Accurate 3D Multi-Marker Tracking in X-ray Cardiac Sequences Using a Two-Stage Graph Modeling Approach

X. Jiang[1], D. Haase[1], M. Körner[1], W. Bothe[2], and J. Denzler[1]

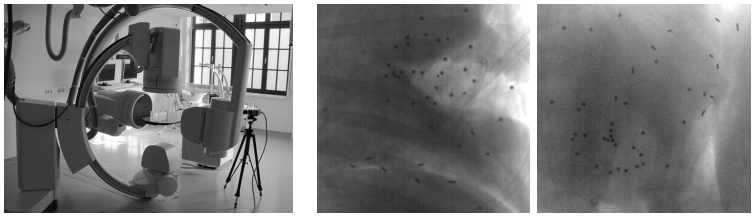
[1] Computer Vision Group, Friedrich Schiller University of Jena
{xiaoyan.jiang,daniel.haase,marco.koerner,joachim.denzler}@uni-jena.de
[2] Department of Cardiothoracic Surgery, University Hospital Jena
bothe@med.uni-jena.de

Abstract. The in-depth analysis of heart movements under varying conditions is an important problem of cardiac surgery. To reveal the movement of relevant muscular parts, biplanar X-ray recordings of implanted radio-opaque markers are acquired. As manually locating these markers in the images is a very time-consuming task, our goal is to automate this process. Taking into account the difficulties in the recorded data such as missing detections or 2D occlusions, we propose a two-stage graph-based approach for both 3D tracklet and 3D track generation. In the first stage of our approach, we construct a directed acyclic graph of 3D observations to obtain tracklets via shortest path optimization. Afterwards, full tracks are extracted from a tracklet graph in a similar manner. This results in a globally optimal linking of detections and tracklets, while providing a flexible framework which can easily be adapted to various tracking scenarios based on the edge cost functions. We validate our approach on an X-ray sequence of a beating sheep heart based on manually labeled ground-truth marker positions. The results show that the performance of our method is comparable to human experts, while standard 3D tracking approaches such as particle filters are outperformed.

Keywords: Multiple object tracking, Directed acyclic graph, Min-cost optimization

1 Introduction

A fully automated system capable of analyzing cardiac movements could significantly help doctors to gain a highly detailed insight into muscular movements under various conditions and to refine surgical strategies for treating heart-related diseases. To analyze heart movements, X-ray recordings are employed in which implanted radio-opaque markers reveal the movement of all relevant cardiac muscles. Fig. 1a shows the biplanar acquisition setup. However, as can be seen in Fig. 1b, X-ray videos of the beating heart containing implanted markers usually have low contrast due to contiguous anatomical structures, and inevitably contain numerous occlusions of the markers. As manually locating these markers is a tedious and time-consuming task, the automatic and accurate tracking and identification of divergently moving markers under severe occlusions is an important, practically relevant, and challenging task.



(a) Acquisition system **(b)** Example images

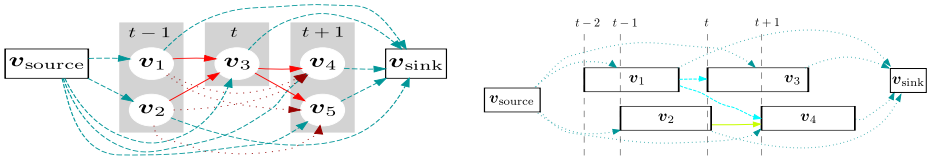
Fig. 1: (a) Biplanar high-speed X-ray acquisition system (Neurostar®, Siemens AG), (b) X-ray images of both camera views showing a sheep heart with implanted markers.

Previous works dealing with heart motion tracking [14,13] involve a lot of manual interactions, yet only sparse marker configurations can be processed. In the computer vision community, multi-object tracking algorithms mainly assume appearance or motion affinity [5,8,19]. However, as is the case in this application, targets are not always distinguishable. Simple online object tracking approaches such as the Kalman [17] or particle filter [5,11] typically fail in such settings due to improper predictions that cause wrong matches between tracks and detections. Local optimization schemes such as the Hungarian algorithm [10], bipartite graph matching [4], or energy minimization [1] consider the best assignment between tracks and detections or detections and detections.

