

Probabilistic Framework for Segmenting People Under Occlusion

Ahmed M. Elgammal

Larry S. Davis

Computer Vision Laboratory
University of Maryland
College Park, MD, 20742 USA

Abstract

In this paper we address the problem of segmenting foreground regions corresponding to a group of people given models of their appearance that were initialized before occlusion. We present a general framework that uses maximum likelihood estimation to estimate the best arrangement for people in terms of 2D translation that yields a segmentation for the foreground region. Given the segmentation result we conduct occlusion reasoning to recover relative depth information and we show how to utilize this depth information in the same segmentation framework. We also present a more practical solution for the segmentation problem that is online to avoid searching an exponential space of hypothesis. The person model is based on segmenting the body into regions in order to spatially localize the color features corresponding to the way people are dressed. Modeling these regions involves modeling their appearance (color distributions) as well as their spatial distribution with respect to the body. We use a non-parametric approach based on kernel density estimation to represent the color distribution of each region and therefore we do not restrict the clothing to be of uniform color. Instead, it can be any mixture of colors and/or patterns. We also present a method to automatically initialize these models and learn them before the occlusion.

1 Introduction

It is desirable for visual surveillance systems to know what people are doing while interacting with each other. Visual surveillance systems are required to keep track of targets as they move through the scene even when they are occluded by or interacting with other people in the scene. It is highly undesirable to lose track of the targets when they are in a group. It is even more important to track the targets when they are interacting than when they are isolated.

The problem that we address in this paper is how to build a representation for people when they are isolated that enables their segmentation when they are interacting as a

group, as well as reasoning about the occlusion. This problem is important not only for visual surveillance, but also for other video analysis application such as video indexing, video archival and retrieval system.

The assumption we make about the scenario is that targets are visually isolated before occlusion so that we can initialize their models. Our approach is based on modeling the major color regions of the body such as head, torso, bottom part (legs) which corresponds basically to different pieces of clothing that a person is wearing. Modeling these regions involves modeling their appearance (color distributions) as well as their spatial distribution with respect to the body. We use a non-parametric approach based on kernel density estimation to represent the color distribution of each region and therefore we do not restrict clothing to be of uniform color. Instead, it can be any mixture of colors and/or patterns. Building these models is performed while targets are isolated by segmenting the body into blobs corresponding to their clothes; this segmentation is initialized based on training data.

Given a foreground region corresponding to a group of people we search for the arrangement that maximizes the likelihood of the appearance of this region given the models that we have built for the individuals. As a result, we obtain a segmentation of the region. The segmentation result is then used to determine the relative depth of each individual by evaluating different hypothesis about the people's arrangement. This allows us to construct a model for occlusion that can be used in the same probabilistic framework to segment foreground regions in subsequent frames.

The problem of tracking groups of people has been addressed recently in the literature. The Hydra [2] systems tracks people in groups by tracking their heads based on the silhouette of the foreground regions corresponding to the group. The Hydra system is able to count the number of people in the groups as long as their heads appear as part of the outer silhouette of the group; it would fail otherwise. The Hydra system was not intended to accurately segment the group into individuals nor does it recover depth information. McKenna *et al.* [6] segment groups of people based on

the individuals color distribution. They represent the color distribution of the whole person by a histogram and use this to segment the group. The color features are represented globally and are not spatially localized; therefore their approach loses the spatial information about the color distributions which is essential discriminant information. The notion of blobs has been used to model humans as a way to spatially localize the color information [10] in the context of tracking individuals. We will briefly mention related work that addressed problems such as representing color regions and occlusion reasoning throughout the paper.

The outline of the paper is as follow: Section 2.1 introduces the color model. Section 2.2 explains how to use these models to segment a foreground region corresponding to a group of people. Section 3 shows how we do occlusion reasoning, construct a model for that occlusion and utilize that model for the segmentation. Sections 4, 5 explain how we automatically initialize the target model.

2 Segmentation under Occlusion

2.1 Representation

People can be dressed in many different ways, but generally the way people are dressed leads to a set of major color regions aligned vertically (shirt, T-shirt, jacket etc., on the top and pants, shorts, skirts etc., on the bottom) for people in upright pose. Our approach is based on representing a person as a set of blobs representing the major parts of the body as the torso, bottom and head. Each blob is represented by its color distribution as well as its spatial location with respect to the whole body. Generally, a person in an upright pose is modeled as a set of vertically aligned blobs $M = \{A_i\}$ where a blob A_i models a major color region along the vertical axis of the person. We have two basic assumptions about the blob structure: First, each blob has the same color distribution everywhere inside the blob, i.e., the color of a pixel $h_A(c)$ within blob A is independent of the location of that pixel within that blob. This can be expressed as:

$$H_A(c | x, y) = h_A(c)$$

where H_A is the conditional probability that a pixel has a color c given its 2D image location x, y . This color-spatial independence assumption is applicable to the majority of clothing people wear and has been used previously in [10]. Notice that the restriction that each blob has a single color distribution does not imply that the blob is uniformly colored.

