

Simultaneous Estimation of Segmentation and Shape

Jens Rittscher Peter H. Tu and Nils Krahnstoeber
General Electric Global Research, One Research Circle, Niskayuna, NY 12309, USA
{jens.rittischer|nils.krahnstoeber|peter.tu}@research.ge.com

Abstract

The main focus of this work is the integration of feature grouping and model based segmentation into one consistent framework. The algorithm is based on partitioning a given set of image features using a likelihood function that is parameterized on the shape and location of potential individuals in the scene. Using a variant of the EM formulation, maximum likelihood estimates of both the model parameters and the grouping are obtained simultaneously. The resulting algorithm performs global optimization and generates accurate results even when decisions can not be made using local context alone. An important feature of the algorithm is that the number of people in the scene is not modeled explicitly. As a result no prior knowledge or assumed distributions are required. The approach is shown to be robust with respect to partial occlusion, shadows, clutter, and can operate over a large range of challenging view angles including those that are parallel to the ground plane. Comparisons with existing crowd segmentation systems are made and the utility of coupling crowd segmentation with a temporal tracking system is demonstrated.

1. Introduction

The focus of this work is the segmentation of groups of people into individuals. Although sophisticated person detectors such as [4] have been developed, they usually assume that people are well separated. Others such as [5, 13] have used mechanisms such as head detectors to segment crowds. However, failure modes occur when these features cannot be directly observed. Alternatively one can extract a large set of image features and partition these into individuals using explicit feature grouping or a form of model based segmentation.

Two examples of the feature grouping approach are Song *et. al.* [9] and Mikolajczyk *et. al.* [6]. These systems define a likelihood function that can determine whether or not a given set of features should be grouped together. Segmentation is achieved through a greedy search where the most likely grouping is first identified and then removed. This process is repeated until all image features have been explained. Since these likelihood functions do not con-

sider all image features simultaneously, difficulties can occur when there is high ambiguity associated with the local context. This can occur in regions such as the center of dense crowds. For this reason a global optimization may be desirable.

Parameter driven generative models, as used in model based segmentation, can completely describe a scene. The parameters may include the number of people, their location and their shape. For a given set of parameters, the grouping of all features can be directly inferred and subsequently evaluated. Hence global optimization can be achieved by searching for maximum likelihood estimates for the model parameters. The main issues are that the parameter space can be very complex. Optimization then requires good initial estimates and expensive search techniques. Elgammal *et. al.* [2] assumed prior knowledge regarding the number of people and their appearance. Optimization was achieved by exhaustive local search. Zhao *et. al.* [13] make no assumptions about the number of people in the scene and use Markov Chain Monte Carlo (MCMC) to perform their optimization.

A new approach that combines the power of the generative models with the simplicity associated with the feature grouping methods is needed. We propose a system that performs global optimization, makes no assumptions about the number of people in the scene, needs only trivial initialization, and does not require random search. The approach is based on an Expectation Maximization (EM) formulation which has shape parameters for all potential individuals and treats feature assignments, similar to Tu and Yuille [12], as hidden variables. A relaxation scheme starts with a uniform distribution for the hidden variables and through an annealing process ultimately arrives at a maximum likelihood estimate for both the shape parameters and feature assignments.

In the following section a formal definition of the hidden variable approach will be given. Implementation details will be discussed in section 3. The important question of how the algorithm can effectively be used within a visual tracking framework is addressed in section 6.

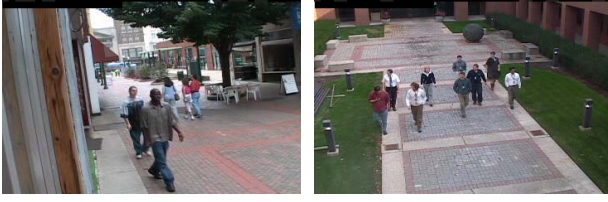


Figure 1: **Crowd Segmentation Problem.** Differences in viewing angle, appearance, and partial occlusion makes crowd segmentation a very challenging problem.