Recently, several global tracking methods based on flow networks have been proposed to avoid local optima and to prevent linking of non-stationary false positive detections [20,2,18,12]. For these approaches, multi-object tracking is solved by conversion into a combinatorial optimization problem which can be solved in polynomial time [6,20]. Generally, for graph-based tracking, observations are represented by vertices, while costs are assigned to edges to denote various levels of support for associations between observations. Solutions of the multi-object tracking problem are then returned as paths with minimum or maximum costs. The popular mass-flow approach presented in [2], however, is not applicable in our scenario, as the graph topology is based on a discrete spatial subdivision which leads to impractically large graphs to achieve sub-pixel accuracy.

Current work on tracklet-based multi-object tracking has shown promising results [19,15,9,16]. Possible approaches are based on tracklet assignment in an iterative [19,10] or non-iterative way [15]. In [16], a sliding window that shifts with every frame is used and tracklets are found via inference from a set of Bayesian networks. Similarly, in [19] particle filters are used to locally generate tracklets from a temporal sliding window. However, the general drawback of extracting tracklets by considering only observations over a short period of time is that useful global information might be lost.

We tackle the dense multiple 3D object tracking problem by a two-stage *Directed Acyclic Graph* (DAG) formulation. Motivated by our medical application scenario, we do not use any appearance consistency or common assumptions such as the homography constraint [1]. In the first stage of our approach, tracklets are generated by finding the shortest path in the graph of all 3D detections. Afterwards, final tracks are found in a similar way by finding the shortest path in the graph of all tracklets. This results in a globally optimal linking of detections



(a) The landmark graph $\mathcal{G}$ includes vertices for all 3D point hypotheses and edges between vertices of succeeding frames.

(b) The tracklet graph $\mathcal{G}'$ contains vertices for all tracklets and edges between all vertices in a temporally consistent order.

Fig. 2: Exemplary graph topologies used in our approach. Additional edges from source and to sink vertex (dotted and dashed lines) allow initiation and termination of tracklets or tracks at any time.

and tracklets, while providing a flexible framework which can easily be adapted to various tracking scenarios based on the edge cost functions.

The structure of the paper is as follows: Section 2 will in detail present our two-stage graph-based method for multi-object tracking. Experimental results on real X-ray recordings of a beating sheep heart are presented in Sect. 3, including qualitative results and a quantitative comparison to ground-truth data provided by human experts. Section 4 concludes this work and discusses further plans.

2 Two-stage Graph-based Tracking

In this section we present our approaches for both extraction and linking of tracklets. We assume to have access to calibration data and detections for the individual views. The application of this approach to our medical scenario of 3D marker tracking in X-ray sequences is described in detail in Sect. 3.

2.1 Tracklet Generation

In order to overcome the problem of trajectory occlusions and interactions, we designed our tracking approach to directly operate on 3D data. We assume to have access to 3D point hypotheses $\boldsymbol{P}_0^t, \boldsymbol{P}_1^t, \ldots \in \mathbb{R}^3$ reconstructed from 2D marker detections $\boldsymbol{p}_0^{t,I_l}, \boldsymbol{p}_0^{t,I_r}, \boldsymbol{p}_1^{t,I_l}, \boldsymbol{p}_1^{t,I_r}, \ldots \in \mathbb{R}^2$ obtained from left and right images I_l, I_r. Using these observations, we construct a landmark graph $\mathcal{G} = (\boldsymbol{V}, \boldsymbol{E}, \boldsymbol{w})$, where each node $\boldsymbol{v}_i^t \in \boldsymbol{V}$ represents one 3D point $\boldsymbol{P}_i^t$ hypothesis. These nodes are connected across neighboring frames $t, t+1$ by directed edges $\boldsymbol{e}_{i,j} = \left(\boldsymbol{v}_i^t, \boldsymbol{v}_j^{t+1}\right) \in \boldsymbol{E}$, and we obtain a bipartite graph topology as outlined in Fig. 2a. Framewise misdetections are handled by creating additional edges $\boldsymbol{e}_{i,j}^{t,t+\Delta t} = \left(\boldsymbol{v}_i^t, \boldsymbol{v}_j^{t+\Delta t}\right)$ across further time steps, which allows skipping certain frames without appropriate detections. The assigned edge weights

