

Automatic Fusion of Segmentation and Tracking Labels^{*}

Cem Emre Akbaş¹, Vladimír Ulman^{1,2}, Martin Maška¹, Florian Jug², and Michal Kozubek¹

¹ Center for Biomedical Image Analysis, Faculty of Informatics, Masaryk University, Brno, Czech Republic {akbas, xulman, xmaska, kozubek}@fi.muni.cz
<https://cbia.fi.muni.cz/>

² Max Planck Institute of Molecular Cell Biology and Genetics, Dresden, Germany
jug@mpi-cbg.de

Abstract. Labeled training images of high quality are required for developing well-working analysis pipelines. This is, of course, also true for biological image data, where such labels are usually hard to get. We distinguish human labels (gold corpora) and labels generated by computer algorithms (silver corpora). A naturally arising problem is to merge multiple corpora into larger bodies of labeled training datasets. While fusion of labels in static images is already an established field, dealing with labels in time-lapse image data remains to be explored. Obtaining a gold corpus for segmentation is usually very time-consuming and hence expensive. For this reason, gold corpora for object tracking often use object detection markers instead of dense segmentations. If dense segmentations of tracked objects are desired later on, an automatic merge of the detection-based gold corpus with (silver) corpora of the individual time points for segmentation will be necessary. Here we present such an automatic merging system and demonstrate its utility on corpora from the Cell Tracking Challenge. We additionally release all label fusion algorithms as freely available and open plugins for Fiji³.

Keywords: Label fusion image annotation segmentation labels tracking labels

1 Introduction

Reliably obtaining high quality segmentation results is, in general, difficult. On biological microscopy data it is common to have segmentation results from multiple sources – either human annotations or automatic segmentations. Multiple solutions have been proposed to merge labels from multiple sources into one consensus labeling [1–3].

^{*} This work has been supported by the German Federal Ministry of Research and Education (BMBF) under the code 031L0102 (de.NBI), and by the Czech Science Foundation (GACR), grant P302/12/G157.

³ <https://github.com/xulman/CTC-FijiPlugins>

The widespread use of deep learning [4] across many bioimage analysis and computer vision tasks has impelled both communities to establish various publicly available repositories of annotated image data for training as well as objective benchmarking of the developed algorithms. Whereas fusion of labels in static images is by now common practice, dealing with labels in time-lapse image data is largely unexplored. This is particularly true in cell tracking applications for which typically no complete gold corpora exist for dense segmentation [5]. Additionally, the few existing gold corpora for cell tracking make use of simplified detection markers instead of complete dense segmentations [6–8]. In order to recover proper segmentation labels of all tracked cells, an automatic solution to merge a simplified gold tracking corpus with dense segmentation corpora for the individual time points is needed.

In this paper, we address the problem of obtaining dense segmentation and tracking results for multidimensional time-lapse light microscopy image data from partial segmentations and detection-based tracking annotations. For a given video, a simplified gold tracking corpus is obtained with a unique detection label for all occurrences of the same cell. Next, silver segmentation corpora are generated for each frame by a set of automatic segmentation methods. This results in the production of image sequences that contain ideally similar but still inconsistent segmentation masks. Finally, we merge those sequences to form a single silver segmentation corpus per frame and merge it with a detection based tracking corpus driven by gold tracking markers to generate a complete and dense tracking corpus. In summary, we present a fully automatic approach to establish a combined silver segmentation and tracking corpus from multiple automatic segmentation results and a detection based gold tracking corpus.

