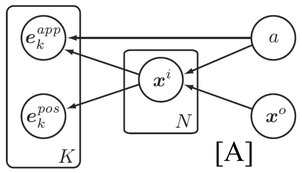
Part-based Human Detection

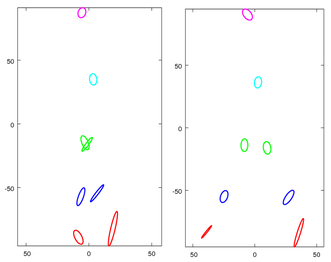
Brandon Tolsch and Joseph Richardson

Part-Based Model

We follow [A] by using a latent variable model to infer the position of the limbs given features observed in the image by maximizing a likelihood function.



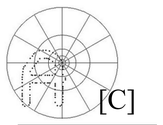
[A]



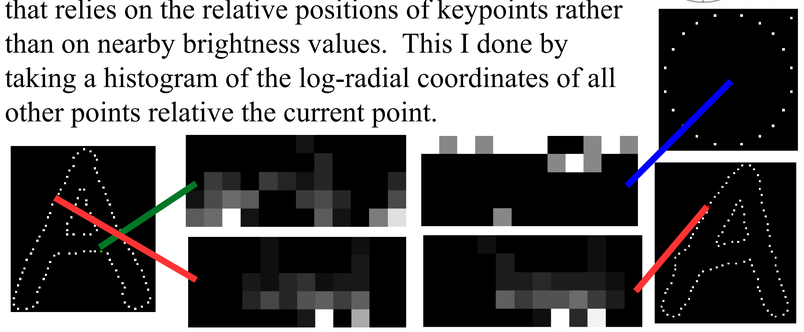
Using an "articulation state" (a) to represent the phase in the walking cycle allows us to make all limb positions (x^i) depend only on the root position (x^0), which allows for more efficient searching for the optimal set of parameters L={xi} (see [E]). e^k_{pos} and e^k_{app} give the position and feature descriptor of each keypoint.

Shape Contexts

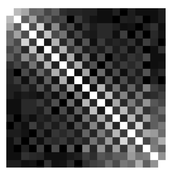
Shape Contexts [B] are a form of feature descriptor that relies on the relative positions of keypoints rather than on nearby brightness values. This I done by taking a histogram of the log-radial coordinates of all other points relative the current point.



[C]



Shape Contexts are generally robust to noise, outliers, and deformations [B]. We repeated some tests from [B], with success, and added our own sanity checks as well. Shown above are some points and their corresponding descriptors; shown right is a similarity matrix between points in the "disc" image.



However, even keypoints that are from the background or other objects can cause trouble, so it is neccecary to limit which keypoints are considered "neighbors". This is difficult when keypoints are too dense...

Overview

We first attempted to implement "People-Tracking-by-Detection and People-Detection-by-Tracking" [A], but quickly found that to be too ambitions for our time constraints. Instead, we opted to implement the first half of their paper, which deals entirely with detection.

While we were able to implement the paper in its basic state, we ran into many problems early on in the pipeline, especially with keypoint detection and feature descriptors, which ultimately resulted in poor results. In addition to discussing the theory of the paper and our implementation of it, we also discuss the reasons for our failures, as well as plans to improve our results.

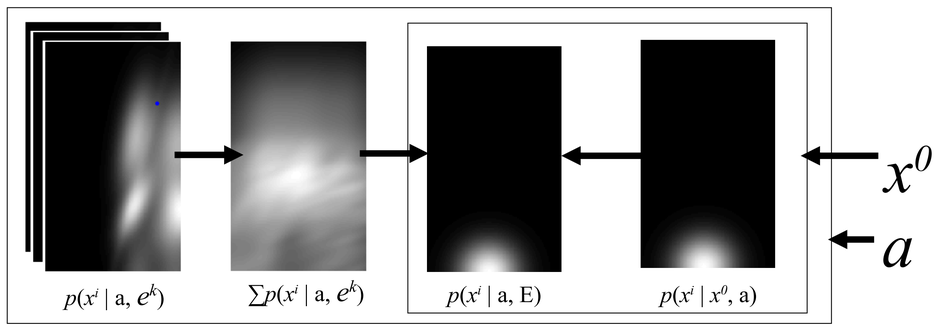
Detection

The final likelihood can be computed from:

$$p(L \mid a, E) \approx \prod_i p(x_i \mid x_0, a)\ [\ \beta + \sum_k p(x_i \mid a, e_k)\ ]$$

