An Interactive Approach to Pose-Assisted and Appearance-based Segmentation of Humans

Zhe Lin, Larry S. Davis, David Doermann, and Daniel DeMenthon
Institute for Advanced Computer Studies, University of Maryland, College Park, MD 20742
{zhelin,lsd,doermann,daniel}@umiacs.umd.edu

Abstract

An interactive human segmentation approach is described. Given regions of interest provided by users, the approach iteratively estimates segmentation via a generalized EM algorithm. Specifically, it encodes both spatial and color information in a nonparametric kernel density estimator, and incorporates local MRF constraints and global pose inferences to propagate beliefs over image space iteratively to determine a coherent segmentation. This ensures the segmented humans resemble the shapes of human poses. Additionally, a layered occlusion model and a probabilistic occlusion reasoning method are proposed to handle segmentation of multiple humans in occlusion. The approach is tested on a wide variety of images containing single or multiple occluded humans, and the segmentation performance is evaluated quantitatively.

1. Introduction

Object segmentation is a long-studied but still difficult problem in computer vision. When the objects of interest and the background have similar color or texture, when the objects are in a cluttered background, or when objects appear under occlusion, segmentation become especially challenging. Recently, the interactive segmentation scheme has become very popular due to its efficiency in handling these difficult cases.

Pairwise potential-based approaches perform figure-ground discrimination by clustering features based on pairwise costs, *e.g*. Normalized Cut [15]. In contrast, object-centered clustering approaches group features with learned parametric or nonparametric densities; typical examples include the *k*-means clustering, and the EM-based clustering with mixtures of Gaussians [6]. EM-based approaches are sensitive to initialization and require appropriate selection of the number of mixture components. It is well known that finding a good initialization and choosing a generally reasonable number of mixtures for the traditional EM algorithm remain difficult problems. In [19], the KDE-EM approach is introduced by applying nonparametric kernel density estimation method in EM-based color clustering. Graph-cut approaches combine the pairwise potential-based scheme with object-centered appearance representation in a unified energy minimization paradigm, *e.g*. Interactive Graph-Cuts [4], and its generalized version, GrabCut [13].

Object segmentation without any prior knowledge is well-known to be an ill-posed problem. Recently a few approaches have concentrated on enforcing global shape priors, top-down reasoning or other higher level knowledge to make the segmentation problem well posed. Object category-specific MRF [11] or pose-specific MRF [5] combines local contrast-dependent MRF with a layered pictorial structure model in [11] or a stickman model in [5] to provide strong global priors. Hence, the resulting segmentations resemble objects of interest. In [16], bottom-up cues are combined with global top-down knowledge for object class learning with unsupervised segmentation. In [12], an appearance learning-based method is proposed for articulated body segmentation and pose estimation; however it focuses on pose estimation and does not compute object segmentation explicitly. In [3], top-down shape cues are used to merge bottom-up over-segmentation to generate an object-like segmentation. In [18], the KDE-EM approach is combined with a shape template-based detection method for object segmentation. Recently, in [17], a layout-consistent random field is employed to provide a preliminary solution to segmentation in the presence of occlusion.

We propose an alternative, more efficient approach to object segmentation capable of handling inter-occlusion between objects.[1] We incorporate local contrast-dependent MRF constraints and global shape priors iteratively into the KDE-EM framework [19] to estimate segmentations and poses simultaneously. There are four important contributions in this paper. First, we represent kernel densities of foreground and background in a joint spatial and color space and update assignment probabilities *recursively* instead of

[1] In this paper, we focus on human segmentation but the approach can be applied to other object categories.

using the direct update scheme in KDE-EM; this modification of feature space and update equations results in faster convergence and better segmentation accuracy. Second, we incorporate contrast-dependent MRF constraints into the KDE-EM scheme to regularize and smooth the segmentation within object and background regions. Third, we build and train a human pose model and perform pose inferences in the iterative clustering stages to enforce global shape priors throughout the segmentation process. This encourages the segmentation of human-like shapes and allows us to optimize segmentations and poses simultaneously. Fourth, and most importantly, we generalize the approach to a multiple occluded object segmentation by explicitly modeling and reasoning about occlusion.

2. Modified KDE-EM Approach

KDE-EM uses nonparametric kernel density estimation [14] for representing feature distributions of foreground and background and performs iterative segmentation using EM. The log-likelihood objective function is similar to the one in the traditional EM-based segmentation, i.e. summation of log likelihoods of all pixels in the image, except that the likelihoods (assignment probabilities) are calculated from kernel densities.

Given a set of sample pixels $\{\mathbf{x}_i, i = 1, 2...N\}$ (with a distribution $\mathcal{P}$), each represented by a d-dimensional feature vector as $\mathbf{x}_i = (x_{i1}, x_{i2}..., x_{id})^t$, we can estimate the probability $\hat{P}(\mathbf{y})$ of a new pixel $\mathbf{y}$ with feature vector $\mathbf{y} = (y_1, y_2, ..., y_d)^t$ belong to the same distribution $\mathcal{P}$ as

$$\hat{P}(\mathbf{y} \in \mathcal{P}) = \frac{1}{N\sigma_1...\sigma_d} \sum_{i=1}^{N} \prod_{j=1}^{d} k(\frac{y_j - x_{ij}}{\sigma_j}), \quad (1)$$

where the same kernel function $k(\cdot)$ is used in each dimension (or channel) with different bandwidth σ_j. It is well known that a kernel density estimator can converge to any complex-shaped density with sufficient samples. Also due to its nonparametric property, it is a natural choice for representing the complex color distributions that arise in real images [7].