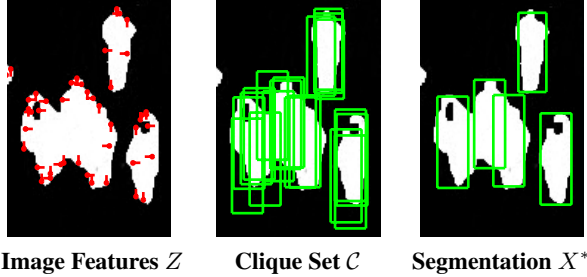


Figure 2: **Image Features, Cliques, and Shape Parameters.** The feature extraction, the computation of the set of cliques \mathcal{C} and segmentation results are shown for one example image. Note that a standard probabilistic background model was used to segment the image foreground. The set of image features Z illustrates that each feature z_i is labeled as being at the top side or bottom of a person. The set of cliques \mathcal{C} illustrates that the algorithm generates a large number of cliques. The segmentation on the right shows a segmentation for the MAP estimate X^* and one sample of V . Relevant details are described in section 4.

2. Problem Formulation

The set of N observations $Z = \{z_i\}$ consists of image features which can include corner points, edges, image regions or the output of application specific detector schemes. A geometric shape model is used to identify which subsets $C \subset Z$ can be associated with a single person. In a pre-processing step a set of K possible groups of features, also referred to as cliques [11], are identified. The set of all cliques is defined as

$$\mathcal{C} := \{C_1, \dots, C_K\}. \quad (1)$$

An assignment vector $V = \{v_i\}$ of length N with $v_i \in [1, \dots, K]$ is used to associate each feature z_i with a particular clique C_k . The association of features to cliques is directly coupled with questions regarding the assumed shape of people. This is why cliques C_k are associated with parameters x_k that encode the location and shape of people. The collection of shape parameters is denoted as

$$X = [x_1, \dots, x_K]^T. \quad (2)$$

The focus of our approach is to model the joint probability of an assignment vector V and a feature set Z , i.e. $p(V, Z; X)$. Here X denotes the parameters of the distribution. The reader should note that the range of the random variable V , \mathcal{V} is defined by the set of cliques \mathcal{C} . The particular challenge here is that the assignments of features z_i to cliques C_k cannot be observed directly. This is why V is treated as a hidden variable. Given a set of image features Z the task is to find the maximum likelihood estimate of X and a distribution for V which can be sampled to generate likely assignments. We address this estimation problem using EM.

One possibility for modeling the joint probability $p(V, Z; X)$ is to define a merit function for a particular feature assignment V , given a set of image features Z . Measuring the affinity of a particular subset of image features (z_i, \dots, z_j) to a particular clique is one way of doing this. Here we limit our analysis to modeling the affinity of single feature assignments as well as pairwise assignments to a given clique C_k with shape parameters x_k . The corresponding affinity functions are denoted as $g(z_i, x_k)$ and $g(z_i, z_j, x_k)$. We formulate the log likelihood of a feature assignment given a set of image features Z as

$$L(V|Z; X) \propto \gamma_1 \sum_{k=1}^K \sum_{i=1}^N g(z_i, x_k) \delta_{|C_k}(v_i) + \gamma_2 \sum_{k=1}^K \sum_{\substack{i,j=1 \\ i \neq j}}^N g(z_i, z_j, x_k) \delta_{|C_k}(v_i, v_j), \quad (3)$$

where $\delta_{|C_k}(\cdot)$ is an indicator function which is one in case $v_i = k$ and set to zero otherwise. In this particular case it is not necessary to compute the normalization constant which depends on \mathcal{V} since the set of cliques \mathcal{C} does not change. Since the value of $p(Z)$ remains constant throughout the formulation, we have

$$p(V, Z; X) = p(V|Z; X) p(Z) \propto \exp(L(V|Z; X)). \quad (4)$$

The EM algorithm estimates the parameters of the model and the distribution of V iteratively starting from some initial guess. Neal and Hinton [7] present a variant of EM in which both the E and the M steps are seen as maximizing, or at least increasing the function $F(\tilde{P}, X)$, which is defined as

$$F(\tilde{P}, X) = E_{\tilde{P}}[\log P(V, Z; X)] + H(\tilde{P}), \quad (5)$$

where \tilde{P} is the probability density function of V . The entropy of this distribution is denoted by H . The function is related to the 'free energy' used in statistical physics. This version of EM justifies incremental versions of the algorithm, which in effect employ a partial E step, as well as sparse versions, in which most iterations update only parts

of the distribution \tilde{P} . An iteration of the standard EM algorithm can be expressed in terms of $F(\tilde{P}, X)$ as follows:

