, the temporal IDs for the predictions were assigned in a post-processing step as described in Section 4.2. It can be seen that although the HOTA and mvHOTA provide similar comparative trends of the different folds, the *mvHOTA* metric provides additional specific information for analysis of the spatial associations in a dataset.

Model 3 has the highest *mvHOTA* score of 0.4339, and the highest *HOTA* score of 0.3398. However, it does not have the highest detection performance, which is reflected in a *detAcc* value of 0.2498 (compared to 0.2629 of Model 1 c.f. Table 2), which is also

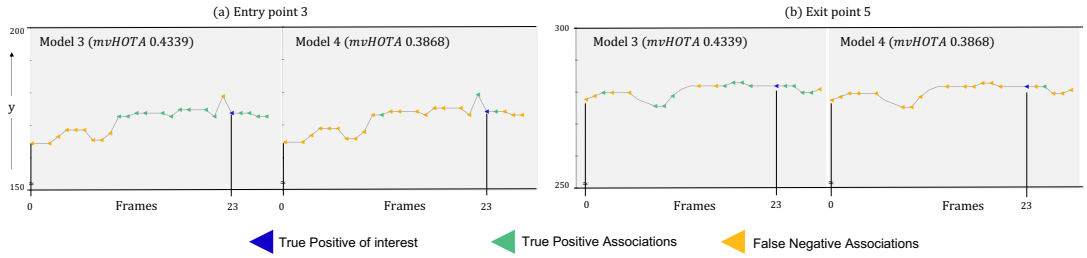

Figure 8.: The association trajectory over time over the first 30 frames, for (a) Entry point 3, and (b) Exit point 5 from Surgery 1 of the test dataset, for the True Positive found in Frame 23. The trajectories of the suture point are projected on the x-t plane, and the True Positive Associations (TPA) and the False Negative Associations (FNA) are shown. Model 3 with $mvHOTA = 0.4339$ outperforms Model 4 with $mvHOTA = 0.3868$.

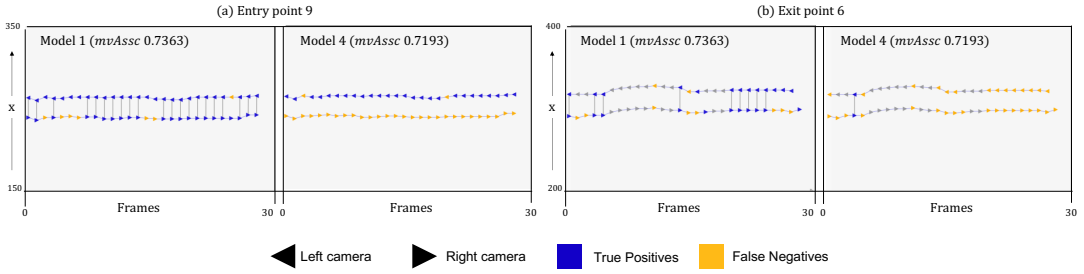

Figure 9.: The trajectory over time of the first 30 frames, for (a) Entry point 9, and (b) Exit point 6 from Surgery 3 of the test dataset. The trajectories of the suture point are projected on the y-t plane. The True Positive Correspondences (TPC) are connected with a line, whereas the False Negative Correspondences (FNC) are not. Model 1 with $mvAssc = 0.7363$ outperforms Model 4 with $mvAssc = 0.7193$.

reflected in the frame level F_1 score of 0.3815 (compared to $F_1 = 0.3902$ for Model 2 c.f. Table 2). In contrast, Model 4 has the lowest *mvHOTA* score of 0.3868 (-0.0471 of Model 3) and the lowest *HOTA* score of 0.3034 (-0.0364 of Model 3). Accordingly, Model 3 has a *tempAssc* of 0.4686, compared to Model 4 with the lowest *tempAssc* of 0.4098 ($+0.0588$ c.f. Table 2). Two such examples, are illustrated in Fig. 8(a) and (b), which indicate the difference in the performance of Model 3 and 4, with respect to temporal associations on the same data, here namely Entry point 3, and Exit point 5 respectively. It can be seen that Model 3 has more temporal associations for a given TP in comparison with Model 4. Similarly, the *mvAssc* of Model 1 is the highest (0.7363), and that of Model 4 the lowest (0.7193, -0.0170 c.f. Table 2). Fig. 9 illustrates this difference, for Entry point 9 (Fig. 9(a)), and Exit point 6 (Fig. 9(b)), which show that Model 1 is able to predict the sutures in both the views more consistently, and as a result has a higher aggregate *mvAssc* compared to Model 4 which has a large number of FNCs Additionally, it can be seen that Model 1 has a negative MOTA score (-0.0256), which means the combined value of FPs, FNs, and *Identity Switches (IDSW)* as defined by (Bernardin and Stiefelhagen 2008) are in this case, more than the number of ground truth detections (Bernardin and Stiefelhagen 2008).