For enhancing the compactness and efficiency of the segmentation, we extend the color feature space in KDE-EM to incorporate spatial information. This joint spatial-color feature space has been previously explored for feature space clustering approaches such as [7, 9]. Each pixel is represented by a feature vector $\mathbf{x} = (X^t, C^t)^t$ in a 5D space, $\mathbb{R}^5$, with 2D spatial coordinates $X = (x_1, x_2)^t$ and 3D normalized rgs color[2] coordinates $C = (r, g, s)^t$. The separation of chromaticity from brightness in the rgs space allows the use of a much wider kernel with the s variable to cope with the variability in brightness due to shading effects. On the other hand, the chromaticity variables r and g are invariant to shading effects and therefore much narrower kernels can be used in these dimensions, which enables more powerful chromaticity discrimination [19]. In Equation 1, we assume independence between channels and use a Gaussian kernel $k(t) = 1/\sqrt{(2\pi)}exp\{-t^2\}$ for each channel. The kernel bandwidths are estimated from the original image as in [14, 19].

[2] $r = R/(R+G+B), g = G/(R+G+B), s = (R+G+B)/3$

Algorithm 1 Modified KDE-EM

Given a set of sample pixels $\{\mathbf{x}_i, i = 1, 2...N\}$ from the image, we iteratively estimate the assignment probabilities $F^t(\mathbf{y})$ and $B^t(\mathbf{y})$ $(t = 0, 1, 2...)$ of a pixel $\mathbf{y}$ belonging to the foreground $\mathcal{F}$ and background $\mathcal{B}$ as follows:

Initialization : Assign initial probabilities to pixels based on a 2D anisotropic Gaussian distribution. The parameters of the distribution are determined by the expected location and sizes (which are assigned via user interaction) of the foreground object.

$$F^0(\mathbf{y}) = e^{-1/2(Y-Y_0)^t V^{-1}(Y-Y_0)}, \quad (2)$$

$$B^0(\mathbf{y}) = 1 - F^0(\mathbf{y}), \quad (3)$$

where Y denotes the spatial coordinates of $\mathbf{y}$, Y_0 denotes expected object center coordinates, and V denotes a 2×2 (diagonal) covariance matrix. The diagonal elements of V are set proportional to the expected sizes of the object.

M − Step : $(Random\ Pixel\ Sampling)$ Randomly sample a set of pixels from the image to estimate foreground and background appearances represented by weighted kernel densities. For computational efficiency, we sample $\eta = 5\%$ of the pixels from the image for density estimation.

E − Step : $(Soft\ Probability\ Update)$

$$F^t(\mathbf{y}) = cF^{t-1}(\mathbf{y}) \sum_{i=1}^{N} F^{t-1}(\mathbf{x}_i) \prod_{j=1}^{d} k(\frac{y_j - x_{ij}}{\sigma_j}), \quad (4)$$

$$B^t(\mathbf{y}) = cB^{t-1}(\mathbf{y}) \sum_{i=1}^{N} B^{t-1}(\mathbf{x}_i) \prod_{j=1}^{d} k(\frac{y_j - x_{ij}}{\sigma_j}), \quad (5)$$

where N is the number of samples and c is a normalizing factor such that $F^t(\mathbf{y}) + B^t(\mathbf{y}) = 1$.

Segmentation : The iteration is terminated when $\frac{\sum_{\mathbf{y}}\{|F^t(\mathbf{y}) - F^{t-1}(\mathbf{y})|\}}{n} < \epsilon$, where n is total number of pixels in the image. $F(\mathbf{y})$ and $B(\mathbf{y})$ denote the final converged assignment probabilities. The segmentation is finally estimated as: $\mathbf{y} \in \mathcal{F}$ if $F(\mathbf{y}) > B(\mathbf{y})$, $\mathbf{y} \in \mathcal{B}$ otherwise.

KDE-EM employs a soft-labelling procedure and weighted kernel density estimation to update the assignment probabilities. For adapting the nonparametric kernel

density estimation to the EM algorithm, a sampling step is substituted for the M-step in EM. In each iteration, samples are independently drawn from a *uniform distribution* and weighted by the assignment probabilities estimated from the previous iteration. The foreground/background assignment probabilities $F^t(\mathbf{y})$ and $B^t(\mathbf{y})$ are updated directly by weighted kernel densities. We modify this by updating $F^t(\mathbf{y})$ and $B^t(\mathbf{y})$ *recursively* on the previous assignment probabilities $F^{t-1}(\mathbf{y})$, $B^{t-1}(\mathbf{y})$ with weighted kernel densities (Equations 4 and 5). This modification results in faster convergence and better segmentation accuracy. An example of the modified KDE-EM approach is shown in Figure 5.

3. Pose-Assisted Segmentation

KDE-EM treats individual pixels separately, hence, the resulting segmentation usually has holes or isolated small regions. In order to obtain a coherent and object-like segmentation, we use higher-order dependencies between pixels. The higher-order dependencies can be exploited in the form of local and global MRFs. Instead of incorporating these priors in the energy function [4, 2, 13, 11, 5], we apply them iteratively and recursively in a single process to force the segmentation result to be a human-like shape. This avoids the need for an extra optimization step such as graph-cut and achieves simultaneous segmentation and pose estimation efficiently. Also, our approach maintains soft labelling throughout the optimization process, while graph-cut is a discrete (labelling) optimization scheme.

