# KEEPING COUNT: LEVERAGING TEMPORAL CONTEXT TO COUNT HEAVILY OVERLAPPING OBJECTS

*Luca Fiaschi* [1] *Gregor Konstantin* [1] *Bruno Afonso* [2] *Marta Zlatic* [2] *Fred A. Hamprecht* [1]

[1] HCI/IWR Heidelberg
[2] HHMI Janelia Farm

## ABSTRACT

When tracking and segmenting multiple objects under heavy occlusion, a large class of algorithms can greatly benefit from a preprocessing that reliably assesses the number of individuals in each cluster. This is a difficult task when relying on local information only, due to scarcity of training examples and lack of strongly predictive features. In this paper, we develop a deterministic graphical model to address the problem of counting the number of objects in each foreground region as global inference across the entire video sequence. We show that global inference improves over local predictions, and is able to produce an accurate and coherent output within an useful runtime.

***Index Terms*— deterministic higher order potential, constraint satisfaction, spatio-temporal segmentation**

## 1. INTRODUCTION

Larvae, such as *Drosophila*, are popular model organisms for behavioral studies. To allow studying the social dynamics of a large population in crowded situations, the ultimate aim is to track single individuals in spite of their non-distinguishable appearance, and mutual overlap. In this tracking scenario, where multiple objects cross each other, all algorithms that we are aware of require the knowledge of the number of individuals in each foreground region: either as a model parameter, as in *tracking by detection* methods [3, 9], or as an initialization, as in *tracking by model evolution* approaches [2, 7, 4]. The goal of this paper is to provide a reliable estimate of the number of individuals in each connected component.

Some approaches which rely only on local information have been proposed [13, 6]. However, in our experiments, appearance-based features that are local in space-time are not sufficient to achieve a high accuracy and lead to several inconsistencies across time. In difficult situations, as shown in Fig. 1, humans are still able to disambiguate and achieve a correct count by browsing the sequence forward and backward in time and looking for the separation of the individuals. Given the assumption that the number of visible individuals is conserved, Henriques et al. [8] obtained the count of pedestrians in merged detections as a minimum cost flow over the detection graph. However, their method requires manual initialization of isolated individuals, and relies on the size of the foreground detection only, which is a weak local cue in case of overlapping objects.

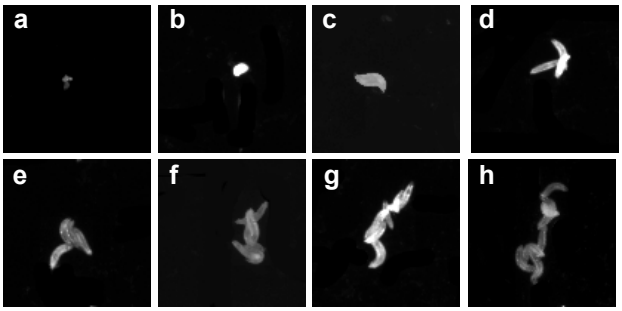


**Fig. 1**. How many larvae are there in each cluster from (a)-(h)? For a human looking at the entire sequence, it is possible to confirm that the true count is: $0, 1, \ldots 7$ from the top left to the bottom right. (a) False positive foreground detection. (b) Single rearing larva. (c) Two overlapping larvae. (d) Three overlapping larvae, etc.

Our approach is to mimic the strategy of the human expert by using a graphical model with deterministic constraints. Our main contribution is to integrate multiple local cues to achieve an object count that is consistent across space and time. The coupling of all estimates across space and time allows to propagate information from simpler parts of a video to complex clusters of larvae.

The relation between constraint satisfaction problems and graphical models has been throughly investigated in the context of Bayesian networks [10]. Akin [3, 9, 11], we formulate the MAP inference step as an integer linear program (ILP) that is solved with a standard software package [1]. We show that this formulation can handle a large volume of high resolution data. Our experimental evaluation demonstrates accurate and coherent results which can be exploited by downstream tracking and segmentation algorithms.