- E-Step: Set \tilde{P} to the \tilde{P} that maximizes $F(\tilde{P}, X)$
M-Step: Set X^* to the X that maximizes $F(\tilde{P}, X)$. (6)

In analogy to the classical mean-field theory we assume that \tilde{p} is independent with respect to the v_i , i.e.

$$\tilde{P}(V) = \prod_{i=0}^N \tilde{P}(v_i) . \quad (7)$$

This independence assumption simplifies the free energy $F(\tilde{P}, X)$ (5). Given a particular clique C_k this independence assumption will not hold and the resulting limitations of the algorithm need to be studied. A similar problem is presented in [10]. One general remark that can be made is that this independence assumption compromises the influence of the underlying shape model. Potentially, image regions that are similar to partial people can generate strong cliques even though the rest of the body is not explained. In addition this assumption does not allow for any explicit occlusion reasoning. In the following $\tilde{P}(V)$ is represented as a matrix $M = \{m_{ik}\}$ where $m_{ik} := \tilde{P}(v_i = k)$.

Based on the model specified in equations (4) and (3), the function $F(\tilde{P}, X)$ is of the form:

$$\begin{aligned} F(\tilde{P}, X) = & \sum_{k=1}^K \sum_{i=1}^N g(z_i, x_k) m_{ik} \\ & + \sum_{k=1}^K \sum_{i,j=1; i \neq j}^N g(z_i, z_j, x_k) m_{ik} m_{jk} \quad (8) \\ & - \sum_{k=1}^K \sum_{i=1}^N m_{ik} \log(m_{ik}) . \end{aligned}$$

The assignment vector V that maximizes $P(V, Z; X)$ for a given set of image features Z and shape parameters X is denoted as V^* , i.e.

$$P(V^*, Z; X) \geq P(V, Z; X) \quad \forall V \in \mathcal{V} .$$

It can be shown that

$$\forall \tilde{P} \log P(V^*, Z; X) \geq E_{\tilde{P}}[\log P(V, Z; X)] .$$

Therefore the expectation term in equation 5 is maximal when $\tilde{P}(V) = \delta(V^*)$. The entropy term on the other hand favors broader distributions which implicitly imposes a lower bound on the variance of \tilde{P} . In order to make this optimization problem more tractable we introduce, similar to [1], an explicit regularization parameter T , i.e.

$$F_T(\tilde{P}, X) = E_{\tilde{P}}[\log P(V, Z; X)] + TH(\tilde{P}) . \quad (9)$$

At a sufficiently high starting temperature \tilde{P} will be uniform since the value of the functional is dominated by the entropy term. Annealing will then be applied to determine the solution by lowering the temperature using an appropriate annealing schedule. The relevant details of this step will be discussed in the following section.

EM Estimate of X

Initialization:

- Compute the set of cliques \mathcal{C} , initial distribution \tilde{P}_0 , and X_0 . Set $T = T_0$.

Until $T = T_\infty$:

- Iterate:
 - E-Step: Set \tilde{P}_s by updating the assignment probabilities m_{ik} according to 10. (see section 3.2)
 - M-Step: Set X_s such that it optimizes $F_T(\tilde{P}_s, X)$. (see section 3.3)
- while $F_T(\tilde{P}_s, X_s) - F_T(\tilde{P}_{s-1}, X_{s-1}) \geq \epsilon$.
- Update T with $T = \alpha T$.

Figure 3: Algorithm Overview.

3. Algorithm Overview

As opposed to the traditional EM algorithm we adopt a variant of EM that maximizes a joint function $F_T(\tilde{P}, X)$ of the parameters, X , and the distribution for the unobserved variables $\tilde{P}(V)$. The implementation of the E-step and M-step will be presented in sections 3.2 and 3.3 respectively.

Any EM parameter estimation critically depends on the initialization method. The entropy term H in $F_T(\tilde{P}, X)$ imposes an implicit lower bound on the variance of \tilde{P} . At sufficiently high temperatures this entropy term H dominates the value of the function $F_T(\tilde{P}, X)$. In consequence the initialization problem becomes trivial since \tilde{P}_0 can be set to a uniform distribution over the range of V , \mathcal{V} . As part of the initialization process, presented in section 3.1, \mathcal{V} is computed by applying a geometric shape model to analyze which image features z_i can be part of the same clique C_k . As part of this initialization, an initial estimate of X is computed directly. An overview of the algorithm is given in figure 3. Section 3.4 contains a brief discussion of the annealing schedule.