The second assumption is that the vertical location of a blob with respect to the person is independent of its horizontal location. This can be expressed as :

$$G_A(y | x) = g_A(y)$$

where G_A is the conditional density for the vertical location of blob A given the horizontal location.

From the previous assumptions it follows that a blob can be represented by three independent density functions:

- Color density function $h_A(c)$
- Vertical density function $g_A(y)$ representing the vertical location of the blob within the body.
- Horizontal density $f_A(x)$

where the spatial density functions $g_A(y), f_A(x)$ are defined relative to some origin. The joint distribution of pixel (x, y, c) (the probability of observing color c at location (x, y) given blob A) is

$$P_A(x, y, c) = f_A(x)g_A(y)h_A(c)$$

Furthermore, since our blobs are aligned vertically, we can assume that all the blobs share the same horizontal density function $f(x)$. Therefore, given a person model $M = \{A_i\}$ $i = 1 : n$ the probability of (x, y, c) is:

$$P(x, y, c | M) = \frac{f(x)}{C(y)} \sum_{i=1}^n g_{A_i}(y)h_{A_i}(c) \quad (1)$$

where C is a normalization factor $C(y) = \sum_i g_{A_i}(y)$.

A pixel $X = (x, y, c)$ can be classified to one of the n blobs using maximum likelihood classification assuming all blobs have the same prior probabilities

$$\begin{aligned} X \in A_k \text{ s.t. } k &= \arg_k \max P(X | A_k) \\ &= \arg_k \max g_{A_k}(y)h_{A_k}(c) \end{aligned} \quad (2)$$

since $P(x, y, c|A_k) \propto g_{A_k}(y)h_{A_k}(c)$ where $k = 1..n$.

This formalization gives us a way to segment foreground regions corresponding to a person into blobs as well as obtaining a probability estimate for a given pixel being part of that person. The following subsection will discuss how to use such estimates to segment foreground regions corresponding to groups of people. Section 5 discusses how to automatically initialize this blob model and how to obtain estimates for the density functions while tracking isolated people. One drawback of this representation is its inability to model highly articulated parts such as hands; but since our main objective is to segment people under occlusion, we are principally concerned with the mass of the body. Correctly locating the major blobs of the body will enforce constraints on the location of the hands which could then be used to locate and segment them.

2.2 Likelihood Maximization

For simplicity and without loss of generality we will focus on the the two people case. Given a person model

$M = \{A_i\}$ where $i = 1 : n$, the probability of observing color c at location x, y is:

$$P(x, y, c | M) = \frac{f(x)}{C(y)} \sum_{i=1}^n g_{A_i}(y) h_{A_i}(c)$$

where x, y and the spatial densities $g_{A_i}(y), f(x)$ are defined relative to an origin o . If the origin moves to x_o, y_o we can shift the previous probability to be:

$$P(x, y, c | M(x_o, y_o)) = \frac{f(x - x_o)}{C(y - y_o)} \sum_{i=1}^n g_{A_i}(y - y_o) h_{A_i}(c)$$

This defines the conditional density as a function of the model origin (x_o, y_o) , i.e., (x_o, y_o) is a parameter for the density and it is the only degrees of freedom allowed.

Given two people in occlusion with models $M_1(x_1, y_1)$ and $M_2(x_2, y_2)$, $h = (x_1, y_1, x_2, y_2)$ is a 4 dimensional hypothesis for the models' origins. We will call h an arrangement hypothesis. For a foreground region $X = (X_1, \dots, X_m)$ representing those two people, each foreground pixel $X_i = (x_i, y_i, c_i)$ can be classified to one of the two classes using maximum likelihood classification (assuming the same prior probability for each person). This defines a segmentation $\omega_h(X) = (\omega_h(X_1), \dots, \omega_h(X_m))$ that minimizes Bayes error where

$$\omega(X_i) = k \text{ s.t. } k = \arg_k \max P(X_i | M_k(x_k, y_k)) \quad k = 1, 2$$

Notice that the segmentation $\omega_h(X)$ is a function of the origin hypothesis h for the two models. i.e., each choice for the targets' origins defined a different segmentation of the foreground region. The best choice for the targets' origins is the one that maximizes the likelihood of the data over the entire foreground region. Therefore, the optimal choice for h can be defined in terms of a log-likelihood function

$$h_{opt} = \arg_h \max \sum_{i=1}^m \log P(X_i | M_k(h))$$

For each new frame at time t , searching for the optimal $(x_1, y_1, x_2, y_2)_t$ solves both the foreground segmentation as well as person tracking problems simultaneously. This formalization extends in a straightforward way to the case of N people in a group. In this case, we have N different classes and an arrangement hypothesis is a $2N$ dimensional vector $h = (x_1, y_1, \dots, x_N, y_N)$.