2 Proposed Method

The proposed method follows a majority or weighted majority voting scheme. A flowchart showcasing the proposed method can be found in Figure 1. The required inputs are (i) expert annotated tracking detection markers (gold tracking corpus). Such markers can either be single pixels or simple objects, like small circles, and (ii) dense segmentations generated by automated segmentation routines (silver segmentation corpora). During merging of these resources, each simplified gold tracking marker at each time-point will consider all dense segmentation masks that cover more than 50% of it as in [9]. The dense segmentation masks that fail to cover more than 50% of any simplified gold tracking marker are discarded. Note that for each simplified gold tracking marker, there can exist at most one such segmentation mask. Consequently, a cumulative gray-scale mask with counts of how many times an image element was observed in the considered masks is computed. This fused mask is thresholded and labeled according to the corresponding gold marker label. Results are put into an output image that accumulates these relabeled dense segmentations. Note that these relabeled segmentation masks can overlap or consist of unconnected components. We are simply removing overlapping areas. Furthermore, if the relabeled segmentation

mask size is reduced by more than 10%, it is removed entirely to prevent from spurious objects. Finally, if the relabeled segmentation masks consist of unconnected components (i.e. isolated islands), these components are also removed and only the largest component is kept. The flow of the proposed merging is illustrated in Figure 2 and its pseudocode can be found in Algorithm 1.

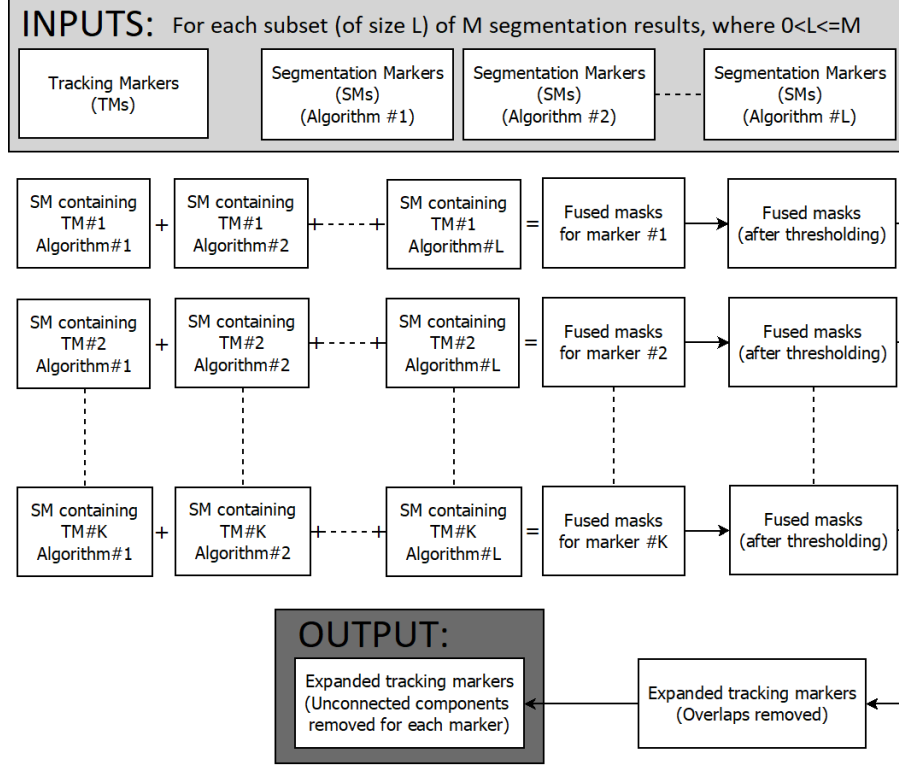


Fig. 1. The flowchart illustrating the proposed merging of a gold tracking corpus with dense silver segmentation corpora.

The silver dense segmentation corpus was created using a traditional majority voting scheme with a threshold value of $2/3$ of the number of input segmentation results. The fused silver segmentation and gold tracking corpora allowed us to calculate various spatio-temporal characteristics (e.g., the average cell overlap due to its movement between consecutive images) of the real videos. The pseudocode of this algorithm is provided in Algorithm 1.

