Estimating Human Pose in Images

Navraj Singh
December 11, 2009

Introduction

This project attempts to improve the performance of an existing method of estimating the pose of humans in still images. Tasks such as object detection and classification have received much attention already in the literature. However, sometimes we are interested in more detailed aspects of objects like pose. This is a challenging task due to the large variety of poses an object can take in a variety of settings. For human pose estimation, aspects such as clothing, occlusion of body parts, etc. make the task even harder.

The approaches taken up in the literature to solve this problem focus on either a top-down approach, bottom-up approach, or a hybrid of the two. The top-down approach involves comparing test images with stored examples of humans in various poses using some similarity measure. This approach might require a very large set of examples of human poses. The bottom-up approach, on the other hand, uses low level human body part detectors and in some manner assembles the information to predict the entire body pose. This project attempts to build upon a mostly bottom-up approach, called LOOPS (Localizing Object Outlines using Probabilistic Shape), that was developed in [1] by G. Heitz, et al. in Prof. Daphne Koller's group. Specifically, we investigate the construction and incorporation of a skin detector into the LOOPS pipeline, and a couple of pairwise features in the appearance model. The overall improvement in the localization is negligible, with some improvement in head localization. Since the improvements considered are within the framework of LOOPS, a brief overview of the LOOPS method is discussed next.

Brief Overview of the LOOPS method as applied to humans

The main random variables defined in the LOOPS method, described in detail in [1], are the locations of a set of key "landmarks" that define an object outline. For humans, we consider the skeleton instead of the outline, and define 14 key landmarks for the skeleton (landmark #1: right foot, #2: right knee, #3: right hip joint, #4: left hip joint, #5: left knee, #6: left foot, #7: right hand, #8: right elbow, #9: right shoulder, #10: left shoulder, #11: left elbow, #12: left hand, #13: neck, #14: head). The method uses a probabilistic model that combines a shape model over the landmark locations with appearance based boosted detectors (for each individual landmark) and some pairwise features over appropriate pairs of landmarks. So, the model defines a joint distribution over the location of these key corresponded landmarks of the human skeleton, as shown next.

The LOOPS Model

Briefly, the shape of an object class is defined by locations of the N object landmarks, each of which is assigned to a pixel in an image. With $\boldsymbol{L}$ denoting the vector of image pixel locations assigned to the landmarks, the probability distribution over $\boldsymbol{L}$, conditioned on a given image I, is a Markov Random Field [1][2]:

$$P(\boldsymbol{L} \mid I, \boldsymbol{w}, \mu, \Sigma) = \frac{1}{Z(I)} P_{Shape}(\boldsymbol{L}; \mu, \Sigma) \prod_i \exp(w_i F_i^{det}(l_i; I)) \prod_{i,j} \exp(w_{ij} F_{ij}(l_i, l_j; I)) \qquad (1)$$

Here, $\mu, \boldsymbol{w}, \sum$ are the model parameters, and i and j index the landmarks of the skeleton. P_{Shape} represents the unnormalized distribution over the object shape, $F_i^{det}(l_i)$ is a detector for landmark i, and F_{ij} are pairwise features over appropriate pairs of landmarks. The notation in the last product term might be a bit misleading, as we can have more than one type of pairwise feature (or even "threewise" and "fourwise" features) over groups of landmarks. The shape model and the detector features are learned in parallel. As it is discussed in [1], in principle the weights $\boldsymbol{w}$ could be learned from data, but the process requires an expensive inference step at each iteration. The authors' experiments indicated that the results are relatively robust to a range of the weights. So, a fixed set of weights is used (e.g. $w_i = 5$, $w_{ij} = 1$, for all i,j).

Shape Model

The shape component of (1) is modeled as a multivariate Gaussian distribution over the landmark locations with mean μ and covariance $\sum$. The Gaussian form decomposes into potentials over only singletons and pairs of variables as described in [1]. During inference, however, a sparse approximation of the shape model is first used since the general Gaussian includes pairwise terms between all landmarks. After this "discrete inference" stage, a "refined inference" step occurs that involves the full Gaussian. The approximation to the maximum likelihood parameters of the full Gaussian (which can be solved analytically) is obtained by minimizing the KL divergence between the sparse and the full maximum likelihood parameters. The details are again given in [1]. The results shown later are all at the end of the refined inference stage.