$$w_{i,j} = w(\boldsymbol{v}_i^t, \boldsymbol{v}_j^{t+\Delta t}) = \begin{cases} d_{\text{spat}}\left(\boldsymbol{v}_i^t, \boldsymbol{v}_j^{t+\Delta t}\right) \cdot d_{\text{temp}}\left(\boldsymbol{v}_i^t, \boldsymbol{v}_j^{t+\Delta t}\right) & \Delta t > 0 \\ \infty & \text{else} \end{cases} \in \boldsymbol{w} \quad (1)$$

are proportional to the product of the Euclidean distance $d_{\text{spat}}(\boldsymbol{v}_i^t, \boldsymbol{v}_j^{t+\Delta t}) = \|\boldsymbol{P}_i^t - \boldsymbol{P}_j^{t+\Delta t}\|_2$ between the two represented 3D points and the number Δt of

skipped frames. Additionally, we connect each vertex $\boldsymbol{v}_i^t$ with the source $\boldsymbol{v}_{\text{source}}$ and the sink $\boldsymbol{v}_{\text{sink}}$ in order to obtain shorter tracklets when a marker was not detected for longer times. The associated edge weights

$$w_{\text{source}}(\boldsymbol{v}_i^t) = (t+1) \cdot d_{\text{penalty}} \qquad \text{and} \qquad w_{\text{sink}}(\boldsymbol{v}_i^t) = (t_{\max} - t) \cdot d_{\text{penalty}} \quad (2)$$

are proportional to the product of the number of skipped frames and the average linking distance to discourage unnecessary shortcuts.

Having such a graph $\mathcal{G}$ as exemplary outlined in Fig. 2a, a consistent tracklet $\tau = (\wp, t_{\tau,0}, t_{\tau,1})$ from frame $t_{\tau,0}$ to frame $t_{\tau,1}$ can be obtained by finding a path $\wp = (\boldsymbol{v}_{\text{source}}, \boldsymbol{v}^{t_{\tau,0}}, \ldots, \boldsymbol{v}^{t_{\tau,1}}, \boldsymbol{v}_{\text{sink}})$ from the source to the sink with minimal cumulated weight. For this purpose we iteratively employ Dijkstra's *shortest path* algorithm [7] until a specified number of iterations is reached or no more optimal paths can be extracted from the graph. Found paths are invalidated by setting the weights of all outgoing edges to infinity for each node of the path. In a post-processing step, missing observations between linked tracklets are interpolated linearly. Furthermore, duplicate tracklets, *i.e.* tracklets with almost identical spatial and temporal extents are merged.

2.2 Tracklet Linking

To fuse individual tracklets into complete tracks, we again formulate this problem in a graph-based way. Each vertex $\boldsymbol{v}'_i \in \boldsymbol{V}'$ in the new tracklet graph $\mathcal{G}' = \left(\boldsymbol{V}', \boldsymbol{E}', \boldsymbol{w}'\right)$ represents a tracklet $\tau_i = (\wp_i, t_{\tau_i,0}, t_{\tau_i,1})$ from frame $t_{\tau_i,0}$ to frame $t_{\tau_i,1}$. All vertices are connected by directed edges $\boldsymbol{e}'_{i,j} \in \boldsymbol{E}'$, which results in a linear graph structure. The associated weights

$$w'_{i,j} = w'(\boldsymbol{v}'_i, \boldsymbol{v}'_j) = \begin{cases} d'_{\text{spat}}\left(\boldsymbol{v}'_i, \boldsymbol{v}'_j\right) \cdot d'_{\text{temp}}\left(\boldsymbol{v}'_i, \boldsymbol{v}'_j\right) & t_{\tau_{j,0}} - t_{\tau_{i,1}} > 0 \\ \infty & \text{else} \end{cases} \in \boldsymbol{w}' \quad (3)$$

