An Adaptive Appearance Model Approach for Model-based Articulated Object Tracking

Alexandru O. Bălan* Michael J. Black
Department of Computer Science - Brown University
Providence, RI 02912, USA
{alb, black}@cs.brown.edu

Abstract

The detection and tracking of three-dimensional human body models has progressed rapidly but successful approaches typically rely on accurate foreground silhouettes obtained using background segmentation. There are many practical applications where such information is imprecise. Here we develop a new image likelihood function based on the visual appearance of the subject being tracked. We propose a robust, adaptive, appearance model based on the Wandering-Stable-Lost framework extended to the case of articulated body parts. The method models appearance using a mixture model that includes an adaptive template, frame-to-frame matching and an outlier process. We employ an annealed particle filtering algorithm for inference and take advantage of the 3D body model to predict self-occlusion and improve pose estimation accuracy. Quantitative tracking results are presented for a walking sequence with a 180 degree turn, captured with four synchronized and calibrated cameras and containing significant appearance changes and self-occlusion in each view.

1. Introduction

The detection and tracking of three-dimensional human body models (using one or more images) has progressed rapidly but successful approaches typically rely on accurate foreground silhouettes obtained using background subtraction. In many practical applications such information is not reliable due to the presence of complicated (cluttered and moving) backgrounds that change in unpredictable ways. This suggests that a reliable human tracker cannot rely on high-quality silhouettes obtained from background subtraction. Instead we need reliable models of image appearance that enable 3D articulated human tracking without accurate background subtraction.

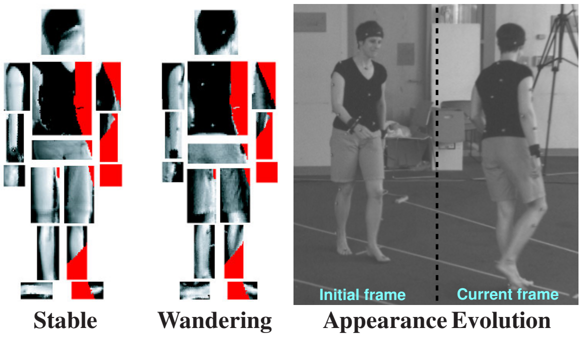

Figure 1. **Exploded person: *RoAM* body model.** The appearance of the body in a camera view is represented by the appearance of the individual parts. Stable and wandering appearance components extracted by the *RoAM* tracker are shown for one frame. The appearance models are initialized based on the first frame and updated over time using a method that takes into account self occlusion (red regions in the appearance models indicate occluded parts in the current frame).

Here we develop a new image likelihood function based on the visual appearance of the subject being tracked. Of course, this appearance is not known *a priori* and it changes over time with viewpoint, clothing deformation, self shadowing, etc. To deal with these complexities we propose a robust, adaptive, appearance model, *RoAM*, based on the Wandering-Stable-Lost ($\mathcal{WSL}$) framework [7] extended to the case of articulated body parts. The method models appearance using a mixture model that includes an adaptive template, frame-to-frame matching and an outlier process. While the $\mathcal{WSL}$ method has been shown to be very reliable for tracking affine or planar image patches with sufficient texture, its application to 3D human tracking presents novel challenges.

In particular, we use a collective set of appearance models for each body part. Self-occlusion however presents a problem in that, in some views, certain body parts may be occluded for long periods of time. In the standard $\mathcal{WSL}$ framework the stable appearance template would adapt to model the appearance of the occluding regions and would

[*] Portions of this work were performed by AOB at Intel Research in Santa Clara, CA.

result in tracking failure. To cope with this we take advantage of the 3D body model to predict self-occlusion and restrict adaptation to visible pixels.

The original $\mathcal{WSL}$ model was developed in the context of parametric motion estimation/tracking. This is possible in the case of low-dimensional linear models (e.g. affine) and small frame-to-frame displacements. Tracking the human body however is more complex and most successful approaches rely on some form of stochastic sampling. For example, we employ an annealed particle filtering algorithm for inference [5]. Our solution extends $\mathcal{WSL}$ to a particle-based tracker by allowing separate models for each particle; making this practical is a challenge.