3.1. Initialization

The decision on which image features can be grouped together depends on the shape model. In this implementation, knowledge of the ground plane is used. The shape of a person varies as a function of its position on the ground plane. Other information such as statistical background models or

object specific image features can be used to derive sophisticated clique nomination schemes - see section 4.3. As a result of this process we obtain a set of cliques \mathcal{C} and an initial estimate of X . The set of cliques \mathcal{C} automatically defines the range of the random variable V . This is why we can compute \tilde{P}_0 , represented by M_0 as follows:

$$M_0 = \{m_{ik}\} \quad \text{with} \quad m_{ik} = \frac{\delta_{|C_k}(z_i)}{\sum_l \delta_{|C_l}(z_i)} .$$

Note that no knowledge of the number of people present in the scene is required. The initial choice of \tilde{P} ensures no bias is given to any particular $V \in \mathcal{V}$.

3.2. E-Step

The objective of the E-step is to maximize $F_T(\tilde{P}, X)$ with respect to \tilde{P} , which is specified using the matrix M (see equation 7). The interpretation of EM [7] adopted here justifies a partial optimization. One possibility is to gradually increase, like in traditional soft assign [1], those assignment probabilities m_{ik} which increase the value of $F_T(\tilde{P}, X)$. In our case the matrix M is not a doubly stochastic matrix, since multiple vertices can be assigned to a single clique C_k . For every $i \in [1, \dots, N]$ and $k \in [1, \dots, K]$ the assignment probability m_{ik} is updated as [11]

$$m'_{ik} = \exp\left(\frac{1}{T} \partial_{m_{ik}} E_{\tilde{P}^s} [\log P(V, Z; X^s)]\right) . \quad (10)$$

The partial derivative $\partial_{m_{ik}} E_{\tilde{P}^s} [\log P(V, Z; X^s)]$ has the form

$$g(z_i, x_k) + 2 \sum_{j=1; i \neq j}^N g(z_i, z_j, x_k) m_{jk}$$

Based on the independence assumption made for \tilde{P} the following normalization is sufficient:

$$m_{ik}^{s+1} = \frac{m'_{ik}}{\sum_l m'_{il}} . \quad (11)$$

Depending on T this update gradually increases the assignment probabilities which increase $F_T(\tilde{P}, X)$. The benefit of this approach is that it is robust and computationally efficient.

3.3. M-Step

The model parameter X encodes the shape of each clique C_k . The set of image features that can be associated with a given clique C_k does not change during this process, which implies that the range of V is fixed. In turn this imposes limits on the rate of change of X . The weaker a particular assignment probability $P(v_i = k)$ gets, the less it influences the shape parameter x_k . In this work, the dimension

of the shape space is small. It should be noted that this optimization is independent of the entropy term H and does therefore not depend on the temperature T .

Since the affinity functions $g(\cdot)$ depends on C_k this optimization problem is non-linear and the parameters are strongly coupled. For a given clique C_k , the shape parameter x_k can only vary within limits. Consequently an exhaustive search is performed for every x_k independently.

3.4. Annealing

The annealing schedule controls the update process. Starting at T_0 the temperature is gradually reduced according to a linear annealing schedule (see figure 3). The initial temperature, T_0 , is determined by calculating the maximal partial derivative

$$m = \max_{i,k} \partial_{m_{ik}} E_{\tilde{P}^s} [\log P(V, Z; X^s)] .$$

In order to insure that the annealing process is slow enough but still sufficient, T_0 is set such that $\exp(m/T_0) = 1$. The lower the final temperature the less the influence of the entropy term H in $F_T(\tilde{P}, X)$.

At T_∞ , the M matrix becomes binary this implies that $\tilde{P}(V^*) = 1$ for some value of V^* . It can be shown that for the given values of Z and X , $\log(P(V, Z; X))$ is locally maximal at V^* .

4. Implementation

In order to implement the formulation defined in the previous section, the following mechanisms and parameterizations must be defined:

- a feature extraction process that provides the observations Z ,
- a shape space parameterized on x_i which defines the apparent contour of clique C_i ,
- a clique nomination scheme that generates \mathcal{C} ,
- the feature to object affinity function $g(z_i, x_k)$ which defines the cost of assigning z_i to clique C_k with shape parameters x_k ,
- the pairwise affinity $g(z_i, z_j, x_k)$ which defines pairwise affinity of z_i and z_j to the clique C_k with shape parameters x_k .