are proportional to the product of spatial distances $d'_{\text{spat}}\left(\boldsymbol{v}_i, \boldsymbol{v}_j\right) = \|\boldsymbol{P}_i^{t_{\tau_j,1}} - \boldsymbol{P}_j^{t_{\tau_i,0}}\|_2$ and temporal distance $d'_{\text{temp}}\left(\boldsymbol{v}'_i, \boldsymbol{v}'_j\right) = t_{\tau_{j,0}} - t_{\tau_{i,1}}$ between two represented tracklets. If two tracklets have a conflicting temporal order or overlap in time, the edge is weighted with infinite costs. Again, each vertex $\boldsymbol{v}'_i \in \boldsymbol{V}'$ is directly connected to $\boldsymbol{v}'_{\text{source}}$ and $\boldsymbol{v}'_{\text{sink}}$ with associated weights similar to those of Eq. 2. In this case, paths $\wp' = (\boldsymbol{v}'_{\text{source}}, \boldsymbol{v}'_1, \boldsymbol{v}'_2, \ldots, \boldsymbol{v}'_{\text{sink}})$ with minimal cumulated weights within this tracklet graph $\mathcal{G}'$ represent consistent sequences of tracklets $\mathcal{T} = (\wp', t_{\tau,0}, t_{\tau,1})$. The extraction of tracks is performed iteratively until no optimal path can be found or a certain number of paths are found. In a further post-processing step, gaps between linked tracklets are interpolated.

3 Experiments and Results

To assess the general performance and practical applicability of our proposed 3D tracking approach, we conducted experiments on real-world X-ray recordings of a beating sheep heart. Specifically, our goal was to analyze the following questions:

(i) Is our proposed method generally able to deal with the difficulties in this application (*e.g.* 2D marker occlusions, non-distinguishable marker appearance, inhomogeneous marker movement)? (ii) How does our method perform compared to standard 3D tracking approaches such as particle filter based tracking for the scenario at hand? (iii) Is the tracking accuracy of our approach comparable to human experts, and are the results sufficient for medical analyses? In the following, the experimental setup and the according results are presented.

3.1 Experimental Setup

The data used in our experiments was acquired by cardiac surgeons using the biplanar X-ray system shown in Fig. 1a. It shows the beating heart of a sheep, including 42 radio-opaque markers. The markers have a diameter of 2 mm and a spherical (30 times) or cylindrical (12 times) shape. The recorded images have a spatial resolution of 1024×768 pixels and a temporal resolution of 500 Hz. The total length of the sequence is 3,001 frames and covers 8 complete cardiac cycles.

For the calibration of the camera setup, we used a custom-built 140 mm $\times$ 60 mm $\times$ 0.5 mm radio-opaque steel plate containing 18 circular holes of diameter 5 mm. The holes can automatically be detected and identified in the resulting X-ray images. We performed the calibration based on Zhang [21], yielding an average backprojection error of 1.3 pixels. The angle between both cameras was 115°, while the distance between each camera and the heart was about 825 mm.

For each camera, 2D detections were obtained by firstly finding discontinuities in the images, *e.g.* using the Laplace operator. Afterwards, initial detections were extracted using simple blob detection. To reduce the number of false positives, saliency maps were built based on temporal variations throughout the whole sequence. By triangulating all detection pairs of the two camera views which were supported by the estimated epipolar geometry, we obtained 3D marker hypotheses which were then used as input for our proposed tracking approach.

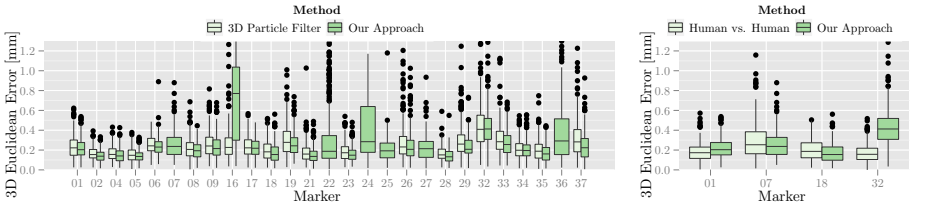
To evaluate the quality of the tracking results, we performed a comparison to ground-truth data provided by human experts, which is available for every 10^{th} frame and all uniquely identifiable markers (37 out of 42). We employ the *multiple object tracking precision* (MOTP) and *multiple object tracking accuracy* (MOTA) metrics [3], which have become the *de facto* standard in the field of multi-object tracking evaluation. The former allows to assess the precision of the tracker independently of correct object matches, while the latter provides information about object misses, mismatches, and false positive tracks.