Quantitative tracking results are presented for a walking sequence with a 180 degree turn, captured with four synchronized and calibrated cameras and containing significant appearance changes and self-occlusion in each view. We combine a standard background subtraction likelihood [5] with the $\mathcal{WSL}$ model in a principled way and then perform experiments where the background data is periodically uninformative. We compare a traditional tracker based on background subtraction to one that includes an adaptive appearance model and find that the $\mathcal{WSL}$ model well improves the stability and accuracy of the tracker.

1.1. Related Work

A variety of image likelihood models have been employed for 2D and 3D human tracking including foreground silhouettes [5, 16], edges [5, 9], brightness constancy [3, 8, 15], optical flow and flow discontinuities [16], image templates [4], color distributions [10, 11], and image filter statistics [13]. In addition there is an enormous literature on image-based tracking. While the recent $\mathcal{WSL}$ model has not been applied to 3D human tracking, previous methods have used pieces of it. We use the Wandering-Stable-Lost idea to categorize previous work.

Wandering. The wandering component corresponds to a brightness constancy assumption between adjacent frames. It implies that the image appearance of the limb at the current frame looks like the image appearance at the previous frame – only the pose has changed. Such a model was first proposed by Ju *et al.* [8] for 2D tracking with a "cardboard-person" model. Bregler and Malik [3] and Sidenbladh *et al.* [15] use the same idea for 3D tracking. Like any optical flow tracker, these methods tend to drift over time and once they do, they have no way of recovering.

Stable. Stable models have appeared in a variety of forms. In particular we can categorize them as adaptive or fixed.

Stable & Fixed. Cham and Rehg [4], for example, used a fixed image template extracted by hand in the first frame of a sequence to track a moving subject. Such an approach is unable to deal with significant changes in viewpoint. To cope with changing and more varied appearance, Sidenbladh *et al.* [14] extended the idea of EigenTracking [2] to cylindrical limbs and learned an eigen-appearance model from multiple training views.

Stable & Adaptive. Howe *et al.* extended [8] by constructing a template for each limb that was the weighted average of several preceding frames. They also accounted for self occlusions and used known occlusion relations to construct support maps. Our $RoAM$ framework extends both of these ideas and provides a more principled way of adapting the templates and combining them with 2-frame tracking. Roberts *et al.* [11] represented the body as a collection of simple geometric primitives and modeled texture regions on the surfaces using color histogram distributions that they adapt over time. In contrast, our model is parametric, view dependent and non-regional. Numerous authors have looked recently at tracking using adaptive appearance models, often based on eigen-representation [12]. Unlike the $RoAM$ framework, these methods do not typically provide a framework for combining different sources of information in a robust probabilistic way.

2. Method

We work in a Bayesian framework using a generative approach. The goal is to estimate the *posterior* probability distribution $\mathcal{P}_t^+ \equiv p(\mathbf{x}_t|\mathbf{y}_{1:t})$ for the state $\mathbf{x}_t$ of the human body at time t given a sequence of image observations $\mathbf{y}_{1:t} \equiv (\mathbf{y}_1, \ldots, \mathbf{y}_t)$.

Making the common assumptions that the state at time t is only dependent on the previous state while the observation is only dependent on the current state, a recursive Bayes formula can be derived and used for inference:

$$p(\mathbf{x}_t|\mathbf{y}_{1:t}) \propto p(\mathbf{y}_t|\mathbf{x}_t) \int p(\mathbf{x}_t|\mathbf{x}_{t-1}) p(\mathbf{x}_{t-1}|\mathbf{y}_{1:t-1}) d\mathbf{x}_{t-1}.$$

The human tracking problem has been shown to be highly under-constrained and ambiguous [6]. The multi-modality of the posterior distribution has pushed the state of the art towards non-parametric approximate methods that represent distributions by a set of N random samples or particles with associated normalized weights $\{\mathbf{x}_t^{(i)}, \pi_t^{(i)}\}_{i=1}^N$. Particles are sampled from the posterior at time $t-1$, propagated over time using the temporal dynamics $p(\mathbf{x}_t|\mathbf{x}_{t-1})$ and assigned new weights according to the likelihood function $p(\mathbf{y}_t|\mathbf{x}_t)$ [6].