In order to obtain the best possible dense silver corpus, this method is applied to all possible combinations of available segmentation results. For N segmentation results, $2^N - 1$ different non-empty input sets for merging are processed. Each merging result is compared to the dense gold segmentation corpus in terms of

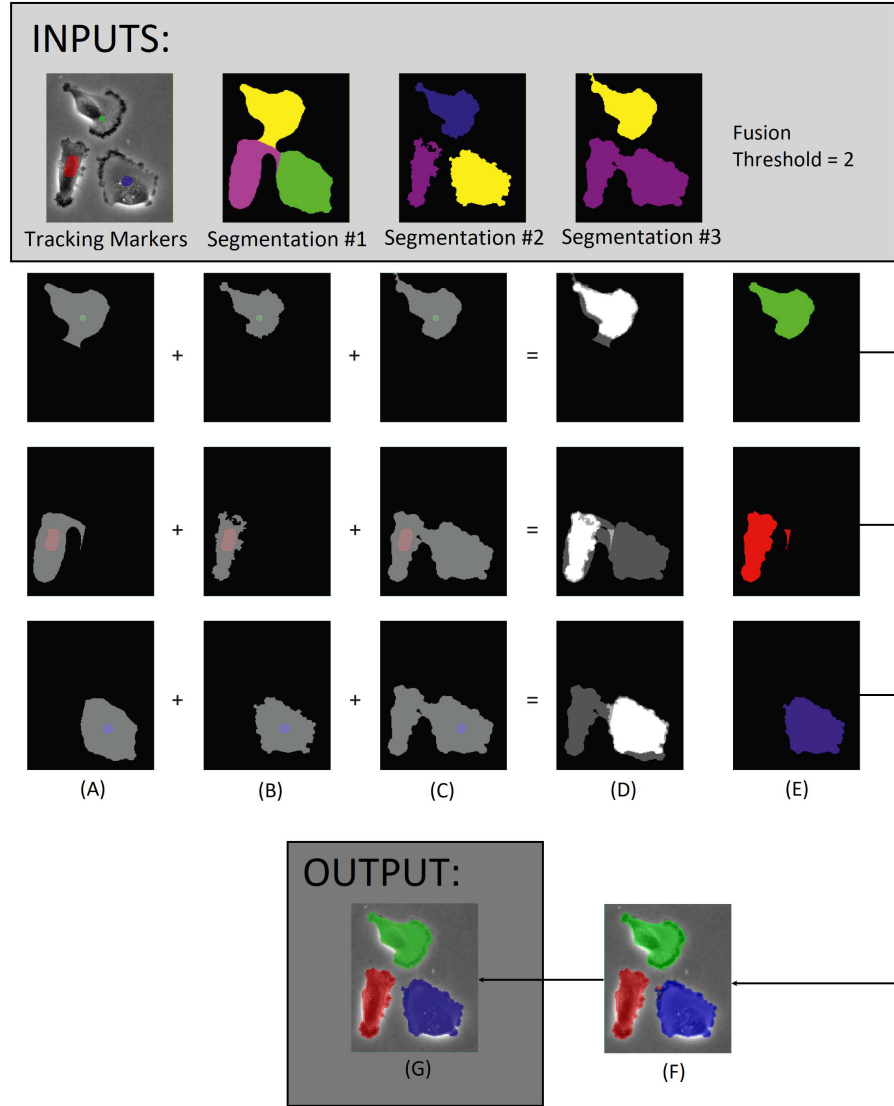


Fig. 2. Illustration of the proposed method with sample inputs from PhC-C2DH-U373 dataset. (A)-(C): Segmentation results from different sources, (D): Fused masks for each marker before thresholding, (E): Fused masks after thresholding, (F): Expanded tracking markers after removing overlaps and (G): Final expanded markers after removing unconnected components for each marker

SEG accuracy measure introduced in [7]. The input combination that produces the highest SEG score is taken as the input set of the dense silver segmentation corpus. The pseudo-code of this algorithm is given in Algorithm 2.

It is important to note that these algorithms are fully automated and require no subsequent manual refinement or checking. Missing objects with respect to simplified gold tracking ground truth are not manually inserted. Therefore, the number of objects in the dense silver segmentation corpus can be lower than the number of objects in the simplified gold tracking corpus.

