# A Cellular Similarity Metric Induced by Siamese Convolutional Neural Networks

Morgan Paull
Stanford Bioengineering
mpaull@stanford.edu

## Abstract

*High-throughput microscopy imaging holds great promise for elucidating complex behavior of cells over time. However with increased throughput comes increased difficulty in managing the identity of large numbers of cells over timecourses. This project presents a neural network method based on visual similarity for determining cell identity across frames, which can work in concert with or instead of existing distance-based methods [7]. Using a siamese neural net architecture, we produce a visual similarity distance metric between pairs of input cells. This visual similarity score can be used to assign the identities of cells between frames of a microscopy time course.*

*Training of the siamese neural net is guided by a discriminative loss function developed by Chopra et al [5], which maximizes the energy score between cell pairs labeled as different cells, and minimizes the energy score between cell pairs labeled as the same cell at different time points.*

*Training and test data are selected from labeled timecourses accounting for in total 16.2 million labeled cell-pairs. The test accuracy of the model is 97.5% on cells one frame apart, 94.5% on cells separated two frames apart, and 95.0% on mixtures of one- and two-frame separated cells.*

## 1. Introduction

### 1.1. Visual similarity-based microscopy cell tracking

The contribution of this project is an improvement of the key step in the cell-tracking problem. The model takes in segmented time-course microscope images, and assigns pairwise identity of cells between frames to create per-cell time courses. The method operates on a cell-by-cell basis, so requires semantic segmentation to be completed upstream, and can be combined with an existing method, the linear assignment problem [7], to further increase single-cell time course accuracy after the initial frame-by-frame joining.

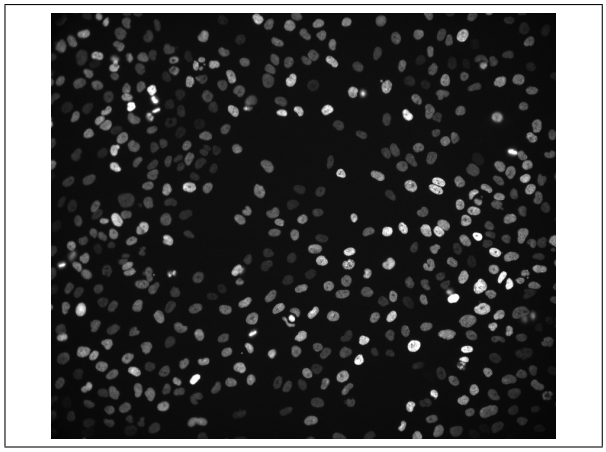

Figure 1. An example microscopy image frame. A typical time-course experiment in the dataset for this project runs 125 frames, each separated by 15 minutes in time.

### 1.2. Microscopy Semantic Segmentation

Microscopy experiments generate very large numbers of cell images. In a typical experiment from the type used for training and testing in this project, 96 positions are imaged once every 15 minutes for 125 total images. Each frame captures on the order of 200 cells, so a single experiment generates approximately 2.5 million cell images. Segmentation of the cells preceeds identification and tracking in this project. A variety of computer vision techniques are available for segmenting biological cells [21]: intensity thresholding is a common first-step [14], and for lower-magnification cells gaussian filters and laplacian-of-gaussian filters can be effective [8]. For higher-resolution images, edge detectors based on first-order and second-order derivatives are a common method [15]. VanValen et al present a method for instance segmentation using convolutional neural networks [20] - this technique is in use in the lab generating training and test data for this project.

### 1.3. Linear Assignment Problem

Existing methods for assigning cell paths through time rely on simple metrics such as x,y position and light intensi-

ties. These methods are used in cominbation with information about the structure of the problem (the same cell can't appear in the same frame twice, cells will rarely disappear and then reapear in a later frame, etc) to maximize expectation of creating a correct reconstruction. Collectively, this current state-of-the-art framework is called the linear assignment problem [7]- it is a greedy local approximations of the computationally infeasible multiple-hypothesis tracking method, which considers every possible cell identity combination to maximize probability of correct assignments. The multiple-hyopthesis tracking method is computationally infeasible for any but the smallest datasets. The linear assignment problem starts by making pairwise assignment of cell identities between adjacent frames, therefore the global accuracy of the method can be improved by increasing the pairwise cell identification accuracy.

This assignment is completed by minimizing the cost associated with joining two cells:

$$\hat{A}_{argmin} = \sum_{i=1}^{Rows} \sum_{j=1}^{Columns} A_{ij} C_{ij} \tag{1}$$

Where $A$ is a matrix of 0s and 1s, where a 1 at row $i$ and column $j$ indicates that the cell with index $i$ in frame $t$ is joined to the cell with index $j$ in frame $t+1$, and a zero indicates no link. Because a cell can only be linked to one cell in a future frame, the sum of the rows of $A$ is one, and the sum of the columns of $A$ is also one.