3.2 Tracking Results

We extracted 3,000 tracklets using the first stage of our algorithm. In the second stage, we extracted all tracks whose length was at least 50% of the total sequence length, resulting in 39 tracks. In order to assess the quality of our results with respect to standard 3D tracking approaches, we performed a comparison to the particle filter based tracking approach presented in [11]. To ensure a fair comparison, we extended the method of [11] to 3D in a straightforward way (3D

Table 1: Multiple object tracking accuracy (MOTA) for the sheep heart sequence of our approach and the 3D extension of the particle filter approach presented in [11]. While the mismatch rate is equally low for both methods, our approach clearly outperforms [11] in terms of false positives, misses, and MOTA.

Method	Miss Rate	False Pos. Rate	Mismatch Rate	MOTA
Our Approach	**26.02%**	**6.23%**	1.26%	**66.49%**
3D Extension of [11]	43.78%	16.22%	**0.37%**	39.64%



(a) Multi object tracking precision (MOTP)

(b) Comparison to precision of human experts

Fig. 3: Quantitative results for the sheep heart sequence. (a) Multi object tracking precision (MOTP) in comparison to the 3D extension of the particle filter approach of [11]. While the precision is comparable for both methods, in six cases our approach reliably tracks markers which can not be tracked by [11]. (b) Comparison to precision of human experts for four markers having ground-truth data provided by multiple persons.

state vectors and updates instead of 2D). Furthermore, we used identical 3D detections and selected the same amount of best tracks as for our algorithm.

The MOTA results of both approaches are shown in Tab. 1. It can be seen that our approach has a moderate miss rate, while the false positive rate and the mismatch rate are relatively low. This behavior can be explained by the fact that the tracklet association step of our approach favors long and reliable tracks, while short and unreliable tracks are discarded at the expense of *full misses*, *i.e.* markers for which no track is present throughout the whole sequence. In the medical scenario at hand, this property is to be favored, as it is more reliable to have no tracks instead of wrong tracks for certain markers. While the mismatch rate of the approach of [11] is even lower than for the proposed method, it is clearly outperformed by our approach in terms of false positives and misses.

The results of MOTP evaluation are presented in Fig. 3a, separately for each marker. Only results are included for markers whose ground-truth data could be obtained and no full misses occurred. We can state that the average 3D precision of our approach is about 0.2 mm, with only minor differences between markers. A notable exception is marker 16, which is located at the end of a cardiac valve and suffers from the very abrupt movements. The particle filter approach of [11] gives comparable results on many markers. However, in six cases our approach is able to reliably track markers which can not be tracked using [11]. Qualitative results, namely 3D surface reconstructions for one cardiac cycle based on tracking results of our algorithm are shown in Fig. 4.

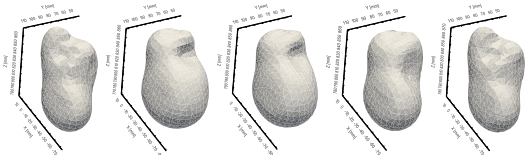


Fig. 4: 3D surface reconstruction of the sheep heart sequence based on our marker tracking results for one cardiac cycle (approximately 350 frames).

Given the qualitative and quantitative evaluations, it can be stated that our approach is able to deal with the difficulties in the data and provides promising results while outperforming standard 3D tracking approaches.

3.3 Comparison to Human Experts

In order to relate the tracking precision of our approach to human experts, for four representative markers ground-truth data was provided by more than one person. The results of the comparison are presented in Fig. 3b. In three out of four cases, our approach is able to compete with the precision of human experts. Only for marker 32, the human results are more precise, which might be caused by the fact that the marker is partly occluded by an anatomical structure in one camera view. All in all, however, we can state that the results are very promising and indeed comparable to human experts. Thus, our approach is clearly suited for practical applications. The fact that it can be used fully automatically supports above argumentation, and shows its applicability for medical marker tracking.

3.4 Complexity and Runtime

The entire system was implemented in `C++`. Our method has a complexity of $\mathcal{O}(k \cdot (n \log n + m))$, where n is the number of nodes, m is the number of edges, and k the number paths to be found in the respective graph. For the extraction of 3,000 tracklets from a 1.16×10^6 node 3D detection graph, our algorithm needs approximately 78 minutes, while the extraction of 39 final tracks from the tracklet graph takes about 31 seconds. All measurements were conducted on a standard desktop computer with an Intel® Core™ i5-760 CPU (2.80 GHz).