Algorithm 1: *Merge_Segmentation_and_Tracking_Labels*

```

Inputs: Number of frames in the processed 2D or 3D video: N
          Number of available segmentation approaches: M
          Images containing simplified gold tracking labels
          (e.g., circles)  $T(i)$ ,  $i = 1 \dots N$ 
          Images with temporally inconsistent segmentation labels
           $S(i,j)$ ,  $i = 1 \dots N$ ,  $j = 1 \dots M$ 
          Minimal expected size of an object in pixels/voxels V
Output: Images with temporally consistent merged segmentation
          and tracking labels  $R(i)$ ,  $i = 1 \dots N$ 
for  $i = 1$  to  $N$  do
     $R(i)$  = zero-initialized image of size like  $T(i)$ 
     $K$  = number of labels in  $T(i)$ 
     $t$  = majority voting threshold
    for  $k = 1$  to  $K$  do
         $A$  = zero-initialized image of size like  $T(i)$ 
        for  $j = 1$  to  $M$  do
            if some object in  $S(i,j)$  overlaps >50% of  $k$ -th object in
             $T(i)$  then
                add the mask of this object (add 1 to all pixels within
                this mask) to  $A$ 
            end for
             $B$  = binary image obtained by thresholding  $A$  with a threshold  $t$ 
            delete all connected components in  $B$  that do not overlap with
             $k$ -th object in  $T(i)$ 
            if object in  $B$  overlaps with some object in  $R(i)$  then
                remove overlapping part in both  $B$  and  $R(i)$ 
                if the object reduction in  $B$  was >10% then  $B$  = zero image
                if the object reduction in  $R(i)$  was >10% then
                    remove this object in  $R(i)$ 
                add object from  $B$  (if any) into  $R(i)$  with label  $k$ 
            end for
        remove all unconnected object components in  $R(i)$  whose size is
        smaller than  $V$ 
    end for
return  $R$ 

```

Algorithm 2: *Find_Optimal_Combination_of_Segmentation_Approaches*

Input: The same as in Algorithm 1 plus gold segmentation labels for selected frames or selected objects only $G(i)$, $i = 1 \dots N$, where $G(i)$ can be empty or incomplete for some frames
Output: The same as in Algorithm 1 but for the optimal subset of segmentation approaches
 $SEG_{best} = 0$
for each non-empty subset Z of $1 \dots M$ **do**
 R = result of Algorithm 1 run with input segmentation approaches restricted to set Z (the input parameter M will be equal to $|Z|$)
 SEG = average Jaccard index for all objects in G compared to corresponding objects in R
 if $SEG > SEG_{best}$ **then** $SEG_{best} = SEG$, $R_{best} = R$
end for
return R_{best}

3 Experimental Results

In the experiments, simulated datasets are used due to the availability of complete and dense gold segmentation and tracking corpora. Experiments are carried out using two time-lapse videos of Fluo-N2DH-SIM+ training dataset from Cell Tracking Challenge [8], one of the few resources for which a large simplified gold tracking corpus is available. For both videos, segmentation results of 13 different algorithms from Cell Tracking Challenge are available. In the experiments, six segmentation results that perform above a certain threshold are used for merging. First video is a sequence of 65 images and the second video is a sequence of 150 images. Therefore, segmentation markers are merged using $2^6 - 1$ (empty set is excluded) different combinations of inputs to get the best possible dense silver segmentation corpus and compared to the gold segmentation corpus in this case of simulated datasets with complete and dense ground truth. The combination that gives the highest SEG score is selected as the input set of dense silver segmentation corpus. For the first video, merging outputs that produce the highest SEG score is obtained using four segmentation results; HD-Hau-GE, KTH-SE (1), FR-Ro-GE and LEID-NL. While on the second video, three segmentation results, KTH-SE (1), PAST-FR and FR-Ro-GE, produce the optimal merging outputs in terms of SEG score over reference objects. Experimental results obtained on the first video and the second video are presented in Table 1. Computation time was 12 hours for the first video and 21 hours for the second video for 63 different combinations of six available segmentation results. Experiments are carried out on a Linux SMP Debian 4.9.65 machine that runs on Intel(R) Core(TM) i7 CPU 920 with 12 GB RAM.