4.1. Feature Extraction

There are numerous approaches to extracting informative features from images. A common approach for selecting features which are indicative of a person is to make use of the bounding contour of the silhouette associated with the foreground (e.g. [14]) as shown in figure 2. It is often assumed that silhouette information is available, but the extraction of such features is challenging since the bounding contours can be corrupted by image noise.

label	d_i	r_k	σ_k
top	v_i	$v_k + h_k$	$\alpha_h * h_k$
bottom	v_i	v_k	$\alpha_h * h_k$
left	u_i	$u_k - 0.5w_k$	$\alpha_w * w_k$
right	u_i	$u_k + 0.5w_k$	$\alpha_w * w_k$

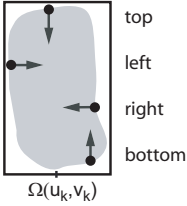


Table 1: **Feature Labels.** This table is used to define the values in equation (12). The values (u_i, v_i) are the coordinates of z_i . The values (u_k, v_k) are the current values of x_k . The values of h_k and w_k are the width and height of the rectangle $\Omega(u_k, v_k)$. The figure on the right illustrates features and their associated labels.

As opposed to using the entire bounding contour a number of segments s_n are taken and feature points z_i are sampled from these segments. Based on local contour information each segment is labeled as being on the top(t), bottom(b), left(l) or right(r) of a person. A point on a given line segment can have multiple labels. Given the label of the segment and the shape space information (see section 4.2), the potential width w and height h of an individual in the region of the line segment s_n can be computed. This estimate is then used to dynamically adjust the sampling rate of feature points z_i . Hence the feature extraction process is robust to a change of scale. This process is illustrated in figure 2.

4.2. Shape Space

A rectangular shape space is used to approximate the shape of a person in the image. The shape space is parameterized by $x = (u, v, w, h)$ where (u, v) are the coordinates of the center of the base of the rectangle $\Omega(u, v)$ which should be coincident with the feet of the person. It is assumed that people are standing vertically on a flat horizontal surface and that they are of uniform height. Thus there is a foot plane and a head plane. The 2D location of a person's feet in the foot plane is assumed identical to the 2D location of the person's head in the head plane. There is a homography H_f that maps the image plane to the foot plane and a homography H_h that maps the head plane back to the image plane. Under these imaging conditions there is a homography $H_{fh} = H_h * H_f$ that maps foot pixels to head pixels [11]. The height h of $\Omega(u, v)$ is then $|\mathcal{H}(u, v) - (u, v)|$. Assuming a constant aspect ratio r the initial width w of $\Omega(u, v)$ is set to $r * h$. The standard deviation of the difference in height and width of the person in the image is modeled as $\alpha_w * w$ and $\alpha_h * h$. The coefficients of \mathcal{H} are obtained in a calibration phase. The remaining parameters are kept site independent.

4.3. Clique Nomination

Given the foreground image and Z , a set of cliques \mathcal{C} must be nominated. The clique nomination process is intended to generate all likely groupings of image features. For a clique to be valid, every pair of its features must satisfy the constraint that they could both have been generated by a single individual. There are many possible criteria for this constraint. In this application we use geometric constraints based on the shape model. For any given rectangle $\Omega(u, v)$ all features that are contained in a slightly dilated versions of $\Omega(u, v)$ can form a valid clique. An exhaustive approach would be to consider all possible values of (u, v) and then prune based on redundancy. Since the complexity of the algorithm is directly affected by the number of nominated cliques, measures should be taken to restrict the nomination process. For a given foreground patch, all bounding pixels are identified. All rectangles $\Omega(u, v)$ such that (u, v) or $H_{fh}(u, v)$ is a bounding pixel are initially considered. All candidates with less than fifty percent overlap with the foreground patch are rejected. A chamfer map is constructed based on the bounding pixels and the chamfer distance for each of the remaining candidates is calculated. Non-maximal suppression is used to nominate the final set of cliques.

4.4. Assignment Affinity Functions

The merit of assigning z_i to a clique with shape parameters x_k is defined as

$$g(z_i, x_k) = 1 - (d_i - r_k)^2 / (\sigma_k), \quad (12)$$

where the values for d_i, r_k and σ_k depend on the label associated with z_i and are defined in table 1. If $g(z_i, x_k)$ falls below zero, it is set to zero.