2.3 Origin Detection Solution

Finding the optimal hypothesis for N people is a search problem in $2N$ dimension space and exhaustive search for this solution would require $O(w^{2N})$, where w is 1-Dimensional window for each parameter (i.e., the diameter

of the search region in pixel). So, finding the optimal solution this way is exponential in the number of people in the group, which is non-practical. Instead, since we are tracking the targets throughout the occlusion and targets are not expected to move much between consecutive frames, we can develop a more practical solution based on direct detection of an approximate solution \hat{h}^t at frame t given the solution \hat{h}^{t-1} at frame $t-1$. Let us choose a model origin that is expected to be visible throughout the occlusion and can be detected in a robust way. For example, if we assume that the tops of the heads are visible throughout the occlusion, we can use them as origins for the spatial densities. Moreover the top of the head is a shape feature that can be detected robustly given our segmentation. Given the model origin location $\hat{h}^{t-1} = (x_i, y_i)^{t-1}$ at frame $t-1$, we can use this origin to classify each foreground pixel X at frame t using the maximum likelihood of $P(X | M(x_i, y_i)_{t-1})$. Since the targets are not expected to have significant translations between frames, we expect that the segmentation based on $(x_i, y_i)_{t-1}$ would be good in frame t except possibly at the boundaries. Using this segmentation we can detect new origin locations (top of the head) i.e., $(x_i, y_i)_t$. We can summarize this in the following steps:

1. $h_o^t \leftarrow \hat{h}^{t-1} = (x_1, y_1, \dots, x_N, y_N)^{t-1}$
2. Segmentation: Classify each foreground pixel X based on $P(X | M_k(x_k, y_k))$
3. Detection: Detect new origins (top of heads) $\rightarrow \hat{h}^t$

Conducting repetitive segmentation-detection might lead to a better solution in the sense of maximizing the likelihood of the data, i.e, we can write an iterative version of this algorithm where at each segmentation step the new solution is evaluated based on the likelihood function and another iteration is performed as long as we improve the likelihood. In practice, we found that one step is enough to reach a good segmentation as will be shown in section 6.

3 Modeling Occlusion

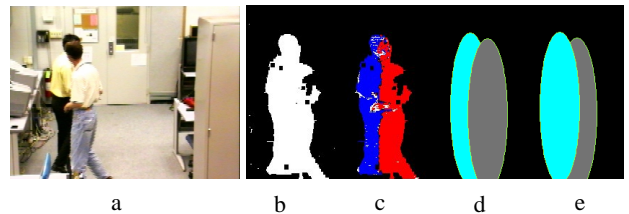


Figure 1. a- Original image. b- Foreground region. c-Segmentation result. d,e-Occlusion model hypotheses.

By occlusion modeling we mean assigning a relative depth to each person in the group based on the segmentation result. Several approaches have been suggested in the literature to solve this problem. In [3] a ground plane constraint was used to reason about occlusion between cars. The assumption that object motion is constrained to the ground plane is valid for people and cars but would fail if the contact point on the ground plane is not visible because of partial occlusion by other objects, or because contact points are out of the field of view (for example, see figure 1). McKenna *et al* [6] define the visibility index to be the ratio between the number of pixel visible of each person during occlusion to the expected number of pixels for that person when isolated. They use this visibility index to measure the depth (higher visibility index indicates that the person is in front). While this can be used to identify the person in front, we can easily construct examples to show that the visibility index does not correspond to depth for more than two people. The solution we present here does not use the ground plane constraint and generalizes to the case of N people in a group.

3.1 Occlusion Reasoning

Given a hypothesis h about the 3D arrangement of people along with their projected locations in the image plane and a model of their shape, we can construct an occlusion model $O_h(x)$ that maps each pixel x to one of the tracked targets or the scene background. Let us consider the case of two targets as shown in figure 1. The foreground region is segmented as was shown in section 2.2, which yields a labeling $\omega(x)$ for each pixel (figure 1-c) as well as the best location for the model origins. There are two possible hypotheses about the depth arrangement of these two people and the corresponding occlusion models are shown in parts d and e of the figure assuming an ellipse as a shape model for the targets. We can evaluate these two hypotheses (or generally N hypotheses) by minimizing the error in the labeling between $O_h(x)$ and $\omega(x)$ over the foreground pixels. i.e., $error(h) = \sum_{x \in FG} (1 - \delta(O_h(x), \omega(x)))$ for all foreground pixels¹. We use an ellipse with major and minor axes set to the expected height and width of each person estimated before the occlusion. Figures 5,6 show some examples of the constructed occlusion model for some occlusion situations.