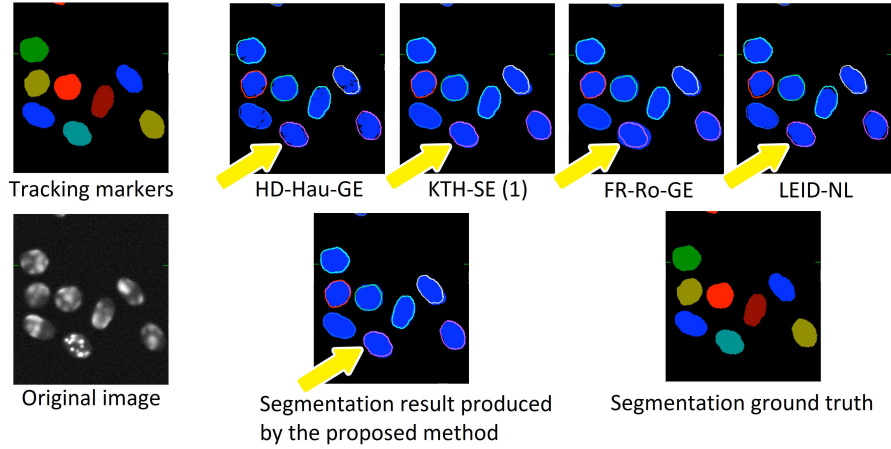


Fig. 3. A sample set of gold tracking markers, segmentation sources that achieve the best possible merging result, (visually enhanced) original image, segmentation result produced by the proposed method and the segmentation ground truth of the 41-st frame of the first video. The segmentation results are shown in blue and are overlaid with contours of the ground truth to facilitate the comparison. Yellow arrows point on the same nuclei that is under-/over-segmented in the results.

Table 1. SEG scores for segmented objects in the first video (second row) and in the second video (third row). First column denotes the merging outputs that are obtained using results of HD-Hau-GE, KTH-SE (1), FR-Ro-GE and LEID-NL for the first video and using results of KTH-SE (1), PAST-FR and FR-Ro-GE for the second video. Rest of the columns present individual algorithm results.

Merged	HD-Hau-GE	KTH-SE (1)	UZH-CH	PAST-FR	FR-Ro-GE	LEID-NL
0.873	0.798	0.865	0.807	0.844	0.848	0.864
0.657	0.579	0.657	0.574	0.629	0.640	0.605

It is shown in Table 1 that on the first video, our merging tool outperforms segmentation results of individual algorithms in terms of SEG score. This improvement can be observed in Figure 3. On the second video (Table 1), merged segmentation result produces the same SEG score as KTH-SE (1) does. The SEG score is known to permit various sources of segmentation errors that, however, lead to the same coefficient value in the case of the second video (Table 1). The number of not expanded markers are more in KTH-SE (1) segmentation (203 markers due to not found, 96 markers due to unresolved collision; 299 markers in total) than in the merged segmentation (128 markers due to not found, 163 markers due to unresolved collision; 291 markers in total). Moreover, the merged segmentation contains more expanded markers (3072 markers) than KTH-SE (1) segmentation does (3064 markers). Therefore, the merged seg-

mentation is not identical to the original KTH-SE (1) segmentation despite the SEG score values are the same. Since the other inputs scored lower SEG score values, they must be deviating from the segmentation ground truth in more regions than KTH-SE (1) does. We also observed a similar performance on real datasets. The proposed method is voting-based, suggesting that most of individual over-segmentations will be stripped away unless majority supports them. Similarly, most of individual under-segmentations will be recovered. This leads to a merged segmentation that is more compact in shape (compare, e.g., top row with column (E) in Fig. 2), and increases the SEG score. On the other hand, sometimes majority of input results misses a cell or nuclei largely or completely, leading to a decrease of the overall SEG score. Similarly, removing overlapping parts of colliding markers decreases the SEG score while removing unconnected components (i.e. isolated islands) increases it. Therefore, our tool provides an increase in the segmentation accuracy for the images, for which removed areas of unconnected components are larger than overlapping parts of colliding markers.