2.1. Appearance Model

The objective of the *Wandering-Stable-Lost* model ($\mathcal{WSL}$) [7] is to track an image region while adapting to slowly changing appearance and maintaining robustness to partial occlusions, natural appearance changes and image deformations. By maintaining a natural measure of the stability of the observed image structure, appearance properties for motion estimation can be weighted according to the corresponding level of stability. Our treatment closely follows that in [7].

The $\mathcal{WSL}$ framework assumes that each dimension of the appearance model is independent of the other dimensions; this simplifies the mathematical formulation greatly as the appearance model is defined in a 1D space. When tracking using a multi-dimensional appearance model, each 1D $\mathcal{WSL}$ component casts a weighted vote according to its level of stability.

Consider a single real-valued data observation d_t at each time t. The method models appearance using a mixture model that includes an adaptive template (stable component $\mathcal{S}$), frame-to-frame matching (wandering component $\mathcal{W}$) and an outlier process (lost component $\mathcal{L}$).

The mixture probability density for a new data observation d_t conditioned on the past observations is given by

$$p(d_t|\mu_{s,t-1}, \sigma^2_{s,t-1}, d_{t-1}, \mathbf{m_{t-1}}) = m_{w,t-1}p_w(d_t|d_{t-1}) + \\ + m_{s,t-1}p_s(d_t|\mu_{s,t-1}, \sigma^2_{s,t-1}) + m_{l,t-1}p_l(d_t).$$

The stable component $p_s(d_t|\mu_{s,t-1}, \sigma^2_{s,t-1})$ is intended to identify appearance properties that remain relatively stable over long time frames. It is modeled as a Gaussian density function with parameters $\mu_{s,t-1}, \sigma^2_{s,t-1}$ which adapt slowly over time. The wandering component $p_w(d_t|d_{t-1}; \sigma^2_w)$ supports frame-to-frame tracking when the appearance changes rapidly compared with the slow adaptation of the stable component, or for initialization when there is no history with which to identify stable properties. This term is modeled as a Gaussian density function relative to the previous data observation, and its variance σ^2_w is fixed. An outlier process accounting for tracking failures, occlusion and noise is given by the lost component $p_l(d_t)$, which is modeled with a uniform distribution. The mixing probabilities of the three components are given by $\mathbf{m_t} = (m_{w,t}, m_{s,t}, m_{l,t})$.

The parameters $\mu_{s,t}, \sigma^2_{s,t}$ of the appearance model and mixing probabilities $\mathbf{m_t} = (m_{w,t}, m_{s,t}, m_{l,t})$ are updated over time using an online Expectation-Maximization (EM) algorithm. During the E-step, the ownership probabilities of each component $i \in \{w, s, l\}$ for each observation d_t are computed by

$$o_{i,t}(d_t) = \frac{m_{i,t-1}p_i(d_t|\mu_{s,t-1}, \sigma^2_{s,t-1}, d_{t-1})}{p(d_t|\mu_{s,t-1}, \sigma^2_{s,t-1}, d_{t-1}, \mathbf{m_{t-1}})}. \quad (1)$$

The impact of previous observations on predicting current observations is assumed to fall exponentially with the time difference between them, making the recent past more relevant than the distant past. An exponential envelope is used to define the weight $S_t(k) = \alpha e^{-(t-k)/\tau}$ of an observation at time k with respect to an observation at time t. Here α is a normalizing constant to make the weights integrate to 1, $\tau = n_s/\log(2)$, and n_s is the half-life of the exponential envelope.

During the M-step, the maximum likelihood estimates of the mean and variance of the stable component are computed using the moments of the past observations, weighted by the stable ownership probability. If $M^{(j)}_t$ is the jth-order data moment weighted by the current ownership probability

$$M^{(j)}_t = \sum_{k=t}^{-\infty} S_t(k) d^j_k o_{s,t}(d_t), \quad (2)$$

then the stable mean and variance can be updated using the standard formulas

$$\mu_{s,t} = \frac{M^{(1)}_t}{M^{(0)}_t}, \quad \sigma^2_{s,t} = \frac{M^{(2)}_t}{M^{(0)}_t} - \mu^2_{s,t}. \quad (3)$$

A recursive expression for the moments can be derived from (2) which allows for the moments to be updated without the need to retain past information:

$$M^{(j)}_t = \alpha d^j_t o_{s,t}(d_t) + (1 - \alpha) M^{(j)}_{t-1}. \quad (4)$$

Here α acts as an adaptation factor. The higher the α value, the faster the model adapts to the new observations. Similarly, the mixing probabilities are updated using

$$m_{i,t} = \alpha o_{i,t}(d_t) + (1 - \alpha) m_{i,t-1}, \quad i \in \{w, s, l\}. \quad (5)$$

For a complete derivation and justification, please refer to [7].

