Clothed People Detection in Still Images

Nathan Sprague
Computer Science Department
University of Rochester
Rochester, NY 14627
sprague@cs.rochester.edu

Jiebo Luo
Imaging Research and Advanced Development
Eastman Kodak Company
Rochester, NY 14650
luo@image.kodak.com

Abstract

We present a trainable system for locating clothed people in photographic images. People detection is a particularly challenging image understanding problem; as a result of variations in clothing and posture, the appearance of people may vary enormously from image to image. Our approach attempts to construct a maximally person-like assembly of image regions, where candidate regions are provided by color-based segmentation followed by non-purposive grouping. A tree structured probability model is employed to allow efficient searches. This structure represents the pairwise configuration of body parts as a function of relative position, relative size, and adjacency. Face and skin detection is also used to help the search. The problem of occlusion is addressed through a mixture of trees, where the different mixture components represent the possible subsets of visible parts. Different clothing styles are accounted for by separate models. Experimental results are shown to demonstrate the promise of and challenges for the current system.

1. Introduction

The purpose of this work is to detect clothed people in digital images. The problem of people detection warrants special attention because people are the primary objects of interest in a majority of photographs. It is necessary to consider the problem of people detection separately from general object detection, because the appearance of people varies so widely. For many object recognition tasks, simple appearance models based on low level image features suffice. However, for detection of human figures, which are compound objects, local appearance information is clearly insufficient; variations in the appearance of components (body parts), in particular clothing, as well as variations in pose, result in a nearly endless array of color and shape profiles. Therefore, a successful people detection system requires a high degree of semantic understanding.

2. Related work

We are unaware of any previously published work that addresses exactly the problem discussed here: single image clothed people detection in unrestricted poses. Existing systems generally tackle a more constrained version of the problem, such as limiting detection to naked people [2], or upright pedestrians [6]. Systems that do not enforce such constraints take advantage of additional information such as motion [1] or depth [7].

3. Detection by assembling regions

Our person detection system works as follows. First, an image is partitioned into regions using color-based segmentation [5] and non-purposive grouping [3]. These regions are the basic building blocks of the search mechanism. The goal is to select an assembly of regions that is maximally person-like. The region-based approach is attractive because it vastly reduces the search space relative to dealing directly with pixel-level information. In addition, region segmentation facilitates the definition of geometric attributes.

The search is guided by the output of face-detection and skin-detection modules, each of which produces a binary result for each image region. Face detection plays an important role because the face is the only body part that can be reliably detected using existing algorithms in photographs whose content is unrestricted. We take advantage of this by constraining the solution to be consistent with the output of the face detection module. The data that skin detection provides is less reliable, but still serves as a useful cue. In the current implementation, skin detection is handled by a color based algorithm, while face detection is simulated by referring to human generated labels.

The proposed algorithm for person construction is based largely on the thesis work of Sergey Ioffe [4]. A more detailed derivation of the basic algorithm is available there, as well as examples of its application to the problems of tracking naked people and face detection. We extend his work to tackle the more challenging problem of single image people

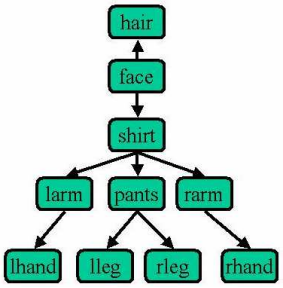

Figure 1. The nodes of this tree represent potentially segmentable human parts, while the edges represent distributions over the configurations of those parts.

detection.

The search algorithm employs a tree structured probability model to represent person-like assemblies of regions. Nodes in the tree represent segmentable body parts such as face, hair, and shirt (torso), while the edges in the tree represent the distributions over relative configurations of those parts. Tree structured models are attractive for two reasons. First, such models mirror the actual structure of a human body, which is a tree structure of more or less rigid segments connected by joints. The second benefit is that a tree structured model allows efficient searches based on dynamic programming.

Assume that the appearance of a person is represented as an assembly of K image regions, $\{X_1, ..., X_k\}$ where each X_i can be treated as a vector that represents the configuration[1] associated with a particular body part. We can then define a distribution $P(X_1, ..., X_k)$ that represents the range of assemblies that are "person-like"; the value of $P(X_1, ..., X_k)$ is high for assemblies of image regions that look like a person, and low for assemblies that do not. The goal of person detection is then to search through the set of *all* regions in an image for the K regions that maximize the value of $P(X_1, ..., X_k)$. For an arbitrary representation of $P(X_1, ..., X_k)$ this search is prohibitively expensive; to check all assemblies in a brute force fashion takes $O(M^K)$ time, where M is the total number of image regions.