Landmark Detector Features and Pairwise Features

To construct the landmark specific detectors, the well known Boosting is used. That is, the feature in the MRF for the assignment of landmark i to pixel l_i is then given by:

$$F_i^{det}(l_i; I) = H_i(l_i)$$

Here, $H_i(l_i)$ is a strong classifier whose output is proportional to the log-odds of landmark i being at pixel l_i. The construction of this boosted classifier from weak detectors is described in [1]. The main weak feature detectors used are randomly extracted patches from object bounding boxes in filtered versions of training images. The patch is matched to a test image using cross-correlation. More details on how the weak detector is applied to a test image are given in [1]. Importantly, after learning the boosted classifiers, we obtain a response map for each landmark for a given test image. These response maps are then used in conjunction with the skin detector as described in the next section.

The pairwise features we tested were a feature between some adjacent pairs of landmarks that encodes the color variance along the line segment connecting the two landmarks (since usually human figures have low color variance along adjacent landmarks), and a feature encoding similarity in color space of symmetric landmarks (e.g. left side hands, feet/shoes, knees, etc. usually have similar color appearance to their right side counterparts).

Localization/Inference

Using the MRF definition of the distribution over the assignments of the model landmarks to pixels, we can outline objects by finding the most probable assignment:

$$\boldsymbol{L}^* = argmax_{\boldsymbol{L}} P(\boldsymbol{L} | I, \boldsymbol{w})$$

The method is not that straight forward, however, as this involves inference over a very dense MRF. A method that involves combination of pruning down the interesting pixels to consider and performing a "discrete inference" is discussed in [1].

The Dataset

For an arbitrary object class, the task of deciding on key landmarks of the object outline and obtaining corresponded training outlines is a tedious one, and [1] describes a way to do this automatically. However, for simplicity, for localizing humans we just use manually labeled training skeletons. In particular, the manually labeled people dataset made available by Deva Ramanan at http://www.ics.uci.edu/~dramanan/papers/parse/index.html is used. The dataset consists of 305 images, all scaled to contain people of roughly 150 pixels in height. We use the first 100 images for training, and the remaining 205 for testing.

Incorporating Skin Detection

One specific aspect of human figures that is missing from the model above is skin color. To learn a good skin detector in a supervised manner, however, one needs labeled training examples of skin and non-skin pixels, which can be a tedious task. While we couldn't find a properly labeled dataset to train a skin detector, we did get access to the H3D dataset from UC Berkeley, which contains several extensively annotated images containing humans. While the pixels are not given any labels, the dataset does contain ground truth segmentations of human body parts. The skin pixels were therefore obtained by extracting the patches corresponding to faces and hands, as these parts are almost always exposed and show some skin. A total of about 2 million skin pixels were obtained in this manner. Random non-skin patches were used to obtain about 10 million non-skin training examples. We note that the skin labels were indeed a bit noisy, as some hands might be covered by gloves, some faces might have sunglasses on, etc. A quick qualitative look at the extracted patches, however, showed that most labels were correct.

Instead of learning a hard skin classifier, in our case we only needed a classifier that outputs the log-odds of a pixel belonging to a skin patch. The first approach tried was to learn a generative model using a single Gaussian to model the skin distribution. The features tried were color-based. Once the skin pixel data was extracted from the H3D dataset, finding the maximum likelihood parameters for the Gaussian Discriminant Analysis model was easily done in closed form. The color spaces investigated were RGB, CIE-LAB, and HSV, with HSV providing slightly better results than the other two. To improve the model further, however, we used a mixture of two Gaussians to capture a larger variation of skin color seen in human images. The standard EM algorithm was used to learn the mean and covariance of the two Gaussians. A brief comparison of the ROC curves between using a single Gaussian and a mixture of two Gaussians (both in HSV color space) is show below. The improvement is marginal, and both models result in a satisfactory skin detector.