3.2 Utilizing depth information

Consider the situation where two targets are being tracked through an occlusion and we are able to determine the depth index for each target, and therefore we have an

¹In the two person case an efficient implementation for this error formula can be achieved by considering only the intersection region and finding the target which appears most in this region; that corresponds to the one in front

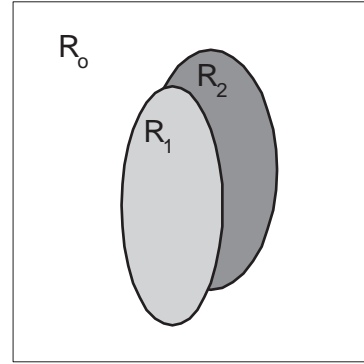


Figure 2. Occlusion model example.

occlusion model similar to the one in figure 2 where R_1, R_2 represents targets 1,2 respectively and R_0 represents the scene background. Clearly, pixels in region R_1 are more likely to be part of the first target. In the following we discuss how to utilize this information in the segmentation process.

Generally, if we have m targets, we would have $m + 1$ regions (layers) similar to those in figure 2 where region R_i represents the visible parts of target i . Let the layer probability $\pi(x) = (\pi^0(x), \dots, \pi^m(x))$ be an $m + 1$ probability vector where $\pi^i(x)$ is the probability that pixel x belongs to layer i and $\sum_{i=0}^m \pi_i(x) = 1$ In other words, $\pi^i(x)$ is the probability that the ray from pixel x through the optical center will hit target i first. Any target arrangement hypothesis h would define different layer probabilities so we will use the notation $\pi_h(x)$ to denote the layer probability defined by such hypothesis. The likelihood probability of a pixel given target i would be

$$P_h(x | target_i) = P(x | M_i(h)) \cdot \pi_h^i(x)$$

Therefore we can use the same framework as in section 2.2 to find the hypothesis that maximizes the likelihood of the foreground region. This way we extend the notion of arrangement hypothesis to include both target 2D location and relative depth.

The question is how we can obtain an estimate for π_h . Here we present a heuristic method to obtain such estimates for each new frame t based on the previous frame, $t - 1$. Let $\omega^{t-1}(x)$ be the segmentation result and let $O^{t-1}(x)$ be the occlusion model for frame $t - 1$. Let $\lambda^{t-1}(i, j)$ be the probability of observing object i in region R_j where $\sum_i \lambda^{t-1}(i, j) = 1$. We can calculate these probabilities from the segmentation result and the occlusion model of frame $t - 1$ as

$$\lambda^{t-1}(i, j) = \frac{\text{N. of pixels labeled } i \text{ in } R_j}{\text{N. of pixels in } R_j}$$

Given this estimate we can assign the layer probabilities for

hypothesis h at frame t as

$$\pi_h^i(x) = (1 - \alpha) \lambda^{t-1}(i, O_h(x)) + \frac{\alpha}{m+1}$$

where $O_h(x)$ is the occlusion model defined by hypothesis h and α is a parameter representing the uncertainty in the process resulting from using inaccurate shape models. Note that if α is set to 1 all the layer will be equiprobable.

4 Blob Modeling

The framework presented in section 2 for segmentation is applicable to any method to estimate the color density function and the spatial density function(s) for each blob. A variety of parametric and non-parametric statistical techniques have been used to model the color and the spatial properties of colored regions. In [10] the color properties of a blob were modeled using a single Gaussian in the three dimension YUV space. The spatial properties of a blob were modeled using pixel support maps. Fitting a mixture of Gaussian using the EM algorithm provides a way to model blobs with a mixture of colors. This technique was used in [7, 8] for color based tracking of a single blob and was applied to tracking faces. Mixture of Gaussian techniques face the problem of choosing the right number of Gaussian for the model. Non-parametric techniques using histograms have also been used in [6]. In this work they used 3-dimensional adaptive histograms in RGB space to model the color of the whole person and therefore no color-spatial localization was used in their model. Color histograms have also been used in [5] for tracking hands. The major drawback with color histogram is the lack of convergence to the right density function if the data set is small.