The key advantage of the tree structure is that it allows the distribution to be factored:

$$P(X_{\text{root}}) \prod_{k \neq \text{root}} P(X_k) P(X_k | X_{\text{Pa}_k}) \tag{1}$$

Here $P(X_k)$ represents the distribution of configurations associated with node k of the tree, while $P(X_k|X_{\text{Pa}_k})$ represents the distribution of configurations for node k *relative* to the configuration of k's parent in the tree. This factored form allows efficient search, and greatly simplifies the problem of learning the distribution.

Specifying the distribution requires both selecting the graphical structure of the tree, and the parameters of the priors $P(X_k)$ and conditionals $P(X_k|X_{\text{Pa}_k})$. It is possible to learn both the structure and parameterization of the tree directly from the data (as is done in [4]). However, the structure of the tree is specified by hand for the system presented in this paper. As can be seen in Figure 1, we select a graphical structure that is consistent with the physical structure of the human body. The specific parameterizations and learning procedures used for the distributions $P(X_k)$ and $P(X_k|X_{\text{Pa}_k})$ will be described in later sections. The remainder of this section will describe the search procedure, assuming that these distributions are known.

In practice, we do not want to maximize $P(X_1, ..., X_k|\text{human})$ as suggested above, but instead the Bayes factor

$$\frac{\text{P}(X_1, ..., X_k|\text{human})}{\text{P}(X_1, ..., X_k|\text{background})}. \tag{2}$$

The intuition is that we wish to find a set of regions that is likely to appear in a random view of a person, but unlikely to appear in a random view of the background.

For the time being, the model of the background distribution is simple; it is assumed that every individual region has a fixed probability β of belonging to the background. Therefore, the maximum of the Bayes factor can be written as:

$$BF^*_{ij}(X_j) = \max_{X_i} \frac{\text{P}(X_i)\text{P}(X_i|X_j)}{\beta} \prod_{k:X_i=\text{Pa}_k} BF^*_{ki}(X_i) \tag{3}$$

This function expresses the maximum at a node i as a function of all possible values of the node's parent j, taking into consideration the values of i's children. A search algorithm based on dynamic programming follows immediately from this formula. This procedure allows us to efficiently discover the global maximum of the Bayes factor.

3.1. Mixtures of trees

A weakness of the search procedure outlined above is that it does not make allowances for missing nodes. It is possible that some of a person's body parts will not be visible, either due to occlusion, or because of errors in the segmentation phase. One way of dealing with this is to use a mixture of trees instead of a single tree. Each tree in the mixture represents some subset of visible nodes, and the choice variables represent the probability of observing each of the given subsets. The problem with this approach is that

[1] By configuration we mean the properties of a region, including size, position, shape, orientation, or whatever other properties we choose to encode.

it would require a large number of mixture components to represent all possible subsets of visible nodes.

To address this, Ioffe introduces the elegant idea of a *generating tree* [4]. The generating tree compactly represents a large number of mixture components by making the assumption that different components differ only in the set of visible nodes. The generating tree not only represents the conditional distribution of a child node relative to a parent, but also the distribution of the child node relative to the grand-parent, the great-grand-parent, etc. This way, if a node's parent is missing, it can be adopted by one of its ancestors.

Search in the generating tree can be performed using a similar dynamic programming approach as was presented above. The difference is that in the generating tree we not only search for the best configuration for each node, but also whether or not each node should be visible to create the most person-like assembly.

The maximum of the Bayes factor in the generating tree can be written as:

$$BF^*_{ij}(X_j) = \max\Big(\mathrm{P}([\overline{X_i}]|[X_j]) \prod_{k:X_j=\mathrm{Pa}_k} BF^*_{kj}(X_j),$$
$$\mathrm{P}([X_i]|[X_j]) \max_{X_i} \frac{\mathrm{P}(X_i)\mathrm{P}(X_i|X_j)}{\beta} \prod_{k:X_i=\mathrm{Pa}_k} BF^*_{ki}(X_i)\Big) \quad (4)$$

Bracketed values in this expression represent the probability that a given node will be visible. For example, $\mathrm{P}([X_i]|[X_j])$ encodes the probability that node i will be visible given that node j is visible. These values are learned by directly counting occurrences in the training set. This equation differs from equation (3) in that node j need not be the parent of node i; it can be any ancestor.

Informally, this equation says that we should trade off the quality of the best match at a given node (the bottom line), against value of leaving the node out entirely (top line). The running time of the associated algorithm is $O(M^2\#\mathrm{edges})$, or equivalently, $O(M^2K^2)$.

The value β, introduced earlier, plays an important role in equation (4). Varying this value allows us to control the degree to which we find more complete assemblies at the expense of accepting lower quality matches. An appropriate value for this parameter must be determined experimentally.