3.1. Incorporating Local MRF Constraints

Let $\Psi^t_{\mathcal{F}}$ and $\Psi^t_{\mathcal{B}}$ represent the probabilities of a pixel $\mathbf{y}$ being labelled as the foreground and background according to local contrast-dependent MRF constraints which are defined as:

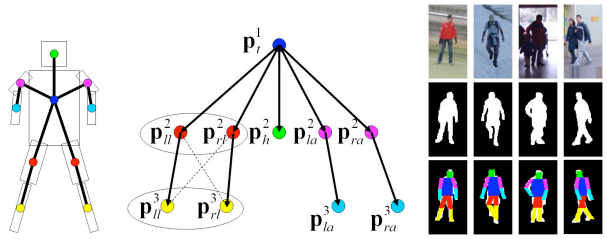
$$\Psi^t_{\mathcal{F}}(\mathbf{y}) = \sum_{\mathbf{z}\in\mathcal{N}_\mathbf{y}} \phi(\mathbf{I}|\mathbf{y},\mathbf{z})F^{t-1}(\mathbf{z}), \tag{6}$$

$$\Psi^t_{\mathcal{B}}(\mathbf{y}) = \sum_{\mathbf{z}\in\mathcal{N}_\mathbf{y}} \phi(\mathbf{I}|\mathbf{y},\mathbf{z})B^{t-1}(\mathbf{z}), \tag{7}$$

where $\mathbf{I}$ denotes the original image, $\mathcal{N}_\mathbf{y}$ denotes the neighborhood (8-neighborhoods) of pixel $\mathbf{y}$, and $\phi(\mathbf{I}|\mathbf{y},\mathbf{z})$ represents the contrast-dependent MRF induced likelihood for pixel $\mathbf{y}$. The likelihood $\phi(\mathbf{I}|\mathbf{y},\mathbf{z})$ is defined as:

$$\phi(\mathbf{I}|\mathbf{y},\mathbf{z}) = \frac{1}{dist(\mathbf{y},\mathbf{z})} e^{-\frac{1}{2}\left((\frac{r_\mathbf{z}-r_\mathbf{y}}{\sigma_r})^2+(\frac{g_\mathbf{z}-g_\mathbf{y}}{\sigma_g})^2+(\frac{s_\mathbf{z}-s_\mathbf{y}}{\sigma_s})^2\right)}. \tag{8}$$

To incorporate the local contrast-dependent MRF constraints into our iterative segmentation scheme, the recursive assignment probability update step (Equations 4 and 5) is extended by the local MRF terms (Equations 9 and 10). This can be explained as follows: the current foreground/background assignment probabilities are updated recursively by combined evidence from the spatial neighborhood and current foregournd/background appearance estimates (weighted kernel densities); in other words, the local MRF terms are incorporated to *smooth* the pixel-wise soft labelling at each iteration. We refer to this modified approach as CDMRF-KDE-EM. An example of the CDMRF-KDE-EM approach is shown in Figure 5.



(a) PS pose model (b) Tree-like structure (c) PS pose model fitting

Figure 1. An illustration of the pose model and training examples. (a) 10-parts PS model, (b) Simplified tree-like struture, (c) Examples of the training images, hand-segmented silhouettes, and PS pose model fitting results.

3.2. Enforcing Global Shape Priors by Poses

We next describe how to incorporate prior shape information into the segmentation process. We build a Pictorial Structure (PS) pose model similar to the model in [8] and enforce global shape priors based on adaptive pose inference on soft segmentations at each iteration.

3.2.1 PS Pose Model

We chose 808 images from the INRIA person dataset [1] as training images (some of them are shown in Figure 1(c)). Human poses are modeled as a 10-part pictorial structure (Figure 1(a)) of which each part is represented as a horizontal parallelogram with five degrees of freedom (position $\mathbf{p}$, orientation α, sizes $\mathbf{s}$). Hence, the model (represented by parameters θ) has a total of $5 \times 10 = 50$ degrees of freedom. For simplicity, we assume independence between head, arms and legs and assume the pair of arms are also independent (the pair of legs are still correlated). This enables us to simplify the model to a tree-like structure (Figure 1(b)) on which the root node is chosen as the torso.

The PS model has many degrees of freedom and the parameter space is huge, while possible human poses form a low-dimensional manifold in this space. Hence, for efficiently searching the parameter space, we train the pose model and estimate its joint parameter distribution $l(\theta)$ from the set of best matching poses which are estimated using

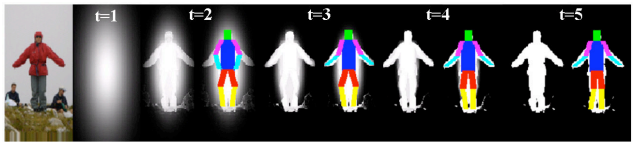


Figure 2. The iterative process of pose-assisted segmentation. Each frame represents the current soft segmentation overlaid with MAP fitted pose.

MLE by fitting the PS pose model to the binary silhouette images (obtained by manual segmentation of the training images) individually (Figure 1(c)). In our implementation, based on the above independence assumption, the joint distribution is marginalized as a set of individual joint distributions for different parts (head, torso, arms and legs). Hence, as a result of training, $l(\theta)$ is represented as a set of probability mass functions on low dimensional parameter spaces.