4.1 Blob Color Model

Our approach is to model the color density of a blob using non-parametric kernel density estimation. Given a sample $S = \{x_i\}$ where $i = 1 \dots N$ and x_i is a d -dimensional vector, kernel density estimation [9] can be used to estimate the probability that a sample $y = (y_1, \dots, y_d)$ is from the same distribution as S

$$\hat{P}(y) = \frac{1}{N\sigma_1 \dots \sigma_d} \sum_{i=1}^N \prod_{j=1}^d K\left(\frac{y_j - x_{ij}}{\sigma_j}\right) \quad (3)$$

where the same kernel function is used in each dimension with different bandwidth σ_j .

We represent the color of each pixel as a 3-dimensional vector $X = (r, g, s)$ where $r = \frac{R}{R+G+B}$, $g = \frac{G}{R+G+B}$ are two chromaticity variables and $s = (R+G+B)/3$ is a lightness variable. Given a sample of pixels $S_A = \{X_i = (r_i, g_i, s_i)\}$ from blob A , an estimate $\hat{h}_A(\cdot)$ for the color density $h_A(\cdot)$ can be calculated as

$$\hat{h}_A(r, g, s) = \frac{1}{N} \sum_{i=1}^N K_{\sigma_r}(r - r_i) K_{\sigma_g}(g - g_i) K_{\sigma_s}(s - s_i)$$

where $K_\sigma(t) = 1/\sigma K(t/\sigma)$. We use Gaussian kernels, i.e., $K_{\text{sigma}}(t) = \frac{1}{\sqrt{2\pi}\sigma} e^{-1/2(t/\sigma)^2}$ with different bandwidth in each dimension.

Theoretically, kernel density estimators can converge to any density shape with enough samples [9]. Unlike histograms, even with small number of samples, kernel density estimation leads to a smooth continuous density estimate. If the underlying distribution is a mixture of Gaussians, kernel density estimation converges to the right density with a small number of samples. Unlike parametric fitting of a mixture of Gaussians, kernel density estimation is a more general approach that does not require the selection of the right number of Gaussian to fit. One other important advantage of using kernel density estimation is that the adaptation of the model is trivial by adding new samples. The major drawback of non-parametric kernel density estimator is its computational cost, which becomes less of a problem with the available computational power and efficient computational methods that have been developed recently [4]

The separation of chromaticity from lightness in the rgs space allows the use of a much wider kernel with the s variable to cope with the variability in this variable due to shading effects as well as small changes in lighting condition due to target motion. On the other hand, the chromaticity variables r, g are invariant to shading effects and therefore much narrower kernels can be used in these dimensions, which enables more powerful chromaticity discrimination. We cannot discard the lightness information since it is essential for discriminating non colored objects (objects on the gray line)

The estimation of appropriate bandwidths is done offline by considering batches of single colored regions taken from images of people's clothing and estimating the variance in each color dimension. Theoretically, for the Gaussian case the bandwidth can be estimated as $h \approx 1.06\hat{\sigma}n^{-1/5}$ [9] where $\hat{\sigma}$ is the estimated standard deviation and n is the sample size.

4.2 Blob Spatial Model

Estimates for the blob vertical density $g_A(y)$ and horizontal density $f(x)$ are learned while tracking the target before occlusion. Each blob vertical density $g_A(y)$ is represented non-parametrically by a histogram as will be shown in section 5. The horizontal density is assumed to be a Gaussian whose parameters are determined by fitting a Gaussian to the vertical projection of target pixels centered at the median.

5 Blob Extraction

The blob extraction process is performed when people are isolated before occlusion. Generally, the way people dress leads to three or four color regions along the vertical axis for upright pose. We consider here the case where the

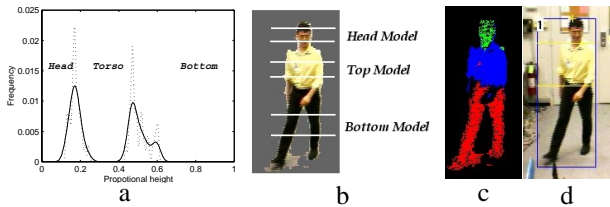


Figure 3. a- Blob separator histogram from training data. b- Confidence bands. c - Blob segmentation. d- Detected blob separators

person model consists of three blobs corresponding to a person’s head, torso and bottom. We denote the horizontal line that roughly separates two consecutive blobs as “blob separator”. A set of training data² is used to learn the location of blob separators (head-torso, torso-bottom) with respect to the body for a set of people in upright pose. Figure 3-a shows a histogram of the locations of head-torso (left peak) and torso-bottom (right peak) in the training data. Based on these separator location estimates, we can determine regions proportional to the height (confidence bands) where we are confident that they belong to head, torso and bottom.