5. Discussion and Conclusion

The proper choice of metrics for a particular task is non-trivial, especially in the case of multi-point detection due to objects moving in and out of the scene, and endoscopic artefacts, leading to spatial and temporal occlusions. Incorporating these associations provides a more complete picture in evaluating the performance of multi-point detection methods. While a range of metrics have been previously proposed to evaluate a MOT task (Bernardin and Stiefelhagen 2008; Dave et al. 2020; Ristani et al. 2016), in this work we propose a metric to track temporal and multi-view spatial associations in multi-point tracking. Beyond the use case of mitral valve repair, different endoscopic datasets, be it laparoscopy (Bodenstedt et al. 2018) or heart surgery (Sharan et al. 2021), contain varying amounts of temporal or spatial occlusions depending on the moving or static nature of the camera and the scene. The *mvHOTA* metric is especially useful in this case, as it enables analysis and benchmarking of different aspects of a model performance, while at the same time can be decomposed into different aspects of tracking for a multi-camera setup.

Alternatively, computing HOTA (Luiten et al. 2021) for each view and averaging them, does not embed the multi-view spatial associations (see Fig. 2 (IIb)). Luiten et al. (2021) suggest a multi-camera extension by stacking the frames of the corresponding view together with the temporal stack to evaluate the total associations for the sequence. However, this precludes an analysis of the temporal and spatial associations in a separate manner while combining them in a single metric.

However, there exist some limitations of *mvHOTA*. Luiten et al. (2021) compute the localisation accuracy for a particular threshold α, to measure the extent of spatial alignment between a ground-truth and predicted point for the object detection case, based on the IoU metric. A similar metric can also be computed for this use-case as the distance cost of the points that lie within the threshold α. However, it is not shown in this work, since the goal is to quantify the detection and tracking performance for a particular threshold. The distance cost between the ground-truth and predicted points is however used to optimise the matching. A future work is to generalise this method to object detection use-cases where the detection and tracking performance across a multi-view setup can be quantified.

Funding

This work was supported in part by Informatics for Life funded by the Klaus Tschira Foundation and the German Research Foundation DFG Project 398787259, DE 2131/2-1 and EN 1197/2-1.).

References

Bernardin, K. and R. Stiefelhagen (2008, December). Evaluating Multiple Object Tracking Performance: The CLEAR MOT Metrics. *EURASIP Journal on Image and Video Processing 2008*(1), 1–10.

Bodenstedt, S., M. Allan, A. Agustinos, X. Du, L. Garcia-Peraza-Herrera, H. Kenngott, T. Kurmann, B. Müller-Stich, S. Ourselin, D. Pakhomov, R. Sznitman, M. Teichmann, M. Thoma, T. Vercauteren, S. Voros, M. Wagner, P. Wochner, L. Maier-Hein, D. Stoyanov, and S. Speidel (2018, May). Comparative evaluation of instrument segmentation and tracking methods in minimally invasive surgery. *arXiv:1805.02475 [cs]*. arXiv: 1805.02475.

Brandao, P., O. Zisimopoulos, E. Mazomenos, G. Ciuti, J. Bernal, M. Visentini-Scarzanella, A. Menciassi, P. Dario, A. Koulaouzidis, A. Arezzo, D. J. Hawkes, and D. Stoyanov (2018, June). Towards a Computed-Aided Diagnosis System in Colonoscopy: Automatic Polyp Segmentation Using Convolution Neural Networks. *Journal of Medical Robotics Research 03*(02), 1840002. arXiv:2101.06040 [cs].

Cao, Z., G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh (2019). Openpose: Realtime multi-person 2d pose estimation using part affinity fields.

Chandra, V., P. G. Sarkar, and V. Singh (2020, January). Mitral Valve Leaflet Tracking in Echocardiography using Custom Yolo3. *Procedia Computer Science 171*, 820–828.

Dave, A., T. Khurana, P. Tokmakov, C. Schmid, and D. Ramanan (2020). TAO: A Large-Scale Benchmark for Tracking Any Object. In A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm (Eds.), *Computer Vision – ECCV 2020*, Volume 12350, pp. 436–454. Cham: Springer International Publishing. Series Title: Lecture Notes in Computer Science.