2.2. Body Model

The skeleton of the body is represented as a kinematic tree having tapered cylinders with elliptical cross-sections around the limbs (Figure 2 Left). We consider 15 body parts: pelvis area, torso, upper and lower arms and legs, hands, feet and the head. A given pose configuration is given by the relative joint angles between connected limbs and the position and orientation of the pelvis in the global coordinate system, for a total of 40 dimensions. The intrinsic parameters such as the length and width of the limbs are provided and we do not optimize over them.

2.3. Likelihood Formulation

The success of any tracking method depends very much on the ability of the image likelihood function to discriminate between poses that fit the image well and those that do not. Ideally a wealth of image cues would be used to make this evaluation. While foreground silhouettes have been shown [1, 5] to be a powerful feature to use, they may not always be available. For these cases object appearance may prove a useful component of the likelihood.

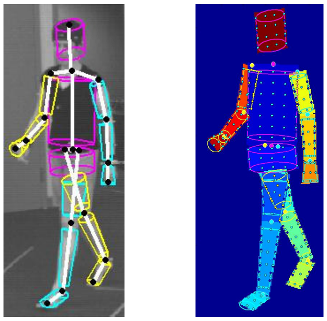

Figure 2. **Left: Body Model.** Kinematic tree with fifteen body parts consisting of 40 degrees of freedom. Six degrees of freedom are given to the pelvis consisting of global positioning and orientation. Torso, shoulders, hips and neck have all 3 freedoms, while the elbows and knees have 2 degrees of freedom. Wrists and ankles have 1 degree of freedom. **Right: Appearance Extraction.** A regular rectangular grid is used to extract pixel values for each body part in each camera view. Usually a dense grid is used that covers every pixel.

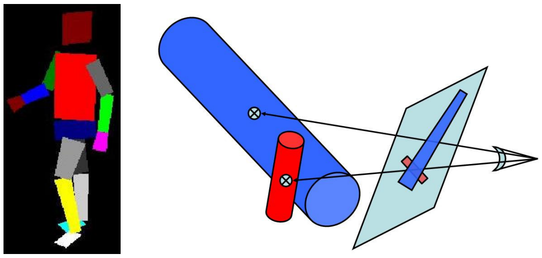

Figure 3. **Visibility Map. Left:** A label is assigned to each pixel in the map, denoting the body part visible at that location or the background. **Right:** Even if the red cylinder is occluded by the blue one, it is still possible however that its center is closer to the camera eye if the two cylinders are disproportionate.

We introduce a new image likelihood function for articulated objects based on the visual appearance of the subject being tracked. We use a robust, adaptive, appearance model *RoAM* based on the Wandering-Stable-Lost framework, extended to the case of articulated body parts. The likelihood function $p(\mathbf{y}_t|\mathbf{x}_t)$ computes a measure of how well a pose hypothesis $\mathbf{x}_t$ fits the image observations. By projecting the body model into the images, we can uniformly extract pixel information for each body part as shown in Figure 2 Right. We assign a 1D $\mathcal{WSL}$ model to each pixel on each limb of interest on each camera view.

In fact, for a given pixel we can assign multiple 1D $\mathcal{WSL}$ models corresponding to different image filter responses. A wide variety of image properties can be used for learning an appearance model including: image brightness, steerable derivative filters, image gradients, color components in various color spaces, wavelet phases, image statistics, filter pyramids at different scales, etc.

The *RoAM* appearance model $\mathcal{A}_t$ at time t consists of all 1D $\mathcal{WSL}$ model parameters indexed by r

$$\mathcal{A}_t = \{(\mu_{s,t}, \sigma_{s,t}, d_{t-1}, \mathbf{m}_t, M_t^{(0)}, M_t^{(1)}, M_t^{(2)})_r\}, \qquad (6)$$

assigned to each pixel in a grid belonging to each body part in each view for every type of image filter response.