4 Discussion and Future Directions

We have presented a method for creating large, dense tracking labels by merging existing corpora of various partial dense segmentations and a detection based tracking corpus. This method has the potential to save impossible amounts of manual human data annotation time when creating dense training data for microscopy datasets. We demonstrated the proposed method on datasets from the Cell Tracking Challenge [8], showing that it generates high quality labels even on large bodies of data. The fused silver segmentation and simplified gold tracking corpora allowed us to calculate more precise and more complete spatio-temporal characteristics. Such characteristics cannot often be computed from pure tracking results due to simplified markers. Additionally, the merged labels can now be used to train various (end-to-end) processing routines.

In our experiments, simulated datasets are used due to the availability of full segmentation results for all frames. Additionally, this allowed us to evaluate the performance of the proposed method more accurately. Each possible combination of available segmentation sources is used as the input set in order to obtain the best possible merging result. Therefore, the proposed method is capable of producing more accurate segmentation results than individual segmentation sources. We also showed that the proposed method improves the quality of the final segmentation during merging in terms of SEG accuracy measure. While the proposed method may not always provide more accurate segmentation results than any individual segmentation source does, it ideally provides the most complete tracking result compared to any single silver segmentation corpus.

Future extensions can make use of more involved merging schemes such as STAPLE [1], SIMPLE [2], or image-based alternatives [10, 11]. This could further improve the quality of merged segmentation labels. A comprehensive study using a large collection of CTC participant results and all CTC datasets will be performed.

References

1. Warfield, S.K., Zou, K.H., Wells, W.M.: Simultaneous truth and performance level estimation (STAPLE): An algorithm for the validation of image segmentation. *IEEE Transactions on Medical Imaging* **23**(7) (2004) 903–921
2. Langerak, T.R., van der Heide, U.A., Kotte, A.N.T.J., Viergever, M.A., van Vulpen, M., Pluim, J.P.W.: Label fusion in atlas-based segmentation using a selective and iterative method for performance level estimation (SIMPLE). *IEEE Transactions on Medical Imaging* **29**(12) (2010) 2000–2008
3. Lampert, T.A., Stumpf, A., Gañarski, P.: An empirical study into annotator agreement, ground truth estimation, and algorithm evaluation. *IEEE Transactions on Image Processing* **25**(6) (2016) 2557–2572
4. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. *Nature* **521**(6) (2015) 436–444
5. Coutu, D.L., Schroeder, T.: Probing cellular processes by long-term live imaging—historic problems and current solutions. *Journal of Cell Science* **126**(17) (2013) 3805–3815
6. F. Amat, W.L., Mossing, D.P., McDole, K., Wan, Y., Branson, K., Myers, E.W., Keller, P.J.: Fast, accurate reconstruction of cell lineages from large-scale fluorescence microscopy data. *Nature Methods* **11**(9) (2014) 951–958
7. Maška, M., Ulman, V., Svoboda, D., et al.: A benchmark for comparison of cell tracking algorithms. *Bioinformatics* **30**(11) (2014) 1609–1617
8. Ulman, V., Maška, M., Magnusson, K.E.G., et al.: An objective comparison of cell-tracking algorithms. *Nature Methods* **14**(12) (2017) 1141–1152
9. Matula, P., Maška, M., Sorokin, D.V., Matula, P., Ortiz-de Solórzano, C., Kozubek, M.: Cell tracking accuracy measurement based on comparison of acyclic oriented graphs. *PLOS ONE* **10**(12) (2015) 1–19
10. Schlesinger, D., Jug, F., Myers, G., Rother, C., Kainmueller, D.: Crowd sourcing image segmentation with iaSTAPLE. In: *IEEE International Symposium on Biomedical Imaging*. (2017) 401–405
11. Liu, X., Montillo, A., Tan, E.T., Schenck, J.F.: iSTAPLE: Improved label fusion for segmentation by combining STAPLE with image intensity. In: *SPIE Medical Imaging*. (2013) 866–920