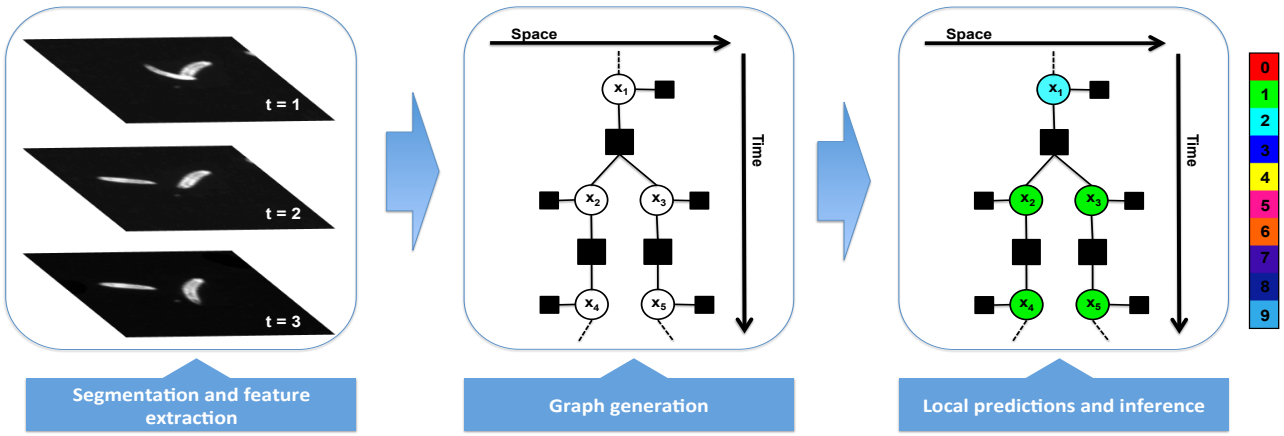


**Fig. 2**. An overview of the workflow. Each image is first thresholded into foreground vs. background. Then, a random variable representing the object count is associated with each spatially connected component. The probability distribution over the true object count is estimated, based on purely local features extracted from each connected component, by unary factors (represented by small black boxes). The random variables are also connected in time by deterministic higher order factors (large black boxes) that encapsulate the "conservation of objects" constraints. The connectivity of these higher order factors follows from spatio-temporal overlap. Finally, global inference determines the consensus estimate of the true object count.

## 2. METHODS

### 2.1. Model definition

Our workflow is depicted in Fig. 2. Firstly, we produce a set of candidate segmented foreground regions ($N$ in total) as detailed in Sect. 3.1. Secondly, for each region $i$, we assign a random variable $x_i \in \{0, ..., M\}$ expressing the number of contained larvae. We include the label $k = 0$ since we want to handle cases in which debris and other contaminations are segmented as foreground. $M$ represents the maximum number of objects in a foreground region. Thirdly, as detailed in Sect. 3.1, we link neighboring foreground regions from consecutive time steps in order to build an undirected detection graph $G = (\mathbf{X}, \mathbf{E})$, where $\mathbf{X} = \{x_i\}_{i=1}^{N}$ is the collection of all random variables and $\mathbf{E} = \{e_i\}_{i=1}^{L}$ is the set of edges. Our assumption that larvae cannot enter or leave the field of view[1] naturally imposes a set of constraints that relates all variables in $\mathbf{X}$. To make the structure of the problem explicit, we define the factor graph $G' = (\{\mathbf{X}, \phi, \Phi\}, \mathbf{E}')$ where two types of potentials are present (as depicted by small and big black squares in the central part of Fig. 2). First order potentials $\phi(x_i, f_i)$ express our belief that the current region contains a certain number of larvae based only on a local feature vector $\mathbf{F} = \{f_i\}_{i=1}^{N}$; higher order potentials are deterministic functions of the variables in their scope that enforce consistency. We define $G_{t,t+1} \subseteq G$ as the subgraph comprising the variables at time $t$ and $t+1$, and $S_{t,t+1}^{j}$ as the $j^{th}$ connected component of this subgraph:

$$G_{t,t+1} = \bigcup_{j} S_{t,t+1}^{j} \quad ; \quad S_{t,t+1}^{j} \cap S_{t,t+1}^{k} = \emptyset \quad \forall j \neq k$$