4 Conclusions

In this work, we presented a two-stage graph based approach for multiple marker tracking in X-ray recordings of beating hearts. The first stage of our approach consists of constructing a 3D observation graph, from which tracklets are extracted via shortest path optimization. Similarly, in the second stage, full tracks are found by constructing a tracklet graph. This process allows for a global linking of detections and tracklets and can easily be adaptapted based on the edge cost functions. We evaluate our approach on a sequence of a beating sheep heart and achieve a results which are comparable to human experts.

As next steps, we would like to incorporate efficient motion models. Also, adapting the edge cost functions based on additional knowledge such as occupancy maps or geometric information could improve the tracking performance.

Acknowledgements

Animal experiments adhered to relevant regulations and were approved by the federal state of Thuringia, Germany. We thank Christoph Bettag for his valuable comments and Rommy Petersohn for technical support during the experiments.

References

1. Andriyenko, A., Schindler, K.: Multi-target tracking by continuous energy minimization. In: CVPR. pp. 1265–1272 (2011)
2. Berclaz, J., Fleuret, F., Turetken, E., Fua, P.: Multiple object tracking using k-shortest paths optimization. TPAMI 33, 1806–1819 (2011)
3. Bernardin, K., Stiefelhagen, R.: Evaluating multiple object tracking performance: The clear mot metrics. EJIVP 246309 (2008)
4. Bredereck, M., Jiang, X., Körner, M., Denzler, J.: Data association for multi-object tracking-by-detection in multi-camera networks. In: ICDSC (2012)
5. Breitenstein, M., Reichlin, F., Leibe, B., Koller-Meier, E., Gool, L.: Online mutiperson tracking-by-detection from a single, uncalibrated camera. TPAMI 33(9), 1820–1833 (2011)
6. Collins, R.T.: Multitarget data association with higher-order motion models. In: CVPR. pp. 744–751 (2012)
7. Dijkstra, E.W.: A note on two problems in connexion with graphs. Numerische Mathematik 1, 269–271 (1959)
8. Fleuret, F., Berclaz, J., Lengagne, R., Fua, P.: Multi-camera people tracking with a probabilistic occupancy map. TPAMI 30, 267–282 (2008)
9. Ge, W., Collins, R.T.: Multi-target data association by tracklets with unsupervised parameter estimation. In: BMVC (2008)
10. Huang, C., Wu, B., Nevatia, R.: Robust object tracking by hierarchical association of detection responses. In: ECCV. pp. 788–801 (2008)
11. Jiang, X., Rodner, E., Denzler, J.: Multi-person tracking-by-detection based on calibrated multi-camera systems. In: ICCVG. pp. 743–751 (2012)
12. Leal-Taixé, L., Pons-Moll, G., Rosenhahn, B.: Branch-and-price global optimization for multi-view multi-target tracking. In: CVPR. pp. 1987–1994 (2012)
13. Malassiotis, S., Strintzis, M.G.: Tracking the left ventricle in echocardiographic images by learning heart dynamics. IEEE Trans. on Med. Imag. 18, 282–290 (1999)
14. Muijtjens, A., Roos, J., Arts, T., Hasman, A., Reneman, R.: Tracking markers with missing data by lower rank approximation. J. Biomech. 30, 95–98 (1997)
15. Nillius, P., Sullivan, J., Carlsson, S.: Multi-target tracking - linking identities using bayesian network inference. In: CVPR. pp. 2187–2194 (2006)
16. Prokaj, J., Duchaineau, M., Medioni, G.: Inferring tracklets for multi-object tracking. In: CVPR Workshops. pp. 37–44 (2011)
17. Satoh, Y., Okatani, T., Deguchi, K.: A color-based tracking by kalman particle filter. In: ICPR. pp. 502–505 (2004)
18. Wu, Z., Kunz, T.H., Betke, M.: Efficient track linking methods for track graphs using network-flow and set-cover techniques. In: CVPR. pp. 1185–1192 (2011)
19. Xing, J., Ai, H., Lao, S.: Multi-object tracking through occlusions by local tracklets filtering and global tracklets association with detection responses. In: CVPR. pp. 1200–1207 (2009)
20. Zhang, L., Li, Y., Nevatia, R.: Global data association for multi-object tracking using network flows. In: CVPR (2008)
21. Zhang, Z.: A flexible new technique for camera calibration. TPAMI 22(11), 1330–1334 (2000)