The cost matrix $C_{ij}$ is given by

$$\text{Frame t cell index} \begin{bmatrix} l_{11} & l_{12} & \times & \dots & \times \\ l_{21} & l_{22} & l_{23} & \dots & \times \\ \vdots & & & & \\ \times & \times & \dots & l_{nn-1} & l_{nn} \end{bmatrix} \quad \text{Frame t+1 cell index} \tag{2}$$

Where $l_{ij}$ is the cost associated with joining two cells, and $\times$ indicates that a join is impossible, because the cells are beyond some user-defined distance considered plausible for a cell to move. In traditional linear assignment, the cost of joining two cells is simply their euclidean distance between frames:

$$l_{ij} = \sqrt{x_i x_j + y_i y_j} \tag{3}$$

### 1.4. Siamese Neural Network

The applicate area of this project is cell recognition; however the technical approach most closely relates to work in facial recognition. The similarity metric used in assigning cell identities is created by a siamese neural network [2], which takes in pairs of inputs and runs a pair of convolutional neural networks with the same weights over both. The features from the last layer of the neural network are then used to create an energy score between the inputs. Highly similar images have a low energy score, and highly dissimilar images have a high energy score. This score serves as a distance metric [13] [19]. Siamese architecture neural nets have been applied to signature verification [2], facial recognition [5] [17], speaker identification in audio files [3] [4], generalization of object recognition to new categories [10] [18], and even dimensionality reduction [6]. This allows a network to produce a similarity score between any pair of input items. The method is intrinsically pairwise, and so for our model must operate on pairs of cell images, requiring that the images of cells in this project be pre-segmented. The cellular similarity score can then be used to assign identities between cells in different frames.

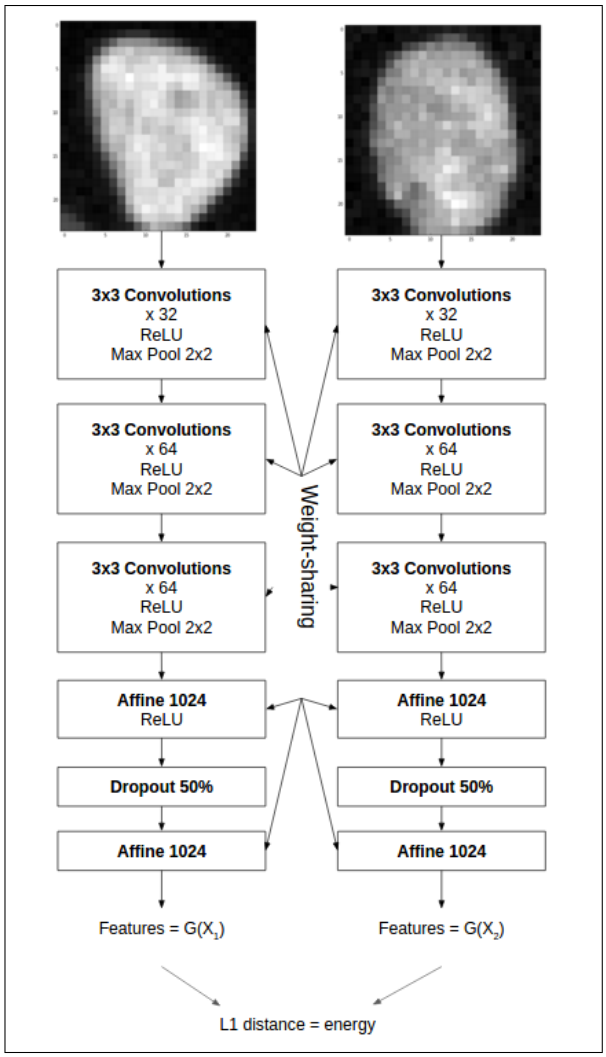


Figure 2. Siamese neural network architecture. Weights are shared between the two halves, such that identical images will produce identical features, and similar images will produce similar features. Inputs are pairs of 31x31 pixel images of cells.

The model developed in this paper improves on linear assignment for frame-to-frame joining of cells by using this visual similarity score instead of euclidean distance as the weight in the cost matrix. This energy score then can replace the euclidean distance given in equation (3) as the cost entries in the matrix from equation (2):

$$l_{ij} = E_{ij} \tag{4}$$

where $E_{ij}$ is the energy score assigned by the siamense neural network between cell $i$ and cell $j$. This means that as long as cells are close enough to be plausibly the same cell between frames (a user-defined paramter), the most similar-looking cell will be chosen instead of the closest.