One advantage for using a kinematic tree model is the ability to determine, for a body configuration, which regions of the limbs are not visible in each view due to self-occlusion. A ray-tracer approach can generate the desired visibility map shown in Figure 3 by computing the depth of the visible surface at each pixel. This can be avoided however and a faster procedure can be employed making the observation that, since the cylinders are convex, there is a guarantee that no two cylinders can occlude each other. This means that there always exists a topological order of the limbs such that rendering the cylinders in this order yields the correct visibility map. We try to obtain such a topological order by sorting the body parts in decreasing distance from the camera. The distance is taken to be between the camera eye and the center of the limb. Figure 3 shows an example were this heuristic could fail, but such failures are unlikely in the case of proportionate body parts.

The goal is to use stable properties of the body appearance to align coherent structures over extended durations, identified by high stable ownership probability. We rely on a good history of stable observations (o_s) to make predictions and we expect the current observation to be consistent with the stable component (p_s). This suggests an aggregate energy function based on the log likelihood of the stable components weighted by the stable ownership probability:

$$E_s(\mathbf{d}_t(\mathbf{x}_t)) = \sum_v o_s(d_{v,t}) \log p_s(d_{v,t}|\mu_{v,s,t-1}, \sigma^2_{v,s,t-1}).$$

Here v indexes the $\mathcal{WSL}$ models corresponding to pixels on visible body parts during both the current and previous frames, and $\mathbf{d}_t(\mathbf{x}_t)$ is the entire set of image observations induced by pose $\mathbf{x}_t$.

Sometimes there are not enough stable components to reliably estimate the matching between a body pose and the learned appearance model. This is true during initialization when there is no history of stable structure, or during rapid changes of appearance. In this case the tracker should gracefully degrade to frame-to-frame matching. We therefore need to incorporate $\mathcal{W}$ constraints into the log likelihood estimation. We use a similar energy function corresponding to the wandering component for visible pixels:

$$E_w(\mathbf{d}_t(\mathbf{x}_t)) = \sum_v o_w(d_{v,t}) \log p_w(d_{v,t}|d_{v,t-1}). \qquad (7)$$

The E_s and E_w energy functions can be combined into an objective function which we seek to maximize. We define the log likelihood of the image observations conditioned on a given pose as

$$\log p_{_{RoAM}}(\mathbf{y}_t|\mathbf{x}_t) \propto \frac{1}{|\{v\}|}(E_s + \epsilon E_w), \qquad (8)$$

where ϵ is a sub-unitary scaling factor to favor stable structure over transient structure.

2.4. Foreground Silhouettes Likelihood

Our results will show that likelihood functions using foreground silhouettes generate better tracking results than our appearance likelihood when the silhouettes are accurate enough. Indeed this was expected; when good edge and silhouette data is available it should be used and has been shown to yield reliable tracking [1, 5]. Our intent is to provide an alternative to foreground silhouettes when they are unreliable, and to supplement them when they are. Foreground silhouettes can be obtained by employing a simple background subtraction method or any other foreground segmentation algorithm. They are represented as binary maps F, with 1 denoting foreground and 0 background.

The negative log-likelihood of a pose is estimated by taking a number of points uniformly distributed inside the cylinders of the model, projecting them into each image view, and computing the mean square error (MSE) of the foreground map responses:

$$-\log p_{_{FG}}(\mathbf{y}_t|\mathbf{x}_t) \propto \frac{1}{|\{\xi\}|}\sum_{\xi}(1-F(\xi))^2, \qquad (9)$$

where ξ denotes points on the grid.

We combine the appearance likelihood in (8) with a foreground silhouette likelihood in (9) using a weighted formulation

$$\begin{aligned}\log p(\mathbf{y}_t|\mathbf{x}_t) \quad &\propto \quad \lambda \log p_{_{RoAM}}(\mathbf{y}_t|\mathbf{x}_t) + \\ &+ \quad (1-\lambda)\log p_{_{FG}}(\mathbf{y}_t|\mathbf{x}_t), \qquad (10)\end{aligned}$$

where λ is a parameter to be determined empirically.