3.2.2 Training the Pose Model from Silhouettes

The degree of fitting $\rho(\theta|S)$ is defined as the similarity of the silhouette image S and the binary model coverage image $M(\theta)$, *i.e.* $\rho(\theta|S) = 1 - \frac{\sum_{\mathbf{x}} \|S(\mathbf{x}) - M(\theta,\mathbf{x})\|}{n}$, where n is the total number of pixels in the image. Then, the problem of model fitting can be formulated as a maximum likelihood estimation: $\theta_i^* = \arg\max_{\theta\in\mathbf{\Theta}} \rho(\theta|S_i)$, where $\mathbf{\Theta}$ denotes the set of all possible model parameters, and θ_i^* is the maximum likelihood estimate for the binary silhouette image $S_i, i \in \{1, 2, ..., N_t\}$ (N_t is the number of training images).

In training, we assume a uniform prior over $\mathbf{\Theta}$. According to the model in Figure 1(b), there are only loops between the pair of legs in the simplified tree-like graph structure. Optimization for matching the PS model to images is performed by belief propagation similar to [8] which is known to achieve globally optimal solutions for tree-structured acyclic graphs. In our approach, parameters for pair of legs are jointly optimized for handling the cases of occlusion between legs. Finally, the configuration corresponding to the maximum overall fitting score is returned as the estimate θ^*. Figure 1(c) shows some examples of PS model fitting results.

3.3. Pose-Assisted Segmentation

Now, we combine the modified KDE-EM scheme with local MRF constraints and global pose priors to form a single iterative algorithm: pose-assisted segmentation. The global shape prior is enforced by iteratively fitting the trained PS model to the current foreground assignment probability map and updating the probability map with the binary model coverage image as an adaptive weighted sum. The segmentation and pose estimation are performed in an interleaved and cooperative manner (Figure 5).

Algorithm 2 Pose-Assisted Segmentation

Initialization : As in KDE-EM.
M − Step : As in KDE-EM.
E − Step I : Incorporating local MRFs.

$$F^t(\mathbf{y}) = cF^{t-1}(\mathbf{y})\Psi_{\mathcal{F}}^t(\mathbf{y}) \sum_{i=1}^{N} F^{t-1}(\mathbf{x}_i) \prod_{j=1}^{d} k(\frac{y_j - x_{ij}}{\sigma_j}), \quad (9)$$

$$B^t(\mathbf{y}) = cB^{t-1}(\mathbf{y})\Psi_{\mathcal{B}}^t(\mathbf{y}) \sum_{i=1}^{N} B^{t-1}(\mathbf{x}_i) \prod_{j=1}^{d} k(\frac{y_j - x_{ij}}{\sigma_j}), \quad (10)$$

E − Step II : Adaptive pose inference on the soft segmentation and assignment probability update by the estimated poses.
1. Fit the PS model $\theta \in \mathbf{\Theta}$ to the current foreground probability map F^t to find the maximum a posteriori (MAP) solution as: $\theta_t^* = \arg\max_{\theta\in\mathbf{\Theta}} l(\theta)\rho_{\mathbf{t}}(\theta)$, where $\rho_t(\theta)$ is calculated as the similarity of the foreground assignment probability F^t and the binary model coverage image $M(\theta)$:

$$\rho_t(\theta) = 1 - \frac{\sum_{\mathbf{x}} \|F^t(\mathbf{x}) - M(\theta,\mathbf{x})\|}{n}. \quad (11)$$

Similar to the PS model fitting scheme, we employ the belief propagation algorithm in the reduced search space for estimating the best fitting model θ_t^*.
2. Use the binary model coverage image $M^t = M(\theta_t^*)$ to update the foreground probability map F^t as follows:

$$F_{new}^t = (1 - \omega_t)F^t + \omega_t M^t,\ F_{new}^t \mapsto F^t, \quad (12)$$

$$B^t = \mathbf{1}_{h\times w} - F^t, \quad (13)$$

where $\mathbf{1}_{h\times w}$ is an all-1 matrix and $\omega_t = \beta^t \rho_t(\theta)^\gamma$ is an adaptive weight to control the iteration based on the current model fitting score.
Segmentation : As in KDE-EM.

4. Segmentation of Multiple Occluded Objects

For the case of multiple objects, a set of bounding boxes are roughly provided around foreground objects by user interaction as in [4, 13]. The order of interactive assignments can be arbitrary and the occlusion ordering is inferred from the segmentation process.

Given an image $\mathbf{I}$ and a set of initial human hypotheses, $(\mathbf{x}_k, \mathbf{s}_k), k = 1, 2, ...K$, where $\mathbf{x}_k$ and $\mathbf{s}_k$ denote the location and scale of each human, the problem of segmentation is the $(K + 1)$-class (K humans and background) pixel labelling problem. The label set is denoted as $\mathcal{F}_1, \mathcal{F}_2, ...\mathcal{F}_K, \mathcal{B}$. Given a pixel $\mathbf{y}$, we denote the probability of the pixel $\mathbf{y}$ belonging to human-k as $F_k^t(\mathbf{y})$, and the probability of the pixel $\mathbf{y}$ belonging to the background as $B^t(\mathbf{y})$, where $t = 0, 1, 2...$ is the iteration index. The

assignment probabilities $F_k^t(\mathbf{y})$ and $B^t(\mathbf{y})$ are constrained to satisfy the condition: $\sum_{k=1}^{K} F_k^t(\mathbf{y}) + B^t(\mathbf{y}) = 1$.