## 2. Methods

### 2.1. Model Architecture

The model used for pairwise similarity scoring of cell images consists of three layers of 3x3 convolutions. The first layer uses 32 kernels, the second and third layers 64 kernels. Each convolution layer is followed by a Rectified Linear Unit (ReLU) nonlinearity [11], and then downsampling with 2x2 max pooling. The convolutional layers are followed by a fully connected layer with 1024 neurons, followed by a ReLU nonlinearity and then a 50% probablity dropout layer [16]. A final 1024-neuron fully connected layer then produces the features for each image. The energy metric is defined as the L1 distance between the feature sets of two images:

$$\text{Distance}(X_1, X_2) = \|G(X_1) - G(X_2)\| \tag{5}$$

Where $G(X_1)$ is defined as the output from the final fully-connected layer of the neural net on image $X_1$.

See figure 2 for a schematic of the neural net architecture used in this project.

### 2.2. Loss Function

A siamese neural net seeks to produce a low energy score for similar inputs, and a high energy score for dissimilar inputs. Accordingly, it must have a loss function which is high for high-energy input pairs labeled as similar, as well as for low-energy pairs labeled as dissimilar, and low for low-energy pairs labeled as similar and high energy pairs labeled as dissimilar. In other words, the loss function must be monotonically decreasing with respect to the energy when the input is labeled as a similar pair, and monotonically increasing with respect to energy for dissimilar pairs. To acheive this, a piecewise loss function is used:

$$L(X_1, X_2, y) = \begin{cases} F_{\text{decreasing}}\left(E\left(X_1, X_2\right)\right) & \text{if } y = 0 \\ F_{\text{increasing}}\left(E\left(X_1, X_2\right)\right) & \text{if } y = 1 \end{cases} \tag{6}$$

Where $L$ is the loss function, $X_1$ and $X_2$ are an input image pair, and $y$ is a label (0 means dissimilar, 1 means similar), and E is the energy between two images. $F_{\text{increasing}}$ and $F_{\text{decreasing}}$ indicate a monotonically increasing or decreasing function.

One particular loss function, proposed by Chopra et al, was proven to have the increasing and decreasing monotonicity property [5]:

$$L(X_1, X_2, y) = y\left(\frac{2}{Q}\right)E^2 + (1-y)\,(2Q)\,e^{\left(\frac{-2.77}{Q}\right)E} \tag{7}$$

Where $E$ is the energy between $X_1$ and $X_2$, $y$ is the label for the $X_1$, $X_2$ image pair, and $Q$ is the maximum possible energy.

Observe that the first term is monotonically increasing in $E$, and contributes to the loss when $y = 1$. The second term is monotonically decreasing in $E$, and contributes to the loss when the label is dissimilar ($y = 0$). This intuitively demonstrates the property desired; see [5] for a proof.

### 2.3. Training

The model was constructed and trained using TensorFlow [1]. Training was performed with Adam parameter updates [9], minimizing the loss function discussed in section 2.2. Batches of 100 labeled examples were used for training, and a total of 336,000 examples were used to train the models. After the loss function and the energy functions stablized, the learning rate was decreased by a factor of ten, and training was continued until a second convergence. In addition to decreasing loss, a successful model should result in diverging energy values for labeled pairs marked as genuine pairs and labeled pairs marked as imposters. As seen in figure 4

## 3. Dataset

### 3.1. Cell Images

The training and test data were produced in a confocal fluorescent microscopy experiment with 96 imaged positions, and 125 frames per position, with frames captured 15 minutes apart. The cells used for the experiment were mammalian cell-line macrophages, with Red Fluorescent Protein labeling the nuclei. See figure 1 for an example of a single field of view (position) from the microsopy experiment.

### 3.2. Labeling

The cells were segmented using a convolutional neural network separate from the one used in this project [20]. Only the Red Fluorescent Protein channel is captured by the experiment, so the cell cytoplasm is not readily visible. Most of the recognition thus takes place based on the appearence of the nucleus alone. In addition to the