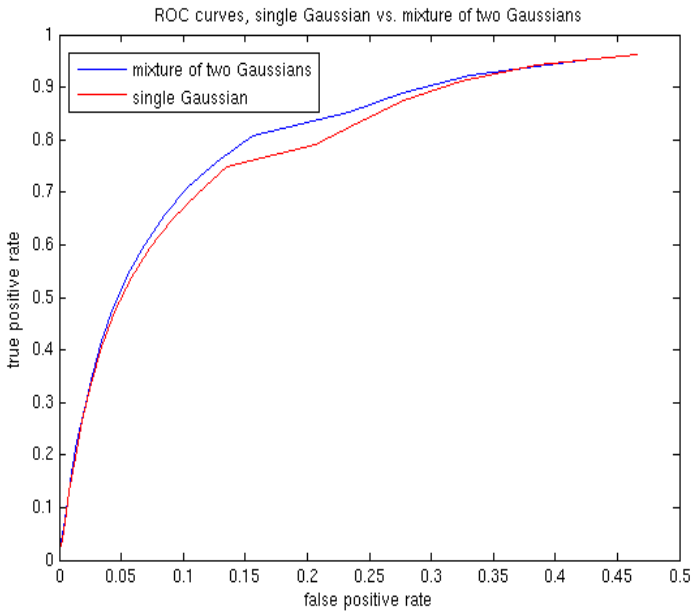

Figure 3. ROC curves for skin detection using single Gaussian (red) vs. a mixture of two Gaussians (blue).

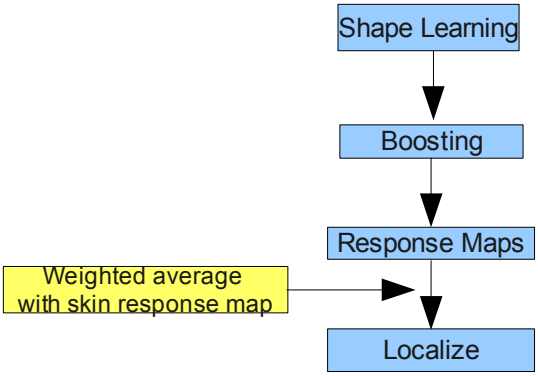

Figure 2. Incorporation of skin response maps into the LOOPS pipeline.

While the ROC curve was obtained by varying the detection threshold and then looking at the output binary skin response maps, for incorporating the skin detection into LOOPS, we only take the soft skin response maps (i.e. the pre-threshold response maps obtained by applying Bayes' Rule to the learned generative model to find the probability of a pixel belonging to a skin patch, and setting the output image's intensity at that pixel location to be this log-probability). In the LOOPS pipeline, after the landmark response maps for test images are computed from the boosted classifiers, we compute a weighted average of the head and hands response maps with a Gaussian-blurred version (to reduce noisy peaks for discrete inference stage) of the soft skin response map for that image. Only the head and hands landmark responses are averaged in this way, since these are most likely to show skin. Skin response weights of 0.2, 0.4, 0.6, and 0.8 were tested in the averaging step, with a weight of 0.6 resulting in the most improvement in the average landmark based error metric that's shown later. A brief illustration of where the skin detection is incorporated into LOOPS is shown in figure 2. Some examples of improved overall localizations as a result of skin detection are shown in Figure 3.

Pairwise Features

Color variance between adjacent landmarks

The original LOOPS method as described in [1] includes a pairwise feature that encodes a preference for aligning outline segments along image edges. Since in the human case we are dealing with skeletons rather than outlines, we don't use this gradient feature. Instead, we introduce a pairwise feature between some adjacent landmarks (such as elbow-shoulder, hip-knee) that encodes a preference for low color variance (in HSV space) along the line segment joining the two landmarks. This is based on the observation that most human figures (at least in the dataset used here), show low color variance along limbs (unless wearing colorful clothes with lots of patterns). To encourage low variance, we set the feature value to be the negative of the sum of the color variance in H and S space of the pixel values along the line segment. The weights tested for this features were 1, 5, and 10, with all three weights resulting in similar performance.

Color similarity between left & right symmetric landmarks

A second pairwise feature tested was the similarity of appearance between left & right hand, left and right foot, etc. So, this feature gets a value that's negative of the difference in color (again, in HSV color space) between the candidate locations of two symmetric landmarks. To reduce noise, we compute the average color in a small 3x3 pixel patch around the candidate location in a blurred version (with a Gaussian of standard deviation of 3) of the test image. The weights tested in the model for this feature were again 1, 5, and 10, with minimal variation between them.

