Base Assumptions:

1. The client has an unlimited budget and wants to get things done efficiently
2. We can obtain as many images of the man as possible.
3. The badge the man is wearing does not track his real time location, only acts as a tool for the sensor to detect.
4. Once the badge has been detected by the counter, the camera will attempt to identify the man.

Challenges:

1. Background clutter
2. Visually similar individuals with different identity
3. Variations in visual appearance (he changes clothing midway)
4. Blurry Video
5. The angle of the video is viewed from the ceiling

Question: Identify which frames in the clip have the staff present?

Solution: Build a Siamese neural network to identify the man if he wears the badge.

Siamese networks are a special kind of neural network architecture, where instead of a model learning to classify its inputs, the Siamese network learns to differentiate between two inputs by learning the similarities between them. The idea behind this is that the network consists of two identical neural networks, each taking one of the two input images, and calculates the similarities between them.

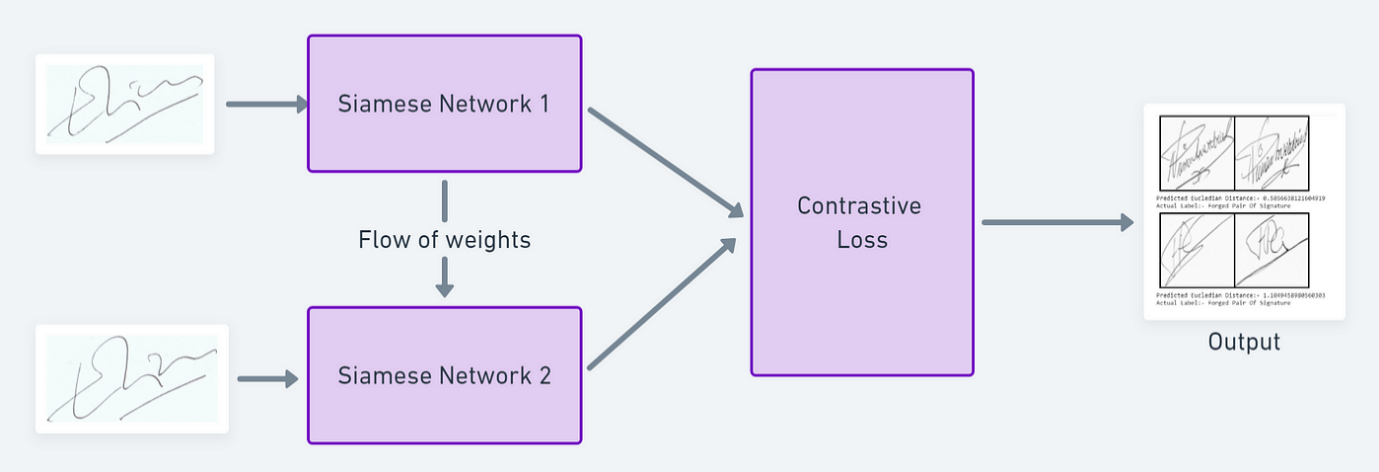


Diagram showing the process of the Siamese network. Image taken from <https://towardsdatascience.com/a-friendly-introduction-to-siamese-networks-85ab17522942>

**Process of training the neural network**

1. First prepare a dataset which contains face images of the man and of other people, which would be used for comparison.

2. Preprocess (resizing, cropping, normalising) the face images such that they are of uniform size and quality.

3. Split the dataset into training, validation and testing sets. The training set will be used to train the neural network, the validation set to tune the hyperparameters of the network, and the test set to evaluate the performance of the network.

4. To train a Siamese network, a pair of images (one of the man and one of another) are picked from the dataset, each image being sent to one of the subnetworks as shown in the diagram above. Each neural network will produce an embedding (feature vector) which captures the unique characteristic of the input images. It then calculates the similarity of the images based on the distance metric used, commonly used metrics are the Euclidean distance or cosine similarity.

5. During testing, the network generates a score based on the chosen distance metric and provides a binary output on whether the faces match.

6. If the faces match, have a feature which allows you to track the man in real time.

An example of a Sianese network can be found here <https://github.com/harveyslash/Facial-Similarity-with-Siamese-Networks-in-Pytorch>

Task 2: Find the man’s xy coordinates.

To start with, if we are only given a single camera, we will not be able to find the 3D coordinates of the man. To achieve this efficiently, a CALIBRATED STEREOVISION SYSTEM can be setup. The idea behind a calibrated stereovision system is to use 2 identical cameras setup in the following way.

If the point P denotes the man in the video, then the points (X,Y,Z) denotes coordinates on the world coordinate system. Suppose we are looking at one point on the left camera, ASSUMING that we already know the correspondence point on the right camera, we could then learn the coordinates of the man.

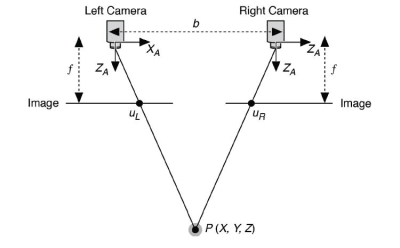


Diagram of a calibrated stereovision system directed towards the man indicated by point P.

Image from <https://www.techbriefs.com/component/content/article/tb/pub/features/articles/14925>