The initialization is done automatically by taking three samples $S = \{S_H, S_T, S_B\}$ of pixels from the three confidence bands corresponding to head, torso and bottom. Given a set of samples $S = \{S_{A_i}\}$ corresponding to each blob and initial estimates for the position of each blob y_{A_i} , each pixel is classified to one of the three blobs based on maximum likelihood classification (equation 2) where $g_{A_i}(y) = N(y_{A_i}, \sigma_{A_i})$ and $h_{A_i}(c)$ is estimated using kernel density estimator as described in section 4.1 .

The actual blob separators are then detected by finding the horizontal line that minimizes the error in classification. Let A and B be two blobs where A is above B . Let $L(X_i) : X_i \rightarrow \{A, B\}$ be the classification result of pixel X_i . A horizontal separator y_{AB} between the two blobs is defined by

$$y_{AB} = \arg_y \min \sum_i 1 - \delta(L(X_i), M_y(X_i))$$

where $M_y(X_i) = \begin{cases} A & X_i \text{ above } y \\ B & X_i \text{ below } y \end{cases}$ and $\delta(a, b) = \begin{cases} 1 & a = b \\ 0 & \text{otherwise} \end{cases}$

Given the detected blob separators, the color model is recaptured by sampling pixels from each blob. Blob segmentation is performed and blob separators are detected at each new frame as long as the target is isolated and tracked.

²The training data consists of 90 samples of different people from both genders dressed in top-bottom manner in different orientations in upright pose

A histogram of vertical location of detected blob pixels $H_{A_i}^t(y)$ is used to update the blob vertical density $g_{A_i}(y)$ at each new frame t for each blob A_i by

$$g_{A_i}^t(y) = (1 - \alpha)g_{A_i}^{t-1}(y) + \alpha H_{A_i}^t(y)$$

where each vertical location y is aligned and scaled to $\hat{y} = \frac{y - y_{top_bottom}}{|y_{top_bottom} - y_{head_top}|} \cdot h$ using the detected blob separators $y_{top_bottom}, y_{head_top}$. Blob separators are robust features that are invariant to partial occlusion as long as they are visible.

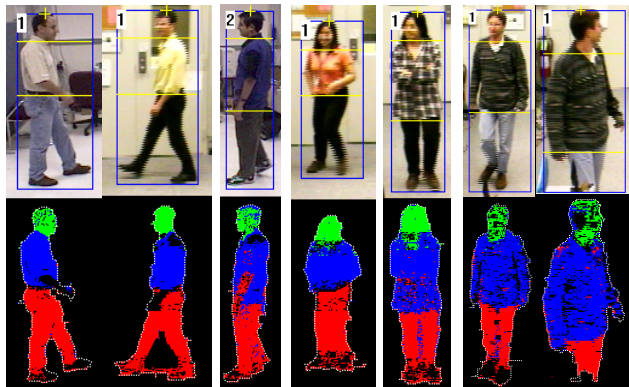


Figure 4. Example results for blob segmentation

Figure 3-b shows initial bands used for initialization where the segmentation result is shown in 3-c and the detected separators are shown in 3-d. Figure 4 illustrates some blob segmentation examples for various people. Notice that the segmentation and separator detection is robust even under partial occlusion of the target as in the rightmost result. Also in some of these examples the clothes are not of a uniform color.

6 Experimental Results

The input to the algorithm are foreground regions corresponding to the moving objects which were extracted from the scene background. Background subtraction is used to extract these regions as a preprocessing step that we do not discuss in this paper, details can be found in [1].

Figure 5 shows some results for segmenting two people in different occlusion situations. The foreground segmentation between the two people is shown as well as blob segmentation. Pixels with low likelihood probabilities are not labeled. In most of the cases, hands and feet are not labeled or are miss-classified because they are not part of the blob representation. The constructed occlusion model for each case is also shown. Notice that in the third and fourth examples, the two people are dressed in similarly colored pants. Therefore, only the torso blobs are discriminating