The higher order factors are defined as follows:

$$\Phi(\mathbf{X}_t^j, \mathbf{X}_{t+1}^j) = \begin{cases} 0 & \text{for } \sum_{x_i \in \mathbf{X}_t^j} x_i - \sum_{x_i \in \mathbf{X}_{t+1}^j} x_i = 0 \\ \infty & \text{for } \sum_{x_i \in \mathbf{X}_t^j} x_i - \sum_{x_i \in \mathbf{X}_{t+1}^j} x_i \neq 0 \end{cases} \quad (1)$$

where $\mathbf{X}_t^j \in S_{t,t+1}^j$ is the set of variables at time $t$ in the $j^{th}$ connected component. The total energy can be expressed in terms of the factor graph as:

$$U(\mathbf{X}) = \sum_{i=1}^{N} \phi(x_i, f_i) + \sum_{t} \sum_{S_{t,t+1}^j \in G_{t,t+1}} \Phi(\mathbf{X}_t^j, \mathbf{X}_{t+1}^j) \quad (2)$$

The Gibbs relation $P(\mathbf{X}, \mathbf{F}) = \frac{1}{Z} e^{-U}$, where $Z$ is the partition function, allows to establish the equivalence between the MAP configuration of the associated probabilistic graphical model and the argmin of Eq. (2).

### 2.2. Integer linear program formulation

We implement the MAP inference as an integer linear program with indicator variables. To each foreground region $i = 1, \ldots, N$ are associated the binary indicator variables $x_i^k \in \{0, 1\}$, $\sum_{k=0}^{M} x_i^k = 1$ encoding the number of objects in the region. We set the unary potential

[1] Except for the borders of the image. Boundary conditions are explained in Sect. 3.1.

$\phi_i^k = -\log p(x_i^k = 1|f_i)$, where $p(x_i^k = 1|f_i)$ is the probability for component $i$ to contain $k$ object instances given the features, as estimated by a local classifier. Then we have:

$$\begin{aligned}
&\min_{x_k^i} \sum_{k,i} \phi_i^k x_i^k \quad \text{s.t.} \\
&x_i^k \in \{0,1\} \quad \forall i,k \\
&\sum_{k=0}^{M} x_i^k = 1 \quad \forall i \\
&\sum_k \sum_{x_i^k \in \mathbf{X}_t^j} k\, x_i^k - \sum_k \sum_{x_i^k \in \mathbf{X}_{t+1}^j} k\, x_i^k = 0 \;\; \forall S_{t,t+1}^j \in G_{t,t+1} \; \forall t
\end{aligned} \tag{3}$$

The third constraint represents the conservation of the number of larvae as enforced by the higher order potentials of Eq. (1).

### 2.3. Unary potentials

For reasons both of accuracy and tractability, it is important to use informative unary potentials. The probability $p(x_i^k = 1|f_i)$ is learned with a classifier from labelled training data. However, this is a strongly unbalanced classification problem since, at least in our data, there are much fewer training examples available for classes $k > 1$. Therefore, we choose the following strategy: firstly, we train a classifier with examples of label $0, 1, 2, 3, many$, where class labelled "$many$" includes all examples of foreground regions containing 4 or more individual instances. Secondly, we take the probability $p(x_i^{many} = 1|f_i) := \beta_i$ and distribute it among the classes $k \geq 4$ according to the parametric function:

$$p(x_i^k = 1|f_i) = \beta_i \exp\Big(-\frac{(\tilde{\tau} - k)^2}{2\sigma^2}\Big)/\alpha \;, \;\; k \geq 4 \tag{4}$$

Here $\tilde{\tau}$ is the size of the foreground region (in pixels) normalized by the average size of all observed foreground regions and $\alpha$ is a normalization constant. In the following experiments we fix $\sigma = 2$. This procedure allows us to obtain a very sharp unary term for foreground regions containing few larvae, while we rely mostly on the global inference step to find the number of instances in big clusters.