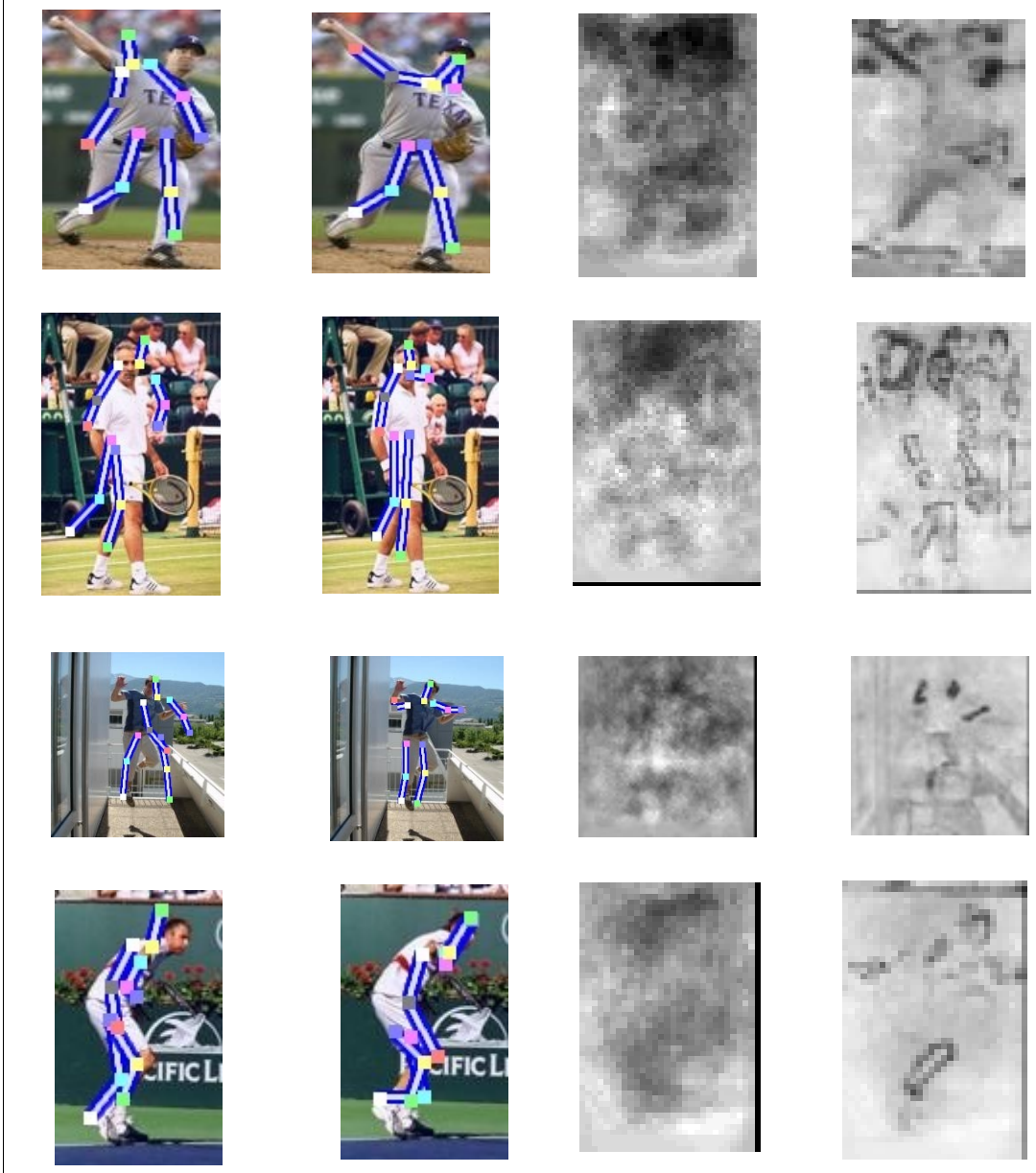

Figure 3. Some improved localizations due to skin detection. First column shows the baseline localization, second column shows localization after averaging the skin response maps (with weight 0.6), the third and fourth columns show as an example the original and skin averaged response maps for the head landmark.

Results

A landmark based error metric was used to see the effects of these methods . Specifically, the metric is the rms distance (mean calculated over the entire test set) between the ground-truth landmark location and the localized landmark location, normalized with respect to the size of the bounding box around the human figure in the test image. The errors for the 14 landmarks, for the baseline (LOOPS method without any of the additions described here), LOOPS + skin detector, LOOPS + skin detector along with the color variance feature, and skin detector along with the symmetry feature, are shown in the next image. The largest errors are seen in localizing both the left and right hands. The largest improvement (close to about 20% over baseline) is seen in locating the head.