The pairwise affinity $g(z_i, z_j, x_k)$ is set to the number of foreground pixels on the line segment l_{ij} divided by the length of l_{ij} . The end points of l_{ij} are the coordinates of z_i and z_j . It should be noted that in this implementation, $g(z_i, z_j, x_k)$ is independent of x_k .

5. Results

In order to provide a comprehensive analysis of the algorithm we compare our results with alternative approaches and analyze the effectiveness of our optimization method. Our final segmentations are based on hypothesized cliques with significant numbers of assigned features. As discussed in section 1 work by Zhao and Nevatia [13] is the most relevant to our approach.

One direct comparison is shown in figure 4. The strength of their approach is a more detailed parametric shape model. Based on a set of interest points found using a head detector they use MCMC estimation to identify the most likely

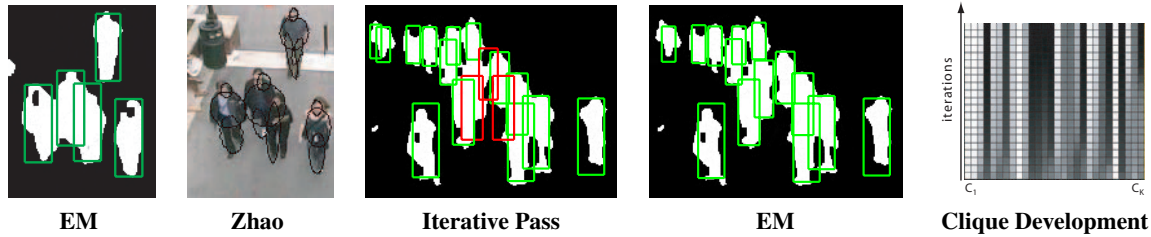


Figure 4: **Method Comparison.** A direct comparison with Zhao and Nevatia’s method [13] is presented. The result obtained by our method is shown in the left image, their result in the second from the left. The binary foreground image illustrates that this is a very challenging example. As our method uses the entire set of image features (shown in figure 2) it is capable of segmenting this group of people. A comparison of a basic iterative method described in section 5 (shown in the middle) and the proposed optimization scheme (shown on second image from the right) is presented. Incorrect detections are highlighted in red. Note that the proposed algorithm resolves the ambiguity at the center the foreground patch. The right image shows the average probability of feature assignment for each clique at each iteration of the relaxation algorithm. Note that it takes many iterations before all the feature assignments are resolved. (Data and results for the second image from the left courtesy of T. Zhao and R. Nevatia)

model parameters. Identifying all the head positions is often challenging (see figure 4). Since our algorithm accounts for all feature points it manages to segment this foreground region correctly. Fifteen similar examples taken from the same video sequence are presented in the accompanying supplementary material. All in all there were 54 possible detections. Our approach missed 4 and the algorithm described in [13] missed 6. There were no false detections for either algorithm.

The merit of performing simultaneous estimation via global optimization given the initial pre-processing steps is now demonstrated. In order to test whether or not a valid segmentation can be computed in a single pass as opposed to a sequence of iterative updates a simple feature grouping approach was implemented. In a single pass it iteratively removes the clique C_i which maximizes the log likelihood (3) assuming that all associated features are assigned to it. Those features are then removed from the set of image features Z . This process is repeated until there are no remaining features. A comparison of the results shown in figure 4 demonstrates that the simultaneous estimation of the segmentation and shape parameters using EM effectively propagates certainty from regions of low to high ambiguity. The ambiguities at the center of the crowd cannot be resolved based on local context information alone. In a separate set of experiments an additional degree of freedom was added to the shape space which allows for much greater variation in height. The results shown in figure 5 demonstrate that the M-step was able to adapt to both children and adults. The distribution of possible head positions in the image varies depending on the camera viewing angle. In figure 4, for example, knowledge of the head location in the image provides a strong estimate for the size of a person. However as the angle between the principle ray and the ground plane decreases, the ambiguity regarding the location of the feet given the location of the head increases dramatically. For



Figure 5: **Shape Estimation.** In a set of experiments one degree of freedom was added to the shape space so that children as well as adults can be segmented correctly. The initial shape of the clique is shown in green and final estimated size in red. Note that the algorithm accurately segments people of different heights.