4.1. Layered Occlusion Model

We introduce a layered occlusion model into the initialization step for segmentation of multiple occluded objects. Layered representation have been used in [10] for motion segmentation. The background is assumed to be in the farthest back layer. Given a hypothesis of an occlusion ordering, we build our layered occlusion representation iteratively by calculating the foreground probability map F_k^0 for the current layer and its residual probability map R_k^0 for pixel $\mathbf{y}$. Suppose the occlusion order (from front to back) is given by $\mathcal{F}_1, \mathcal{F}_2, ...\mathcal{F}_K, \mathcal{B}$; then the initial probability map (Figure 3) is calculated recursively as follows:

Algorithm 3 Initialization by Layered Occlusion Model

initialize $R_0^0(\mathbf{y}) = 1$ for all $\mathbf{y} \in \mathbf{I}$
for $k = 1, 2, ...K$
– for all $\mathbf{y} \in \mathbf{I}$
– $F_k^0(\mathbf{y}) = R_{k-1}^0(\mathbf{y}) e^{-1/2(Y-Y_0)^t V^{-1}(Y-Y_0)}$
– $R_k^0(\mathbf{y}) = 1 - \sum_{j=1}^{k} F_j^0(\mathbf{y})$
endfor
return $F_1^0, F_2^0, ..., F_K^0$ and $B^0 = R_K^0$

4.2. Pose-Assisted Segmentation for Multiple Occluded Objects

We generalize the single-human segmentation scheme presented in the previous sections. We first incorporate the contrast-dependent MRF to regularize the probability maps in the $E - Step\ I$, and perform the PS pose model inference on individual probability maps F_k^t for each object and update the probability maps in the $E - Step\ II$. Based on the pose inference on individual probability maps, we explicitly reason about occlusion status between humans by comparing the assignment probabilities of the pixels in the occluded regions. Our pose-assisted segmentation approach performs segmentation, pose estimation and occlusion reasoning simultaneously in an interleaved, iterative process where occlusion reasoning is applied as a prior to update the assignment probability maps at each iteration.

Occlusion reasoning: the initial occlusion ordering is determined by sorting the hypotheses by their vertical coordinates and the layered occlusion model is used to estimate initial assignment probabilities. The occlusion status is updated at each iteration after the $E - step\ I$ by comparing the evidence of occupancy in the overlap area between different object hypotheses. For two object hypotheses H_i and H_j, if they have overlap area O_{H_i,H_j}, we estimate the occlusion ordering between the two as: H_i occlude H_j if

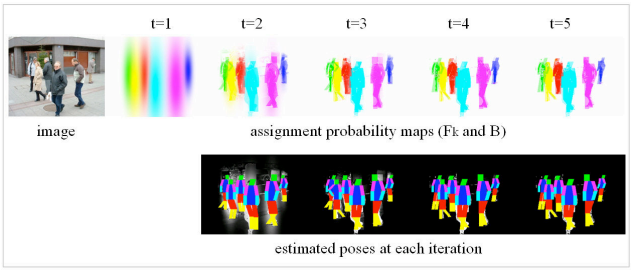


Figure 3. The process of pose-assisted segmentation for multiple occluded objects.

$\sum_{\mathbf{x} \in O_{H_i,H_j}} F_i(\mathbf{x}) > \sum_{\mathbf{x} \in O_{H_i,H_j}} F_j(\mathbf{x})$ (*i.e.* H_i better accounts for the pixels in the overlap area than H_j), H_j occlude H_i otherwise, where F_i^t and F_j^t are the foreground assignment probabilities of H_i and H_j. At each iteration, every pair of hypotheses that have a non-empty overlap area is compared in this way. The whole occlusion ordering is updated by exchanges if and only if the estimated pairwise ordering differs from the previous ordering. Similar reasoning scheme have been explored in [10] using $\alpha\beta$-swap and α-expansion algorithms.

5. Experiments and Evaluation

In this section, we first present experiments on initialization sensitivity and then discuss qualitative and quantitative results for both single and multiple occluded human segmentation. In the experiments, the segmentation accuracy $\gamma[\%]$ is defined as the proportion of pixels correctly classified as foreground or background by comparing the binary segmentation result with the ground truth: $\gamma = \left(1 - \frac{\sum_{\mathbf{x}} |F(\mathbf{x}) - H(\mathbf{x})|}{n}\right) \times 100\%$, where F is the binary segmentation image and H is the hand-segmented ground truth. The constants β and γ are set to $\beta = 0.9$, $\gamma = 4$, and remained constant during the experiments.

5.1. Initialization Sensitivity

The sensitivity of segmentation accuracy with respect to the initialization bias (scale, shift-x, shift-y) is tested for various images and results for a typical example are shown in Figure 4. (Note that results for other examples are very similar). The sensitivity curves show that segmentation accuracy decreases monotonically (but very slowly) with respect to scale and horizontal/vertical shifts. Specifically, the best segmentation accuracy is above 98% which is achieved with the true bounding box, and the accuracy is above 96% when the scale factor is in the range [0.75 1.25], when the horizontal shift factor is below 0.4, and when the vertical shift factor is below 0.42. Also, the accuracy remains above 90% when the scale factor is in the range [0.5 1.5], and remains above 92% and approximately above 90% when the

Algorithm 4 Pose-Assisted Segmentation for Multiple Occluded Objects

Initialization : By the layered occlusion model.
M − Step : As in KDE-EM.
E − Step I : Assignment probability updates for multiple foreground objects and background.