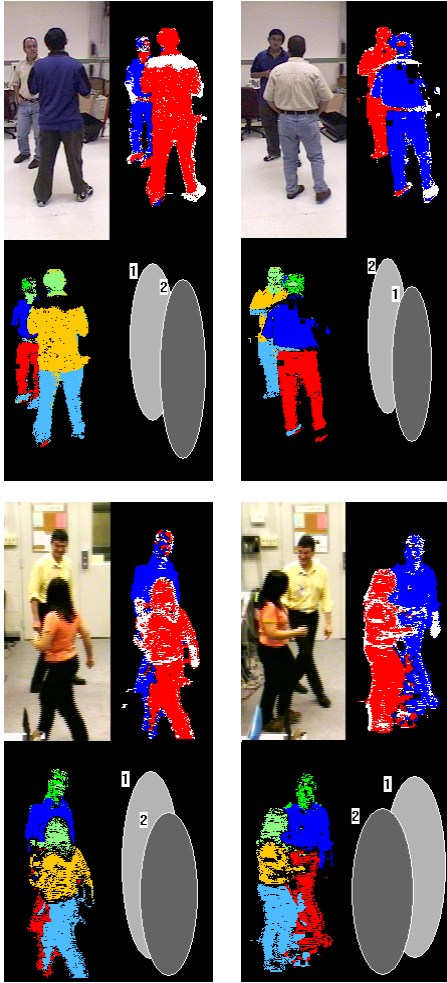


Figure 5. Example results: Top left: Original image. Top right: people segmentation. Bottom left: blob segmentation. Bottom right: Constructed occlusion model

in color. This was sufficient to locate each person’s spatial model parameters and therefore similarly colored blobs (head and bottom) were segmented correctly based mainly on their spatial densities. Still, some miss-classification can be noticed around the boundaries between the two pants which is very hard even for human to segment accurately³.

Figure 6 illustrates several frames from a sequence for two targets being tracked throughout occlusion. The blob segmentation results are shown as well as the constructed occlusion model. This result is obtained using the segmentation-detection solution (section 2.3) with the top of the head used as a reference point for the spatial densities.

³Full video clips showing these results and others can be downloaded from <ftp://www.umiacs.umd.edu/users/pub/elgammal/video/occlusion>

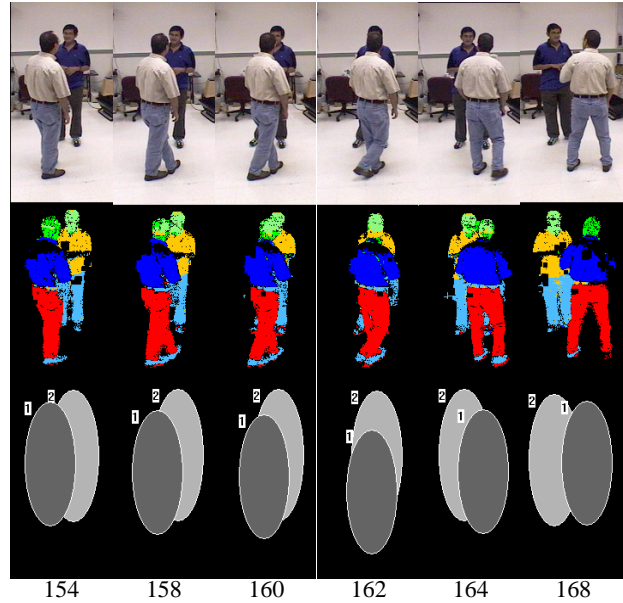


Figure 6. Example results: Top: Original image. Middle: Blob segmentation. Bottom: Occlusion model

In order to evaluate the algorithms we compared the segmentation results with ground truth data that we obtained by manual segmentation at certain key frames. We measure the cross classification error which is defined for each class as number of incorrectly classified pixel / number of foreground pixels for that class. The first result, shown in figure 7, is a comparison between the two methods described in sections 2.2 and 2.3, i.e., a search for the optimal hypothesis that maximize the likelihood versus origin detection solution. For the first method an exhaustive search for the optimal target model parameter hypothesis was performed at each new frame with a window of size 9 pixels in each dimension around the previous frame solution. For the second method the solution was obtained at each new frame by detecting the top of the head based on segmentation using previous frame solution as a hypothesis for the spatial model parameters. The error rates for each target as well as the overall miss-classification rate are shown in figure 7 which also shows (top plot) the ground truth visibility ratio of the occluded person as a measure for the occlusion. As can be seen, the origin detection solution gives similar results to the search solution in most of the evaluation frames except frames 162,164. This is because the heads of the two people were against each other and that caused significant confusion between them. Therefore the tops of the heads were not detected accurately. The segmentation result using the origin detection solution is shown in figure 6.