Figure 6: **Viewing Angle.** Note almost all head locations in these image are close to the horizon. However there is a large variance in size. By taking all image features Z into account accurate segmentation is achieved. See text for further discussion.



Figure 7: **Challenges.** Varying scale, clutter, shadows, as well as partial occlusion make these examples particularly challenging. Note that the algorithm generates plausible results.

algorithms reliant on a strong shape model based on outputs of such mechanisms as head detectors this becomes problematic. Since our algorithm does not rely on any one type of feature and the hypothesis nomination scheme can produce a wide range of possible interpretations, robust performance can be achieved under these challenging viewing conditions. Figure 6 shows several results where all head locations are found near the horizon, the apparent body sizes do however vary significantly.

In figure 7 we show a successful segmentation for a complex scene where there is partial occlusion and like in the previous examples, the viewing angle is almost horizontal. The second image demonstrates the robustness of the algorithm to shadows and foreground clutter. When individuals walk behind one another, an elevated viewing angle is required. The third image demonstrates the ability of the algorithm to function under these viewing conditions.

6. Visual Tracking

The quality of the EM estimate depends of course on the extraction of image features Z and in particular on the quality of the foreground/background estimation process. Another factor are inherent ambiguities. For example, fully occluded individuals can not be detected but challenges are also posed by distractors such as backpacks or other items that may have the appearance of partially occluded people.

To address the above problems we show in this section how the proposed crowd segmentation algorithm can be integrated into the context of a larger tracking system (see Figure 8). The system consists of three main components: A standard low-level foreground estimation algorithm, a template-based tracker as well as the crowd segmentation, as presented in this work. All three components are combined into a tightly coupled framework. In the following, the tracker and the integrated crowd segmentation components are described in more detail.

6.1. Tracker

The tracker uses an adaptive appearance based approach similar to [8, 14]. The tracker is adaptive and can track

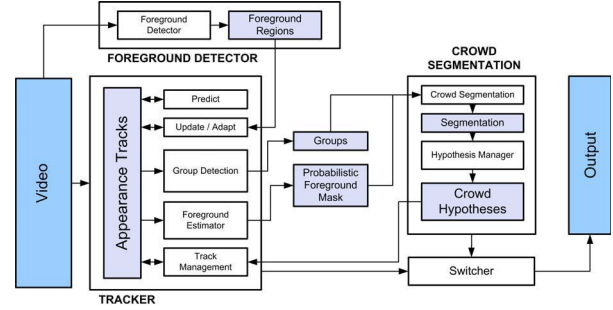


Figure 8: **System overview.** There are three components: foreground detector, probabilistic tracker, and the crowd segmentation algorithm. See text for details.

people and other targets such as vehicles alike. Various algorithms are in place for initiating, merging, splitting and deleting tracks. Each track is modeled by a color signature, an appearance template as well as a probabilistic target mask that is an autoregressive estimate of the foreground information as obtained in the previous stage. The tracker handles short term occlusions between isolated tracks, but groups closely spaced targets together into *group tracks*. Only foreground regions which are large enough to contain a number of people and image regions that contain closely spaced tracks are forwarded to the crowd segmentation algorithm for further analysis. In addition, an improved foreground region image is composed based on the information maintained by the tracker and also supplied to the crowd segmentation algorithm. The motivation for this is the following: The properties of the target masks compare favorably to the direct estimate of the foreground. First, the autoregressive process used to maintain the target masks suppresses high frequency variations and noise in the foreground image. Second, since the target masks are estimated from the foreground image relative to the moving tracks, foreground region information is effectively integrated across multiple images along the motion paths of targets, hence resulting in more accurate overall estimates.

6.2 Crowd Segmentation Component

The crowd segmentation algorithm processes all regions in the image that the tracking component has judged to contain groups of people. The resulting segmentation observation \hat{S}^t at frame t contains information about the detected number of people and their location in the image $\hat{S}^t = \{\hat{n}^t, (\hat{x}_i^t, \hat{y}_i^t), i = 0, \dots, \hat{n}^t\}$. As discussed above, noise in the feature extraction process as well as inherent ambiguities will inevitably lead for the estimate \hat{S}^t to deviate from the true state S^t . To reduce the error in the resulting segmentation, the estimated values are processed by a



Figure 9: **Crowd Tracking.** The yellow bounding boxes visualize the most likely hypotheses maintained by the crowd segmentation tracker.

simplified multiple hypothesis tracker. Within each group individual tracks are smoothed using a constant velocity Kalman filter. Two frames taken from a tracking experiment are shown in figure 9. Please note that the corresponding video sequence has been submitted as supplementary material.