$$F_k^t(\mathbf{y}) = cF_k^{t-1}(\mathbf{y})\Psi_{\mathcal{F}_k}^t(\mathbf{y})\sum_{i=1}^{N} F_k^{t-1}(\mathbf{x}_i)\prod_{j=1}^{d} k(\frac{y_j - x_{ij}}{\sigma_j}), \quad (14)$$

$$B^t(\mathbf{y}) = cB_k^{t-1}(\mathbf{y})\Psi_{\mathcal{B}}^t(\mathbf{y})\sum_{i=1}^{N} B^{t-1}(\mathbf{x}_i)\prod_{j=1}^{d} k(\frac{y_j - x_{ij}}{\sigma_j}), \quad (15)$$

where c is a normalizing constant such that $\sum_{k=1}^{K} F_k^t(\mathbf{y}) + B^t(\mathbf{y}) = 1$.
E − Step II : Adaptive pose inference on the soft segmentation and assignment probability update by the estimated poses.
1. Update the occlusion ordering
2. Fit the PS pose model $\theta \in \mathbf{\Theta}$ to the current foreground probability map F_k^t to find the maximum a posteriori (MAP) estimation as: $\theta_{k,t}^* = \arg\max_{\theta\in\mathbf{\Theta}} l(\theta)\rho_{k,t}(\theta)$, where

$$\rho_{k,t}(\theta) = 1 - \frac{\sum_{\mathbf{x}} \|F_k^t(\mathbf{x}) - M(\theta, \mathbf{x})\|}{n}. \quad (16)$$

We perform MAP optimization for each hypothesis to estimate the set of best fitting models $\theta_{k,t}^*, k = 1, 2, ...K$ for the current iteration step t.
3. Use the set of binary model coverage images $M_k^t = M(\theta_{k,t}^*)$, $k = 1, 2, ...K$ to update the foreground probability maps F_k^t, $k = 1, 2, ...K$ as follows:

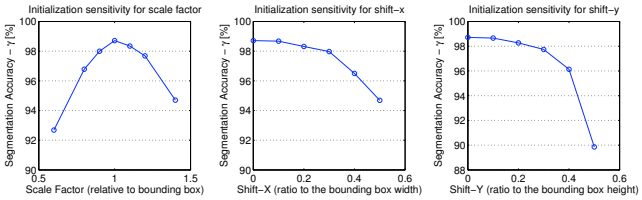
$$F_{k_{new}}^t = (1 - \omega_t)F_k^t + \omega_t M_k^t, \; F_{k_{new}}^t \mapsto F_k^t, \quad (17)$$

$$B^t = \mathbf{1}_{h\times w} - \sum_k F_k^t, \quad (18)$$

where $\omega_t = \beta^t \rho_{k,t}(\theta)^\gamma$.
Segmentation : The iteration is terminated when $\frac{\sum_k \sum_{\mathbf{y}} \{|F_k^t(\mathbf{y}) - F_k^{t-1}(\mathbf{y})|\}}{n} < \epsilon$. We denote $F_k(\mathbf{y})$ and $B(\mathbf{y})$ as the final converged assignment probabilities. Then the final segmentation is determined as: pixel $\mathbf{y}$ belong to human-k, *i.e.* $\mathbf{y} \in \mathcal{F}_k, k = 0, 1, ...K$ (where $k = 0$ corresponds to background $\mathcal{F}_0 = \mathcal{B}$), if $k = \arg\max_{k\in\{0,1,...K\}} F_k^t(\mathbf{y})$.

horizontal and vertical shift factors increase from 0 to 0.5. We only consider sensitivity in these intervals since the initialization rectangle will have less than 50% overlap with the object region for more severe biases. Horizontal shifts tend to be less sensitive than scale change and vertical shifts.

(a) Initialization and Ground Truth Segmentation



(b) Initialization Sensitivity Analysis

Figure 4. Experiments on initialization sensitivity. (a) Ground truth and biased bounding boxes, (b) Sensitivity w.r.t. scale, shift-x, and shift-y.

5.2. Results on Single-human Segmentation

We have tested our approach to single-human segmentation on the INRIA person dataset [1]. Figure 5 shows comparison of the segmentation performances for GrabCut, KDE-EM, CDMRF-KDE-EM, and the proposed approach. KDE-EM resulted in a very inaccurate segmentation with many holes and isolated small regions. GrabCut obtained coherent segmentations but the results are very sensitive to the interactive initialization and does not guarantee a human-like segmentation. CDMRF-KDE-EM obtained a coherent segmentation but incorrectly included background regions in the segmentation. In contrast, with the local MRF and global shape priors provided by the PS pose model inference, our approach achieved the best result, and the segmentation accuracy almost reached the ground truth (98.71%) for this example. Results for more difficult examples are shown in Figure 6.

We also quantitatively evaluated the proposed segmentation approach on a subset of 100 test images from the INRIA person dataset [1] and compared it with KDE-EM [19]. The set of test images are chosen to avoid redundancies of mirror images and overlap with the training set. Figures 7(a) and 7(b) show some examples of test images and the quantitative comparison results. The distribution of the performance is evaluated by sorting the images by segmentation accuracy and number of iterations. The result shows that our proposed approach outperformed KDE-EM significantly in segmentation accuracy. For the number of iterations to convergence 7(c), our approach achieved slightly better convergence (fewer iterations) than KDE-EM (this is mainly due to the recursive soft probability update).