3.2. Handling over-segmentation

The outlined search procedure will only be effective to the extent that the regions produced by segmentation and non-purposive grouping correspond to the body parts that are represented in the person model. In practice this condition is often violated when a single body part is broken up into multiple regions due to shadows or patterns of color in the clothing. This problem is addressed by modifying the set of candidate body parts to include not only individual regions, but also sets of neighboring regions. This allows body parts that are broken into multiple regions to be reconstructed. This can be seen as a kind of purposive grouping; we use the learned person model to group image regions that cannot be grouped based purely on low level information [3].

Unfortunately, the running time of the search algorithm is sensitive to the size of the candidate set. We limit the number of candidates that must be considered by adding only pairs and triplets of neighbors that pass a set of tests for compactness and size.

3.3. Handling different clothing styles

The search mechanism described above is not sufficiently general to handle any style of clothing. Different clothing styles will result in distributions of image regions that are too divergent to be captured by a single model. For example, it would be impossible to define a single tree that is adequate for detecting people in formal gowns as well as for people wearing bathing suits.

This is addressed by training different generating trees for different clothing styles. At present, we have two different models, a shirt model, and a dress model. Detection is performed by running both of these models and selecting the one with a higher match value.

4. Parameterization

The description above does not address the important question of how exactly the distributions $P(X_i)$ and $P(X_i|X_j)$ are parameterized and learned. The following factors are considered in the distribution $P(X_i|X_j)$: size, position of centroid, position of bounding box, and adjacency. $P(X_i)$ is based only on the output of the skin detector.

To minimize the amount of training data that is required, each of the components of $P(X_i|X_j)$ are assumed to be independent. This enables each to be learned separately. Obviously, the independence assumption is not entirely valid; for example, the position of a region's centroid and its bounding box will clearly be related. However, it is hoped that the resulting distribution will serve as a sufficient approximation.

There are many ways to approximate multidimensional distributions: Gaussians, mixtures of Gaussians, neural networks, etc. We chose to use quantized histograms, because they are simple to implement and train, and they do not make strong assumptions about the distribution of the data.

Due to space considerations, this paper will not discuss in detail how the values considered in $P(X_i|X_j)$ are computed. However, it should be noted that each of these values

is computed in a scale invariant fashion, and a global coordinate system is provided by the face detection module.

5. Post-processing

Even if the person detection procedure is successful, in the sense that the major segments of the person are associated with the correct body parts, there will often be some gaps in the result. Since these gaps are much more likely to result from over-segmentation than from occlusion, it is desirable to fill them in. This is currently handled by a simple procedure that iteratively adds regions that are holes or near holes in the person profile.

6. Training

Ground truth data for training the model described above was obtained by hand labeling the specified body parts for a set of 118 images containing people. The bulk of these images (87) were professional photographs obtained from the Corel image set, while the remainder (31) were from a set of cruise vacation photographs. The entire set of photographs contained 153 individuals of which 96 were labeled as wearing shirts, and 60 were labeled as wearing dresses. In order to effectively increase the amount of training data, the training set was augmented with the set of mirror images.

Given the simple form of the distributions mentioned above, the process of training the system is straightforward. Simply iterate through the set of training images creating a data point for every region and every region-ancestor pair that is labeled. For each of these points compute the relative size, position, etc. and fill in the appropriate histograms using these values.

7. Experimental results

It is not entirely straightforward to quantify the success of the people detection system. It is unrealistic to demand a perfect correspondence between the detected body parts and ground truth. Such a perfect correspondence is often impossible due to problems with the segmentation. Even when the segmentation is good there may be some ambiguity concerning exactly where each body part is located.

For many practical applications, we will be happy with the results if the overall outline of the person is mostly correct, even if some individual body parts are missed or incorrectly classified. The following measure will be used to quantify "mostly correct":

$$\frac{\mathrm{AREA}((M \cap \overline{GT}) \cup (\overline{M} \cap GT))}{\mathrm{AREA}(GT)} \qquad (5)$$

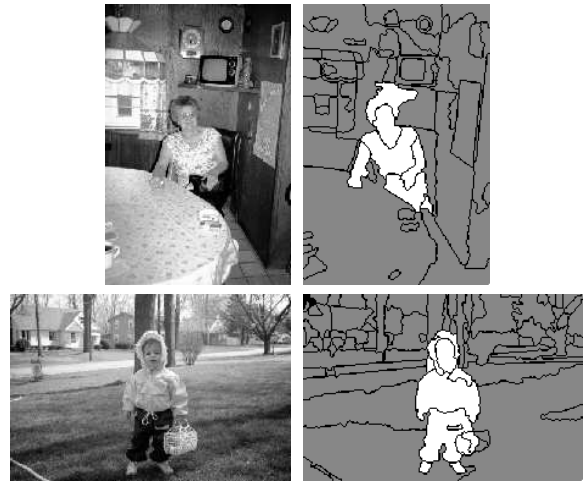

Figure 2. Examples of success. The black lines indicate region boundaries. The white regions indicate the most likely person assembly.

where M is the region that is selected by the person finder, and GT is actual area covered by the person. This error term is equal to zero when the output of the person finder is a perfect match with ground truth. We will consider the output of person detection to be a success when the value is less than 0.4.