Dendorfer, P., H. Rezatofighi, A. Milan, J. Shi, D. Cremers, I. Reid, S. Roth, K. Schindler, and L. Leal-Taixe (2019, June). CVPR19 Tracking and Detection Challenge: How crowded can it get? *arXiv:1906.04567 [cs]*. arXiv: 1906.04567.

Engelhardt, S., R. De Simone, N. Zimmermann, S. Al-Maisary, D. Nabers, M. Karck, H.-P. Meinzer, and I. Wolf (2014). Augmented reality-enhanced endoscopic images for annuloplasty ring sizing. In *Augmented Environments for Computer-Assisted Interventions*, pp. 128–137. Springer International Publishing.

Engelhardt, S., A. Mukhopadhyay, R. D. Simone, L. Sharan, A. Stern, J. Brand, and H. Krumb (2021, March). Deep Generative Model Challenge for Domain Adaptation in Surgery 2021. Publisher: Zenodo.

Geiger, A., P. Lenz, and R. Urtasun (2012). Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Karar, M. E., T. Noack, J. Kempfert, V. Falk, and O. Burgert. Real-Time Tracking of Aortic Valve Landmarks Based on 2D-2D Fluoroscopic Image Registration. pp. 4.

Koehler, S., L. Sharan, J. Kuhm, A. Ghanaat, J. Gordejeva, N. K. Simon, N. M. Grell, F. André, and S. Engelhardt (2022, January). Comparison of Evaluation Metrics for Landmark Detection in CMR Images. *arXiv:2201.10410 [cs]*. arXiv: 2201.10410.

Kuhn, H. W. (1955). The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly 2*(1-2), 83–97. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/nav.3800020109.

Leichter, I. and E. Krupka (2012, June). Monotonicity and error type differentiability in performance measures for target detection and tracking in video. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2003–2009. ISSN: 1063-6919.

Luiten, J., A. Osep, P. Dendorfer, P. Torr, A. Geiger, L. Leal-Taixe, and B. Leibe (2021, February). HOTA: A Higher Order Metric for Evaluating Multi-Object Tracking. *International Journal of Computer Vision 129*(2), 548–578. arXiv: 2009.07736.

Ristani, E., F. Solera, R. S. Zou, R. Cucchiara, and C. Tomasi (2016, September). Performance Measures and a Data Set for Multi-Target, Multi-Camera Tracking. *arXiv:1609.01775 [cs]*. arXiv: 1609.01775.

Russakovsky, O., J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei (2015, December). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision 115*(3), 211–252.

Sharan, L., G. Romano, J. Brand, H. Kelm, M. Karck, R. De Simone, and S. Engelhardt (2021, December). Point detection through multi-instance deep heatmap regression for sutures in endoscopy. *International Journal of Computer Assisted Radiology and Surgery 16*(12), 2107–2117.

Silverstein, E. and M. Snyder (2018, March). Comparative analysis of respiratory motion tracking using Microsoft Kinect v2 sensor. *Journal of Applied Clinical Medical Physics 19*(3), 193–204.

Stern, A., L. Sharan, G. Romano, S. Koehler, M. Karck, R. De Simone, I. Wolf, and S. Engelhardt (2021). Heatmap-based 2d landmark detection with a varying number of landmarks. In C. Palm, T. M. Deserno, H. Handels, A. Maier, K. H. Maier-Hein, and T. Tolxdorff (Eds.), *Bildverarbeitung für die Medizin 2021 - Proceedings, German Workshop on Medical Image Computing, Regensburg, March 7-9, 2021*, Informatik Aktuell, pp. 22–27. Springer.

Yin, T.-K., K.-L. Huang, S.-R. Chiu, Y.-Q. Yang, and B.-R. Chang (2022, April). Endoscopy Artefact Detection by Deep Transfer Learning of Baseline Models. *Journal of Digital Imaging*.

Zia, A., X. Liu, K. Bhattacharyya, Z. Wang, M. Berniker, A. Jarc, C. Nwoye, D. Alapatt, A. Murali, S. Sharma, A. Vardazaryan, N. Padoy, B. V. Amsterdam, D. Psychogyios, E. Colleoni, A. Rau, S. Bano, Y. Jin, J. Cartucho, S. Giannarou, S. Ali, Y. Jin, Y. E. López, E. Buc, B. L. Roy, P. Teoule, C. Reissfelder, A. Bailey, Z. Soonawalla, A. Gordon-Weeks, M. Silva, A. Bartoli, T. Roß, A. Reinke, S. Bodenstedt, D. Stoyanov, L. Maier-Hein, and S. Speidel (2022, March). Endoscopic Vision Challenge 2022. Publisher: Zenodo.