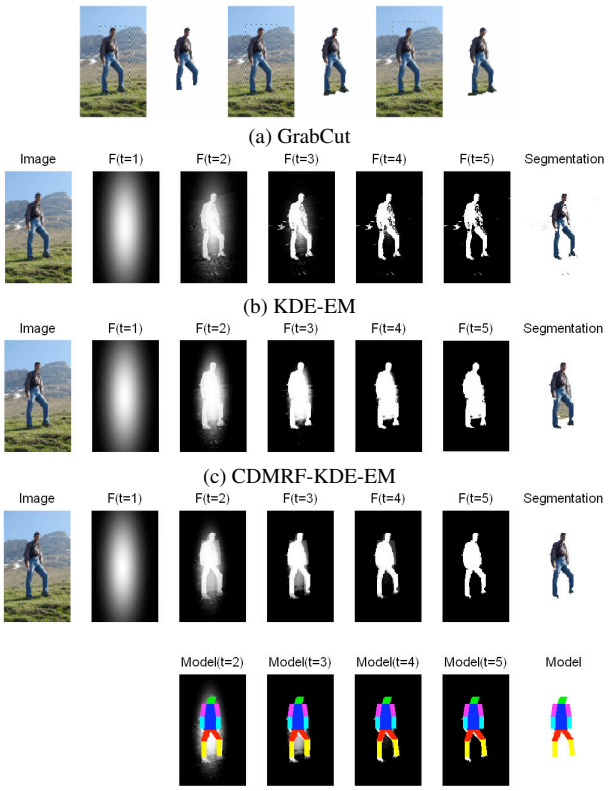


Figure 5. Example processes of segmentation approaches. (a) GrabCut [13] segmentation for three different initializations, (b) KDE-EM: EM soft-labelling using weighted kernel density estimation, (c) CDMRF-KDE-EM: KDE-EM combined with local contrast-dependent MRF constraints, (d) Pose-assisted segmentation.

5.3. Results on Multi-human Segmentation

We compared our multi-human segmentation approach to G-KDE-EM (KDE-EM generalized to the case of multiple objects) on a variety of test images. Figure 8 shows some results on our segmentation and pose estimation results for images with multiple occluded humans. Our approach achieved good segmentation and pose estimation results even with severe inter-occlusions between humans, while KDE-EM resulted in poor segmentations with few human-like segmentations. This is as expected since KDE-EM does not enforce any prior knowledge in the segmentation. Finally, the running time and the number of iterations needed for our multi-human segmentation algorithm are similar to the cases of single human segmentation.

6. Conclusions

The KDE-EM framework has fast convergence and achieves accurate results for color-based segmentation. Our incorporation of local contrast-dependent MRF and PS pose

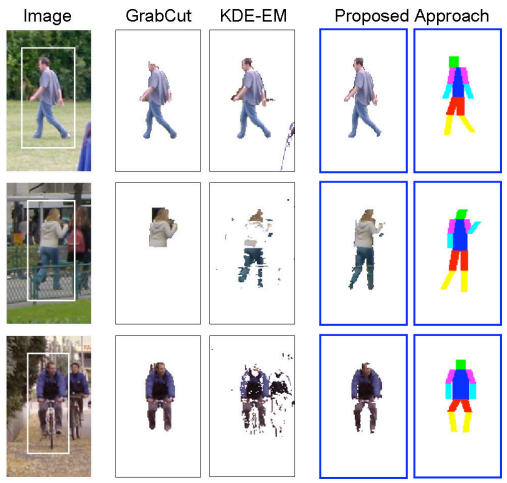


Figure 6. Results for more test images with increasing complexity. From left to right are original image with selected bounding boxes, result using GrabCut, result using KDE-EM, and segmentation and pose estimation results using our proposed method. Note that in these examples, we assume there is single foreground object and only segment the human in the center of the image.

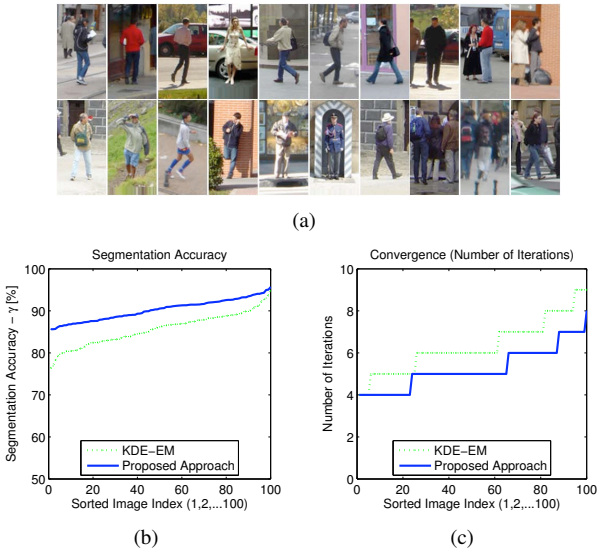


Figure 7. Quantitative performance evaluation. (a) Sample test images, (b) Comparison of segmentation accuracy, (c) Comparison of convergence rates.

model inference shows the combined local and global priors give very accurate segmentations, while human poses are estimated simultaneously. The pose-assisted segmentation approach is also generalized to the case of multiple occluded human segmentation based on a layered occlusion model and a probabilistic occlusion reasoning method. Experiments show that our approach improves KDE-EM to a large extent while preserving the basic computational cost