## 3. EXPERIMENTAL RESULTS

### 3.1. Data and preprocessing

A population of 72 hours old *Drosophila* larvae was filmed for 5 minutes with a temporal resolution of 3.3 frames per second, 1000 frames in total. Images have a spatial resolution of 135.3 $\mu m$/pixel, a size of $1560 \times 1600$ pixels, and contain on average 323 larvae. For detection and segmentation of the foreground regions we use the open-source software ILASTIK [12]. After elimination of tiny isolated objects ($\tau < 15$ pixels), we compute connected components of each thresholded foreground probability image of the series. The graph $G$ is created by linking foreground regions from neighboring timesteps that overlap spatially, by more than 10 pixels. All foreground regions which are not fully inside a margin of 100 pixels from the image borders are excluded to avoid dealing with truncated larvae (cluster).

### 3.2. Training unary potentials

As explained in section 2.3, we obtain the unary term from labelled training data. All images in the sequence were labelled by hand to establish a gold standard. However, for the training of the classifier (a standard random forest [5]) we use only 10 images from the first 250 timesteps of the sequence (every 25 timesteps, 1% of the available data). A set of 21 object features describe the size and the shape of each foreground region: area, convex area, eccentricity, equivalent diameter, axis lengths, perimeter, solidity, mean intensity, variance of the intensity, total intensity and the magnitude of the first 10 Fourier contour descriptors.

### 3.3. Implementation details

In our experiments, we run and evaluate the results on the entire sequence (1000 timesteps). Inference takes 28 minutes. We use the following optimizations: firstly, not every larva interacts with all others over the course of time, therefore we can solve the optimization problem separately for independent connected components of the factor graph. We found our dataset to contain 158 subgraphs, with one huge subgraph consisting of 266215 foreground regions accounting for 93% of all regions (solving the problem for this subgraph takes approximately 14 min). Secondly, we use CPLEX's warm-start interface to initialize the solver with the assignments obtained by minimizing the unaries. Thirdly, we add conservative constraints that rule out improbable assignments. In particular, if $\tau$ is the size in pixels of the foreground region, for $\tau < 50$ we add the constraint $x_i \leq 1$. In a similar vein, $\tau < 120 \Rightarrow x_i \leq 2, \tau > 50 \Rightarrow x_i \geq 1, \tau > 300 \Rightarrow x_i \geq 2$.

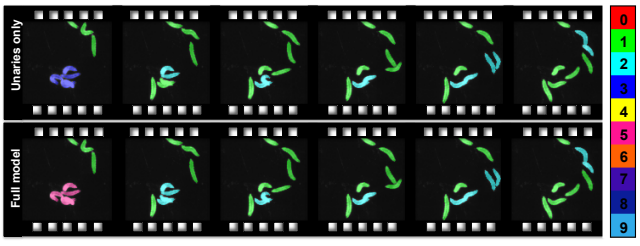


**Fig. 3**. Comparison between the predicted counts obtained by simple minimization of the unary potentials (first row) and the proposed MAP solution (second row). The second row shows the consistent, and correct, labelling.

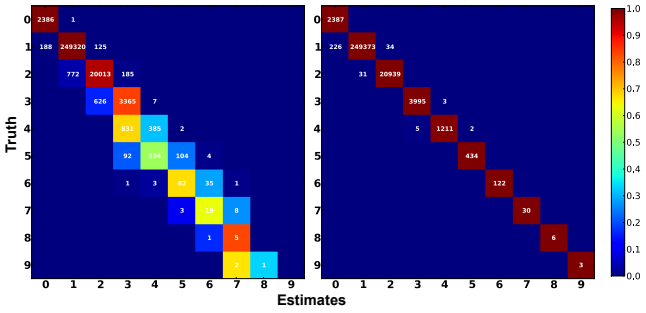


**Fig. 4**. Comparison between the confusion matrix from simple minimization of the unary potentials (left) and the proposed MAP solution (right). The false colors reflect the confusion after row-wise normalization.