2.5. Inference

Inference is achieve using an annealed particle filter [5]. This approach searches for peaks in the posterior distribution using simulated annealing, and tends to concentrate the particles into one mode. It has been shown to be very precise when multiple cameras are considered and a likelihood based on foreground silhouettes is used [1]. Its major drawback is it will often fail to represent multiple hypotheses.

The annealing process consists of several iterations (layers) of the filtering procedure which assumes the posterior distribution is represented by weighted particles. During the filtering procedure, particles are sampled with replacement from the posterior, and temporal-dynamics are used to propagate the particle from one frame to the next. We use the simplest model to make predictions from the posterior, which assumes a Gaussian distribution around the previous time estimates: $\mathbf{x}_t \sim \mathcal{N}(\mathbf{x}_{t-1}, \Sigma)$, where Σ is a covariance matrix which can be learned from training examples. To reduce the search space, we apply a hard prior that eliminates any particle that corresponds to implausible body poses such as having angles exceeding anatomical joint limits or inter-penetrating limbs. Predicted particles are then re-weighted according to an annealed version of the likelihood function. The likelihood distribution is intended to be very broad in the first layer and gradually become more peaked. This is achieved by exponentiating the likelihood responses from (10) according to an annealing schedule.

The weighted particles obtained during the last layer of annealing represent the estimated posterior at the next frame. To extend $\mathcal{WSL}$ to the particle filtering framework, each particle is connected through a genealogical chain to one particle in the previous posterior. The *RoAM* appearance of each particle can then be updated based on the appearance model of the corresponding particle at previous time instance using (3), (4) and (5). In this way every particle carries its own $\mathcal{WSL}$ model through time.

3. Experiments and Results

We consider a test sequence of grey-value images from the Brown database [1] consisting of walking motion with a 180° turn captured from 4 calibrated cameras. The background is stationary, but cluttered; shadows are also present. The person's appearance changes significantly due to the full rotation of the body and also due to temporary self-occlusions. Often there is not enough contrast to differentiate the limbs from the background. All these make the sequence challenging.

Ground truth data obtained with a marker-based motion capture system is available for this sequence. We quantitatively evaluate our results with respect to ground truth data using a measure based on 3D Euclidian average distances at joint locations. We define the frame error as the minimum error of any particle in the posterior distribution or for the expected pose [1]. For an entire sequence we report the average frame error; errors above 200mm are considered failures.

Tracking with and without *RoAM*. In our first experiment, we have performed full body tracking using both the $RoAM$ appearance model and the foreground silhouettes. By adjusting the λ parameter in (10), we are able to compare the two likelihoods separately, and evaluate how they work together. Tracking results as a function of λ are plotted in Figure 4 (thick curve).

We first observe that, due to the small image size of the limbs and the lack of image texture, tracking using only the appearance model is significantly worse than using only foreground silhouettes. However, mixing $RoAM$ appearance with the silhouettes in a 20-to-80 ratio increases tracking accuracy with respect to the silhouette likelihood. Figure 7 illustrates that tracking with only silhouettes permits the legs to switch starting with frame 120, while the arms sometimes "stick" to the torso to fit inside the silhouette. In

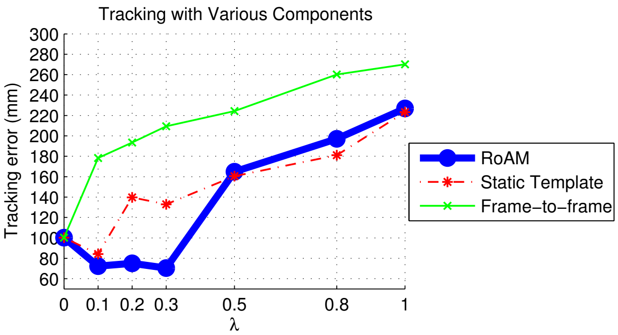

Figure 4. **Appearance vs. Silhouettes.** Tracking using foreground silhouettes corresponds to $\lambda = 0$, while $\lambda = 1$ gives full contribution to the appearance model. The best performance for the $RoAM$ model (thick curve) is obtained when the appearance gives about 20% to the likelihood estimation and silhouettes 80%. Performance for individual components of the model (static template – fixed $\mathcal{S}$, frame-to-frame – $\mathcal{W}$) is displayed the same way.

contrast, Figure 8 shows that the combined likelihood function is able to reasonably address these problems.