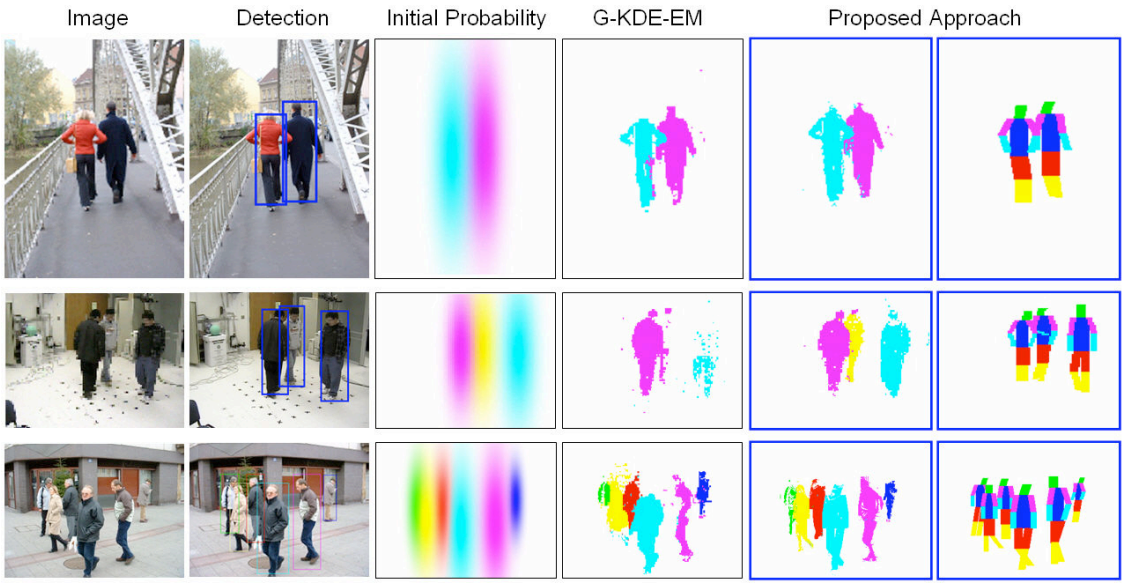


Figure 8. Comparison of segmentation and pose estimation for occluded cases.

and running time. Currently our approach can deal with most standing human poses (front/back and side views) but has limitations on handling self-occlusion and performing inference on more difficult poses. We need to extend our system to incorporate in-process user adjustments (*e.g.* correcting orientation of arms) to handle pose inference in these these cases. Another future direction is to generalize the approach to the cases of other object categories.

Acknowledgement

This research was funded in part by the U.S. Government VACE program.

References

[1] INRIA Person Dataset, http://pascal.inrialpes.fr/data/human/.

[2] A. Blake, C. Rother, M. Brown, P. Perez, and P. Torr. Interactive Image Segmentation using an Adaptive GMMRF model. *ECCV*, pages Vol I:428-441, 2004.

[3] E. Borenstein and J. Malik. Shape Guided Object Segmentation. *CVPR*, 2006.

[4] Y. Y. Boykov and M.-P. Jolly. Interactive Graph Cuts for Optimal Boundary and Region Segmentation of Objects in N-D images. *ICCV*, 2001.

[5] M. Bray, P. Kohli, and P. H. Torr. PoseCut: Simultaneous Segmentation and 3D Pose Estimation of Humans using Dynamic Graph-Cuts. *ECCV*, 2006.

[6] C. Carson, S. Belongie, H. Greenspan, and J. Malik. Blobworld: Image Segmentation Using EM and Its Application to Image Querying. *IEEE Trans. PAMI*, 24(8):1026–1038, 2002.

[7] D. Comaniciu and P. Meer. Mean-Shift: A Robust Approach Toward Feature Space Analysis. *IEEE Trans. PAMI*, 24(5), 2002.

[8] P. F. Felzenszwalb and D. P. Huttenlocher. Pictorial Structures for Object Recognition. *International Journal of Computer Vision*, 61(1):55–79, 2005.

[9] S. Gordon, H. Greenspan, and J. Goldberger. Applying the Information Bottleneck Principle to Unsupervised Clustering of Discrete and Continuous Image Representations. *ICCV*, 2003.

[10] M. P. Kumar, P. H. S. Torr, and A. Zisserman. Learning Layered Motion Segmentations of Video. *ICCV*, 2005.

[11] M. P. Kumar, P. H. S. Torr, and A. Zisserman. Obj Cut. *CVPR*, 2005.

[12] D. Ramanan. Learning to Parse Images of Articulated Bodies. *NIPS*, 2006.

[13] C. Rother, V. Kolmogorov, and A. Blake. Grabcut - Interactive Foreground Extraction using Iterated Graph Cuts. *ACM Transactions on Graphics: SIGGRAPH*, 2004.

[14] D. Scott. Multivariate Density Estimation. *Wiley Interscience*, 1992.

[15] J. Shi and J. Malik. Normalized Cuts and Image Segmentation. *IEEE Trans. PAMI*, 22(8), 2000.

[16] J. Winn and N. Jojic. LOCUS: Learning Object Classes with Unsupervised Segmentation. *ICCV*, 2005.

[17] J. Winn and J. Shotton. The Layout Consistent Random Field for Recognizing and Segmenting Partially Occluded Objects. *CVPR*, 2006.

[18] L. Zhao and L. S. Davis. Closely Coupled Object Detection and Segmentation. *ICCV*, 2005.

[19] L. Zhao and L. S. Davis. Iterative Figure-Ground Discrimination. *ICPR*, 2004.