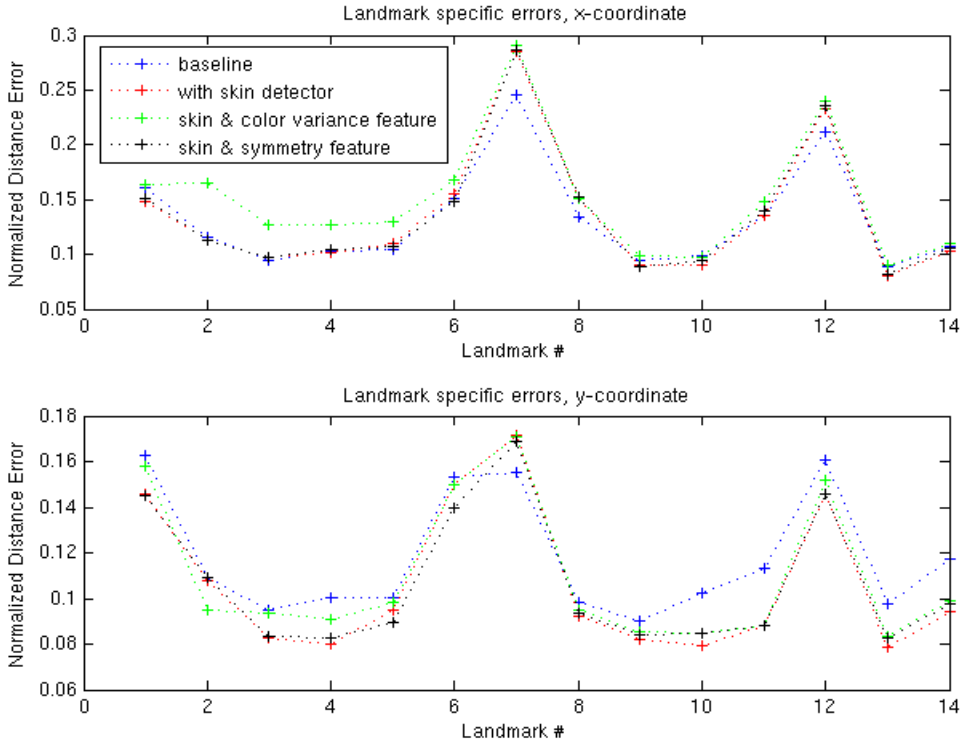

Figure 4. Landmark specific errors in x and y directions. Landmark numbers are as described earlier. The largest errors (the two peaks in the middle) are for the left/right hands. Most improvement due to skin detection is seen in head localization. The pairwise features in general do not lead to improvements.

Conclusions/Future Work

The results show that incorporating a skin detector mostly helps in localizing the head, but doesn't help much for hands on average. A better, adaptive skin detector with a richer set of color-based and geometric features, such as that developed in [4], can be used to see if better skin detection leads to improvements. The pairwise features considered here do not lead to significant improvements, either. Another pairwise feature that could be tried is one that encodes a preference for requiring the localized parts to lie in the image foreground, assuming the image background/foreground can be reliably segmented. In general, the largest errors occur in localizing hands. This perhaps indicates the need for a parts based model as used in [3].

A large part of the effort in this project went into understanding LOOPS and its infrastructure. With more experience with the LOOPS code-base, experimentation could be done with better features in the boosting stage (for example, in addition to using the randomly selected filter response patch features, more human part specific features could be experimented with). On a long term basis, integration of LOOPS with holistic scene understanding seems to be an interesting direction to take. For example, given a test image, if we can determine its class (e.g. Ballet), then a class-specific shape model can be used that has a better tolerance for out-of-the-ordinary articulation of human figures.

Acknowledgements

Thanks to Ben Packer and Tianshi Gao (from Prof. Daphne Koller's group) for providing assistance in understanding the LOOPS pipeline and giving access to their LOOPS code library and assisting in adding pairwise potentials. Also thanks to Stephen Gould for helpful suggestions on skin detection.

References

[1] G. Heitz, G. Elidan, B. Packer, and D. Koller. Shape-based object localization for descriptive classification. NIPS, 2008.
[2] J. Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann. 1988.
[3] L. Bourdev and J. Malik. Poselets: Body Part Detectors Trained Using 3D Human Pose Annotations. ICCV. 2009.
[4] Q. Zhu, K. Cheng, et al. Adaptive Learning of an Accurate Skin-Color Model. IEEE International Conf. on Automatic Face and Gesture Recognition, 2004.