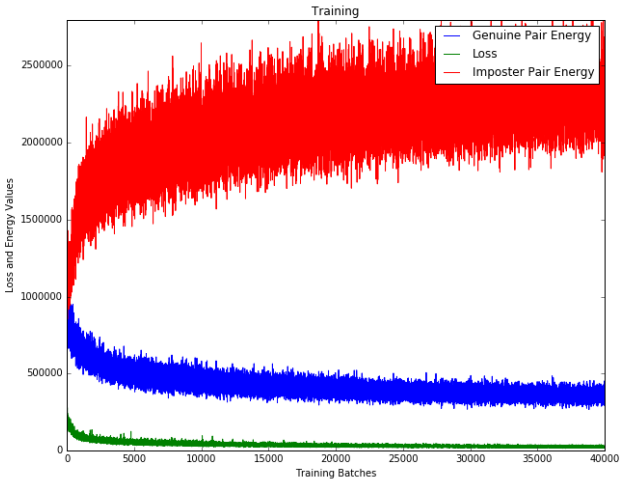


Figure 3. As the loss function decreases, the energy on example pairs which are labeled as the same cell (genuine) decreases, while the energy on pairs marked as different cells (imposter) increases.

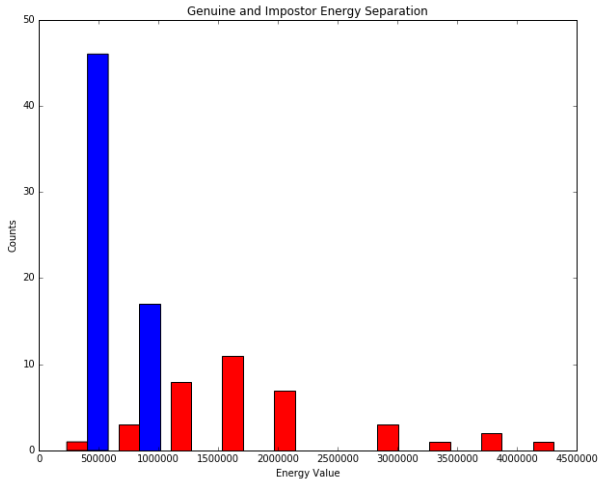


Figure 4. At test time, the model produces good separtion in energy between genuine and impostor pairs.

Red Flourscent Protein channel, the segmentation mask is passed in as a channel to the input. This mask is the output from semantic segmentation, so it marks zeros in parts of the image without a cell, and ones whereever there is one.

The training data were then generated using regular heuristic methods, ie the linear assignment problem with additional methods to increase accuracy at fusion and division points [12], which was then inspected for accuracy. Some of the training data is likely to contain errors, but the testing data is selected from the highest-quality position to ensure that even if some noise is introduced by the rough nature of the dataset, the testing accuracy is believable.

In particular it is worth noting the the chance of errors in the labeling accumulate over longer time periods, so the datasets incorporating cells from greater than one frame separation are likely to contain more noise than those with a separation of only one (see section 3.3). This may account for some of the difference in test accuracy for these models (see tabel 1).

### 3.3. Example Generation

An example image is a 31 by 31 pixel image centered on the computed centroid of a particular segmented cell. The Red Fluorescent Protein channel can contain multiple cells within this field of view, and large cells may be slightly cut off on the sides. See figure 2 for an example of two individual cell input images. Unlike the Red Fluorescent Protein channel, however, the segmentation mask (which is also cut to the same size and centered at the same centroid point) is trimmed so that only the labeled cell is marked. Thus, if additional cells are visible in the sides of the image, the network should be able to learn from the segmented input image which part of the image is the cell under consideration, and ignore or appropriately consider the rest as neighboring cells and background. This will be particularly important in cases where a cell is dividing, as the two daughter cells are likely to be very close together. Clarity about which cell is to be compared is crutial in this case, and is provided by the selective information of the segmentation mask.

Three versions of the dataset were produced and used for training and testing. The first used only cells separated by a single frame as positive examples, the second used a mixture of one-frame separted cells and two-frame separted cells, and the final dataset used only cells separated by two frames. In all cases, an equal number of negative examples as positive examples was produced. Negative examples were generated by pseudorandomly selecting a second cell in the same frame as the first, thus ensuring that the negative pairs are definitely not the same cell at different points in time.

## 4. Results

### 4.1. Model Accuracy

The primary experiment performed in this project is to establish the accuracy of the trained siamese neural network architecture in identifying genuine and imposter pairs of test images. These results are presented in table 1. Overall, the network performs quite well, with 97.5% test accuracy on the best model. As expected, the networks trained on images two frames apart are marginally less accurate

| Dataset | 1 Frame | 1 + 2 Frame Mix | 2 Frame |
|---|---|---|---|
| Test Accuracy | 97.5% | 95.0% | 94.5% |

Table 1. Test accuracy on datasets with examples taken only one frame/timestep apart, taken two timesteps apart, or a mixture. The further apart the test examples are, the more the cells could have changed, making the higher separation examples more challenging.

| Confusion Matrix | | Predicted Class | |
|---|---|---|---|
| | | Genuine | Imposter |
| Actual Class | Genuine | 104 | 5 |
| | Imposter | 4 | 137 |

Table 2. Confusion matrix for the 2-frame model with an energy threshold of 750,000. Area under the curve of the receiver operating characteristic for the was .71.

than those trained on example images generated exclusively from frames which are adjacent in time. Unless otherwise specified, all results and figures are from this final, most challenging dataset and accompanying model.