The system is tested on two sets of images, TRAIN is the training set of images discussed above, and TEST is a separate set of 39 consumer images containing a total of 63 individuals. Both of these data sets consist of images in which people are the main subject. Very difficult cases such as bizarre costumes were excluded. However, the sets do contain a great deal of variation. There are people in a wide range of poses, people are partially occluded, and there is a great diversity of ages, ethnicities, and clothing styles. Many images contain background clutter that accounts for a large majority of the image regions.

The system successfully locates 28.7% of the individuals in the training set, 36.5% in the test set, and 30.9% overall. The performance is higher on the test set than on the training set, possibly because the test set contains fewer difficult cases or the results of the segmentation procedure are slightly better. Figure 2 shows examples of successful detection results after post-processing has been performed.

An overall success rate of 31% is not too discouraging for a first attempt at this difficult problem on images with unconstrained scene content and background clutter, but it is low enough to warrant an examination of why the system fails when it does.

The errors can be grouped into several categories. One

Figure 3. Examples of failure. In the top image the people detector fails due to under-segmentation. The woman's dress is segmented into a single large region that includes much of the background. In the bottom image the failure is a result of over-segmentation. The woman's shirt is broken up into many small regions

major source of error is imperfect segmentation (see Figure 3), e.g., highly patterned clothing that is broken into a large number of regions. Since the system is currently able to group no more than three regions, it is unable to account for such over-segmentation. Failure also occurs when large parts of the person's body are grouped with the background due to under-segmentation. This problem clearly cannot be addressed through the mechanisms described in this paper. Shape based parts decomposition is expected to help break up the body parts from the background to a certain degree.

Another problem with the current system is that there is no mechanism to deal with self occlusion. A typical scenario that causes this problem is a person with arms crossed in front of them. This will tend to break the shirt into two non-contiguous regions, one above the arms and one below. Since only adjacent regions may be grouped, currently there is no way to correctly group the two shirt regions into one.

Finally, the system may fail when a particular configuration is not represented by the trained model. The solution to this class of failures is continued refinement of the model, and the introduction of more representative training data.

8. Future directions

One direction for further research is the development of better models for individual body parts. Although it is probably not possible to model individual body parts (except for faces) well enough for them to be detected in the absence of any additional context, it should be possible to improve the current models. Different body parts have different characteristic shapes, as well as different distributions of colors. For example, an elongated blue region is unlikely to be hair, but could easily be an arm.

Another major opportunity is shape based parts decomposition. Most of the variations in human poses are in the posture of the limbs relative to the torso. The limbs are frequently *not* separated from the clothing if skin is not exposed. Consequently, in the current system, the probability density functions of both the limbs and clothing are not as discriminative as hoped. With parts decomposition, we expect that the region segments will become fairly simply shaped (rectangular in 2D or cylindrical in 3D). Therefore, measures of position can be more accurately computed. The graphical model will become more like a "stick-figure" model, which has been used successfully in robotics, tracking, and computer graphics animation.

References

[1] C. Bregler. Learning and recognizing human dynamics in video sequences. In *Proc. of CVPR*, 1997.

[2] M. Fleck, D. Forsyth, and C. Bregler. Finding naked people. In *Proc. of ECCV*, Cambridge, England, April 1996.

[3] C. Guo and J. Luo. Non-purposive region grouping. In *Proc. IEEE Int. Conf. Image Process.*, 2002.

[4] S. Ioffe. *Managing Correspondence Search in Object Recognition.* PhD thesis, Berkeley, 2001.

[5] J. Luo, R. T. Gray, and H.-C. Lee. Towards physics-based segmentation of photographic color images. In *Proc. IEEE Int. Conf. Image Process.*, 1997.

[6] M. Oren, C. Papageorgiou, P. Sinha, E. Osuna, and T. Poggio. Pedestrian detection using wavelet templates. In *Proc. of CVPR*, San Juan, 1997.

[7] L. Zhao. *Dressed Human Modeling, Detection, and Parts Localization*. PhD thesis, Carnegie Mellon University, 2001.