Figure 8 shows the effect of utilizing the recovered depth

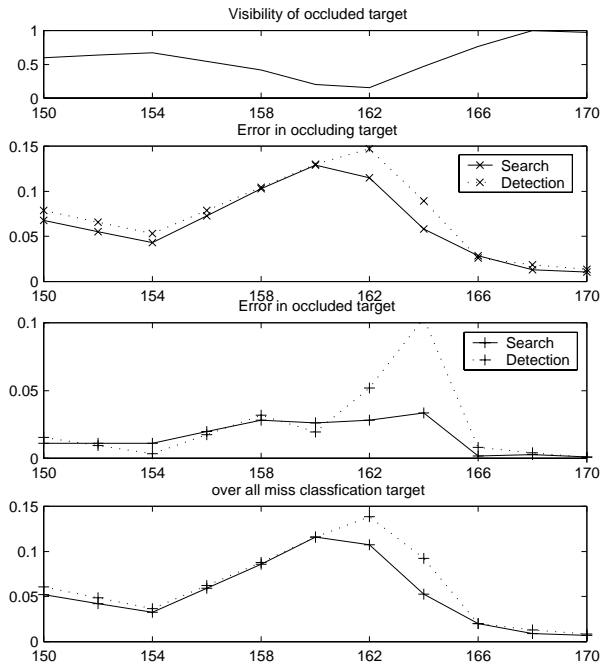


Figure 7. Evaluation: Exhaustive search vs. direct solution. From top to bottom: Visibility ratio of occluded target as a measure for occlusion, Error in occluding target, Error in occluded target, Overall cross classification.

information (layer probability) in the segmentation as was described in section 3.2. We compared the error rates that we obtained by utilizing the layer probability in segmentation to the error rates obtained using color-spatial information only (section 2.2). In both cases we search for the optimal target origin hypothesis. The results show that a slight improvement can be achieved by utilizing layer probabilities with a simple shape model. Further improvement is expected by utilizing a more accurate shape model.

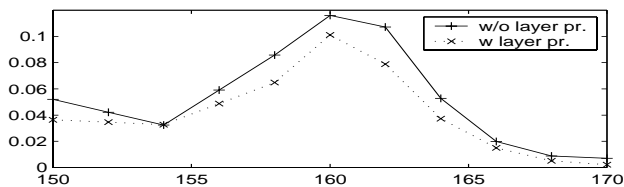


Figure 8. Evaluation: Error rates for segmentation without layer probabilities vs. segmentation with layer probabilities

7 Conclusion

We can summarize our contribution as follows: We introduced a representation of people that spatially localizes color properties in a meaningful way. We presented a general probabilistic framework that uses maximum likelihood estimation to estimate the best arrangement for people in a group in order to segment the foreground regions corresponding to this group. The framework can be used with any density estimation method for color density and spatial densities for each blob. We presented a method to reason about occlusion so we can construct a model of that occlusion and showed how we can utilize such model in the same segmentation framework.

Currently, the automatic initialization of a person's model is restricted to people dressed in a top-bottom manner which yields three color blobs. Immediate future extension is to be able to do automatic initialization in a general way based on pre-trained cloth model specially that our framework is not restricted to a certain cloth model.

Future work includes also segmenting groups of peoples without pre-captured models of their appearance, i.e., build these models simultaneously while solving the occlusion segmentation program.

References

- [1] A. Elgammal, D. Harwood, and L. S. Davis. Nonparametric background model for background subtraction. In *6th European Conference of Computer Vision*, 2000.
- [2] I. Haritaoglu, D. Harwood, and L. S. Davis. Hydra: Multiple people detection and tracking using silhouettes. In *IEEE International Workshop on Visual Surveillance*, 1999.
- [3] D. Koller, J. Weber, and J. Malik. Robust multiple car tracking with occlusion reasoning. In *European Conference of Computer Vision*, 1994.
- [4] C. Lambert, S. Harrington, C. Harvey, and A. Glodjo. Efficient on-line nonparametric kernel density estimation. *Algorithmica*, (25):37–57, 1999.
- [5] J. Martin, V. Devin, and J. Crowley. Active hand tracking. In *IEEE International Conference on Automatic Face and Gesture Recognition*, 1998.
- [6] S. J. McKenna, S. Jabri, Z. Duric, and A. Rosenfeld. Tracking groups of people. *Computer Vision and Image Understanding*, (80):42–56, 2000.
- [7] Y. Raja, S. J. Mckenna, and S. Gong. Colour model selection and adaptation in dynamic scenes. In *5th European Conference of Computer Vision*, 1998.
- [8] Y. Raja, S. J. Mckenna, and S. Gong. Tracking colour objects using adaptive mixture models. *Image Vision Computing*, (17):225–231, 1999.
- [9] D. W. Scott. *Multivariate Density Estimation*. Wiley-Interscience, 1992.
- [10] C. R. Wern, A. Azarbayejani, T. Darrell, and A. P. Pentland. Pfunder: Real-time tracking of human body. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 1997.