### 4.2. Energy Difference

The utility of the model is in its abilty to differentiate genuine from imposter image pairs. For the single accuracy metric, a threshold is set to determine which pairs are predicted to be genuine and which imposters. However in actual applications, the scalar difference value may be of more use. In the cell-tracking application, for example, the actual difference value is used. Therefore, in order to better capture the full information of the visual similarity metric, the area under the curve of the receiver operator characteristic is a useful metric. This value captures the full discriminative value of the model, allowing for preference of false positives versus false negatives. The area under the ROC curve is .71

## References


[1] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, L. Kaiser, M. Kudlur, J. Levenberg, D. Man, R. Monga, S. Moore, D. Murray, J. Shlens, B. Steiner, I. Sutskever, P. Tucker, V. Vanhoucke, V. Vasudevan, O. Vinyals, P. Warden, M. Wicke, Y. Yu, and X. Zheng. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. *None*, page 19, 2015.

[2] J. Bromley, J. W. Bentz, L. Bottou, I. Guyon, Y. Lecun, C. Moore, E. Säckinger, and R. Shah. Signature Verification Using a Siamese Time Delay Neural Network. *International Journal of Pattern Recognition and Artificial Intelligence*, 07(04):669–688, 1993.

[3] K. Chen and A. Salman. Extracting Speaker-Specific Information with a Regularized Siamese Deep Network. *Advances in Neural Information Processing Systems*, pages 1–9, 2011.

[4] K. Chen and A. Salman. Learning speaker-specific characteristics with a deep neural architecture. *IEEE Transactions on Neural Networks*, 22(11):1744–1756, 2011.

[5] S. Chopra, R. Hadsell, and Y. Lecun. Learning a similiarty metric discriminatively, with application to face verification. *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 349–356, 2005.

[6] R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality reduction by learning an invariant mapping. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2:1735–1742, 2006.

[7] K. Jaqaman, D. Loerke, M. Mettlen, H. Kuwata, S. Grinstein, S. L. Schmid, and G. Danuser. Robust single-particle tracking in live-cell time-lapse sequences. *Nature Methods*, 5(8):695–702, 2008.

[8] K. Jiang, Q.-M. Liao, and Y. Xiong. A novel white blood cell segmentation scheme based on feature space clustering . *Soft Computing - A Fusion of Foundations, Methodologies and Applications*, 10(1‘):12–19, 2006.

[9] D. Kingma and J. Ba. Adam: A Method for Stochastic Optimization. *arXiv:1412.6980 [cs]*, pages 1–15, 2014.

[10] G. Koch. *Siamese Neural Networks for One-Shot Image Recognition*. PhD thesis, University of Toronto, 2015.

[11] A. Krizhevsky, I. Sulskever, and G. E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. *Advances in Neural Information and Processing Systems (NIPS)*, pages 1–9, 2012.

[12] T. Kudo. Unpublished work.

[13] Y. LeCun and F. J. Huang. Loss Functions for Discriminative Training of Energy-Based Models. 2005.

[14] E. Meijering. Cell Segmentation: 50 Years Down the Road. *IEEE Signal Processing Magazine*, 29(5):140–145, 2012.

[15] M. E. Sieracki, S. E. Reichenbach, and K. L. Webb. Evaluation of automated threshold selection methods for accurately sizing microscopic fluorescent cell by image analysis. *Appl. Envir. Microbiol*, 5598(11):2762–2772, 1989.

[16] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout : A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research (JMLR)*, 15:1929–1958, 2014.

[17] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. DeepFace: Closing the gap to human-level performance in face verification. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1701–1708, 2014.

[18] G. W. Taylor, I. Spiro, C. Bregler, and R. Fergus. Learning invariance through imitation. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2729–2736, 2011.

[19] Y. W. Teh, M. Welling, S. Osindero, and G. E. Hinton. Energy-based models for sparse overcomplete representations. *Journal of Machine Learning Research*, 4:1235–1260, 2003.

[20] D. Van Valen. Unpublished manuscript.

[21] Q. Wu, F. Merchant, and K. R. Castleman. *Microscope Image Processing*, volume 1. Academic Press, Burlington, MA, first edition, 2008.