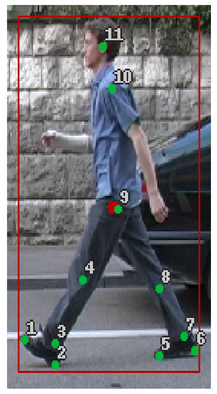
$$p(x_i \mid a, e_k) = \sum\ p(x_i \mid a, c_j, e_k^{pos})\ p(c_j \mid e_k^{app})$$

After finding interest points, feature descriptors, and the probability of cluster membership for each keypoint, we search over the space of possible articulation states and root positions to find the combination with the highest likelihood, $p(L \mid a, E)$. Each limb is placed in the position with the best likelihood, and the overall likelihood is the product of the limb likelihoods. We do this by using "probability images", as shown here.



There are many ways to search over the space of possible x^0 and a. However, we found that even searching exhaustively (ie, naively), our results were very poor, as one would expect by looking at $p(x^i \mid a, e^k)$. Our problems most likely come from an earlier stage.

Training



We were fortunate to have a well-labeled training set (from [A]) giving the L={x^i} set for each person in a series of images, where x^i is the position of the ith part of the person.

In training, we fit several gaussian distributions:

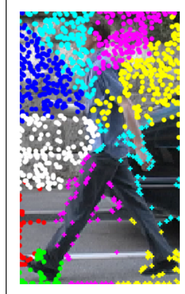
- $p(a)$
- $p(x^i - x^0 \mid a{=}1, 2, \ldots)$
- $p(x^i - e^k_{pos} \mid a{=}1, 2, \ldots, c_j{=}1, 2, ..)$

We also use clustering on shape contexts, and state that $p(c_j \mid e^k_{app})$ is proportional to the L2 norm of $(c_j - e^k_{app})$. The cluster centers c_j are also needed in order to fit $p(x^i \mid a, c_j, e^k_{pos})$.

Problems, and Future Work

Unfortunately, we were unable to obtain any detections that were remotely meaningful. We faced two main problems:

Keypoints and Feature Descriptors



Because keypoints were very dense and uniform, we were unable to properly localize their neighborhoods, and the Shape Contexts were nearly meaningless, resulting inpoor values for $p(x^i \mid a, c_j, e^k_{pos})$.

The next step is to find a better keypoint detector, such as the Harris-Laplace detector [F], and to have more meaningful keypoint sets. Alternately, we could simply abandon Shape Contexts, as in [D].

Speeding up likelihood maximization

Searching over p(L | a, E) can be greatly sped up, something which we did not bother with on account of the previous issue, but which is a major concern once we are able to get meaningful results.

There are many options on how to do this, such as Dynamic Programming (as in [B]), gradient descent, or iteratively narrowing and refining an initially coarse search space.

References

A) Andriluka, Roth, and Schiele, "People-Tracking-by-Detection and People-Detection-by-Tracking", https://www.d2.mpi-inf.mpg.de/andriluka_cvpr08

B) Belongie, Malik, and Puzicha, "Shape Context: A new descriptor for shape matching and object recognition", http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.27.8567&rep=rep1&type=pdf

C) https://en.wikipedia.org/wiki/Shape_context#/media/File:Shapecontext.jpg

D) Leibe, Seemann, and Schiele, "Pedestrian Detection in Crowded Scenes", http://luthuli.cs.uiuc.edu/~daf/courses/AppCV/Papers-2/leibe-crowdedscenes-cvpr05.pdf

E) Felzenszwalb and Huttenlocher, "Pictorial Structurs for Object Recognition", http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.66.5153&rep=rep1&type=pdf

F) Mikolajczyk and Schmid, "Scale & Affine Invariant Interest Point Detectors", http://www.robots.ox.ac.uk/~vgg/research/affine/det_eval_files/mikolajczyk_ijcv2004.pdf