7. Summary and Future Work

This paper presents an EM formulation for crowd segmentation which can be viewed as a combination of feature grouping and analysis by synthesis. The main strength of the approach is that global optimization is achieved without reliance on random search and strong initialisation. The number of individuals present in the scene need not be known in advance and the algorithm is robust with respect to clutter, shadows and partial occlusion. The algorithm has been successful over a variety of scales and challenging camera angles. An integrated approach that fuses tracking and crowd segmentation was proposed and demonstrated.

Although the shape space and features used in this implementation are relatively simple, more expressive models and informative features can be incorporated into this framework. As part of our future work we will develop shape models similar to those proposed by Felzenszwalb [3] and extend the existing M-Step such that it will fit the appropriate shape model to the image features. The set of images features will be extended to include the outputs of part specific detectors. This will overcome the limitations of reliance the silhouette information alone. The segmentation itself will be improved by using temporal priors when available. Complicated scenes such as brawls and dense retail environments, which are confounded with foreground clutter, are still beyond the scope of current systems. The approach presented in this paper could be a step towards solving these challenging problems.

References

[1] H. Chui and A. Rangarajan. A new point matching algorithm for non-rigid registration. *Computer Vision and Image*

Understanding, 89(3):114–141, March 2003.

- [2] A. Elgammal and L. Davis. Probabilistic framework for segmenting people under occlusion. In *Proc. 8th Int. Conf. on Computer Vision, Vancouver, BC*, volume 2, pages 145–152, 2001.
- [3] P. F. Felzenszwalb. Representation and detection of deformable shapes. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 27(3):208–220, 2005.
- [4] D. Gavrila. Pedestrian detection from a moving vehicle. In *Proc. 6th European Conf. Computer Vision, Dublin, Ireland*, volume 2, pages 37–49, 2000.
- [5] I. Haritaoglu, D. Harwood, and L. S. Davis. HYDRA: Multiple people detection and tracking using silhouettes. In *In IEEE International Workshop on Visual Surveillance*, pages 6–13, 1999.
- [6] C. Schmid K. Mikolajczyk and A. Zisserman. Human detection based on a probabilistic assembly of robust part detectors. In *Proc. 8th European Conf. Computer Vision, Prague, Czech Republic*, volume 1, pages 69–82, 2004.
- [7] R. M. Neal and G. E. Hinton. A view of the EM algorithm that justifies incremental, sparse, and other variants. In M. I. Jordan, editor, *Learning in Graphical Models*, pages 355–368. MIT Press, 1998.
- [8] A. W. Senior. Tracking with probabilistic appearance models. In *ECCV workshop on Performance Evaluation of Tracking and Surveillance Systems*, pages 48–55, 2002.
- [9] Y. Song, L. Goncalves, and P. Perona. Monocular perception of biological motion - clutter and partial occlusion. In *Proc. 6th European Conf. Computer Vision, Dublin, Ireland*, volume 2, pages 719–733, 2000.
- [10] J. Sullivan, A. Blake, and J. Rittscher. Statistical foreground modelling for object localisation. In *Proc. 6th European Conf. Computer Vision, Dublin, Ireland*, volume 2, pages 307–323, 2000.
- [11] P. H. Tu and J. Rittscher. Crowd segmentation through emergent labeling. In *Statistical Methods in Video Processing: ECCV 2004 Workshop SMVP2004*, pages 187–198, May 2004.
- [12] Z. Tu and A. Yuille. Shape matching and recognition - using generative models and informative features. In *Proc. 8th European Conf. Computer Vision, Prague, Czech Republic*, volume 3, pages 195–209, 2004.
- [13] T. Zhao and R. R. Nevatia. Bayesian human segmentation in crowded situations. In *IEEE Computer Vision and Pattern Recognition, Madison, Wisconsin*, volume 2, pages 459–466, 2003.
- [14] T. Zhao and R. R. Nevatia. Tracking multiple humans in complex situations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(9):1208–1221, September 2004.