### 3.4. Results

Figs. 3 and 4 show consistent improvements of the full model over local predictions. In particular, Fig. 3 illustrates how the higher order potentials help disambiguating an inconsistent assignment. To summarize our findings with a single number, we define an adjusted precision score for regions containing three or more individuals:

$$rec = \frac{\bar{\text{TP}}}{\text{TP} + \text{FP}} \tag{5}$$

where TP and FP are the true positives and false positives. $\bar{\text{TP}}$ is similar to TP but only considers foreground regions that do not directly violate consistency constraints as in Eq. (1). For the full model $rec$ is equivalent to standard precision, while for the reduced model $rec$ penalizes temporally inconsistent assignments. Under this measure, the score of the prediction obtained by simply minimizing the unaries is 61.1%, while the score of the full model is 99.8%.

## 4. DISCUSSION AND FUTURE WORK

We have demonstrated that the introduction of higher order deterministic consistency constrains outperforms local predictions when estimating the number of individuals in clusters of larvae. Our work builds on two assumptions which allow the introduction of the deterministic constraints: first, that the detection is sufficiently sensitive such that an isolated object cannot disappear for short periods of time. Second, that the temporal resolution of the data is sufficient for the construction of the consistency graph. As it stands, the proposed algorithm allows estimating the rate of larvae encounters, and the tracking of isolated individuals. These measurements could already support behavioural biologists. The principal benefit, however, is to provide input for future algorithms that should allow the tracking of each and every larva, even through complex agglomerates. We are currently developing a downstream tracking and segmentation algorithm exploiting the information from the presented method.

## 5. REFERENCES

[1] Ilog cplex, 2003.

[2] R. Bise, K. Li, S. Eom, and T. Kanade. Reliably tracking partially overlapping neural stem cells in dic microscopy image sequences. In *MICCAI Workshop on OPTIMHisE*, volume 5, 2009.

[3] R. Bise, Z. Yin, and T. Kanade. Reliable cell tracking by global data association. In *ISBI*, pages 1004–1010. IEEE, 2011.

[4] K. Branson and S. Belongie. Tracking multiple mouse contours (without too many samples). In *Computer Vision and Pattern Recognition, CVPR*, volume 1, pages 1039–1046. IEEE, 2005.

[5] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[6] A.B. Chan, Z.S.J. Liang, and N. Vasconcelos. Privacy preserving crowd monitoring: Counting people without people models or tracking. *CVPR*, 2008.

[7] E. Fontaine, A. Barr, and J.W. Burdick. Model-based tracking of multiple worms and fish. In *ICCV Workshop on Dynamical Vision*, 2007.

[8] J.F. Henriques, R. Caseiro, and J. Batista. Globally optimal solution to multi-object tracking with merged measurements. In *International Conference Computer Vision, ICCV*, pages 2470–2477. IEEE, 2011.

[9] B.X. Kausler, M. Schiegg, B. Andres, M. Lindner, U. Koethe, H. Leitte, J. Wittbrodt, L. Hufnagel, and F.A. Hamprecht. A discrete chain graph model for 3d+ t cell tracking with high misdetection robustness. *European Conference Computer Vision, ECCV*, 2012.

[10] R. Mateescu and R. Dechter. Mixed deterministic and probabilistic networks. *Annals of Mathematics and Artificial Intelligence*, 54(1):3–51, 2008.

[11] D. Roth and W. Yih. Global inference for entity and relation identification via a linear programming formulation. In *Introduction to Statistical Relational Learning*, pages 553–580. The MIT Press, 2007.

[12] C. Sommer, C. Straehle, U. Kothe, and F.A. Hamprecht. Ilastik interactive learning and segmentation toolkit. In *ISBI*, pages 230–233. IEEE, 2011.

[13] C. Wählby, T. Riklin-Raviv, V. Ljosa, A. L. Conery, P. Golland, F. M. Ausubel, and A. E. Carpenter. Resolving clustered worms via probabilistic shape models. In *ISBI*, pages 552–555, 2010.