We have also looked at how individual pieces of the model affected performance. We replaced the $\mathcal{WSL}$ model with individual components: a static template appearance model (fixed $\mathcal{S}$) and a frame-to-frame tracker ($\mathcal{W}$). The results in Figure 4 suggest that, as the silhouette contribution to the likelihood is reduced, the frame-to-frame tracker drifts more easily into the background and effectively becomes a penalty for the true pose. On the other hand, the static template is not subject to drifting and copes very well on our short sequence (probably because each body part has the same texture on all sides), nonetheless it would not generalize to longer sequences due to its inability to adapt to drastic appearance changes.

Failures of background subtraction. To simulate drastic failures of background subtraction, we artificially made the silhouette likelihood uniform for 4 frames out of every 5. In these situations the $RoAM$ appearance model reduces tracking degradation due to poor foreground segmentation significantly as shown in Figure 5 Left.

Tracking individual parts. To better understand how the appearance of each body part is represented we have tracked individual body parts and looked at the behavior of the $\mathcal{WSL}$ models. First we tracked only the head in 3D using all 4 views. The head has more features and texture than the arms and legs and can be tracked easily. We show in Figure 6 how the stable and wandering components evolve over time. We note that when the head is not tracked in the conjunction with the torso, the orientation is not retained very well. The stable component in the top row eventually replaces the face skin with hair, while the frontal hair has high stability even when the head turns since the image intensity is the same. The wandering component really shows the estimated pose at the previous frame.

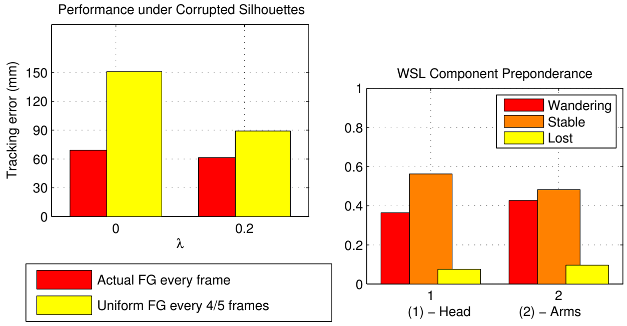

Figure 5. **Left: Corrupted Silhouettes.** The foreground silhouette likelihood is artificially made uniform every 4 frames out of 5 in the sequence. Over 100 frames, the error for using just silhouettes ($\lambda = 0$) is 2.5 times larger, while the error for the $RoAM$ plus silhouette model ($\lambda = 0.2$) is increased by only 50%. **Right: $\mathcal{WSL}$ Component Preponderance.** The mixing probabilities $\mathbf{m_t}$ for the head and lower arms are averaged individually over all $\mathcal{WSL}$ models at every frame. The head has more stable structure in regions with hair, while the arms get lost more easily and wander in the background.

In contrast, it has been noted that lower arms are one of the hardest body parts to track. In comparison with legs and upper arms, they are smaller, they move faster, and their motion is harder to model [1]. To track only the lower arm movement, we have localized the torso with the true position and orientation at each frame. We have averaged over all $\mathcal{WSL}$ models at every frame the mixing probabilities $\mathbf{m_t}$ for the head and lower arms to obtain the bar plot in Figure 5 Right. The head is quite stable since the hair does not change appearance even when the head rotates. The fact that the arms are lost more often explains why its stable component has lower mixing proportion. Whenever an arm is mistracked over a relatively uniform background, the wandering ownership can remain high.

3.1. Limitations

We have observed a number of situations that has made appearance tracking difficult. We currently employ no motion prior to guide the tracker when image observations are ambiguous. Multiple cameras are needed to cope with self-occlusions. Even so, tracking can fail when the stable appearance of the limb is very similar to the background; this is particularly true in low contrast regions. In these situations, as in any tracker, the model is prone to track the background.

Finally, in this sequence there is little texture information on the subject's limbs and the poor contrast between foreground and background permits the appearance model to slide off. We conclude that appearance-based tracking alone requires higher quality image data and higher resolution images. Color image data would provide additional information.

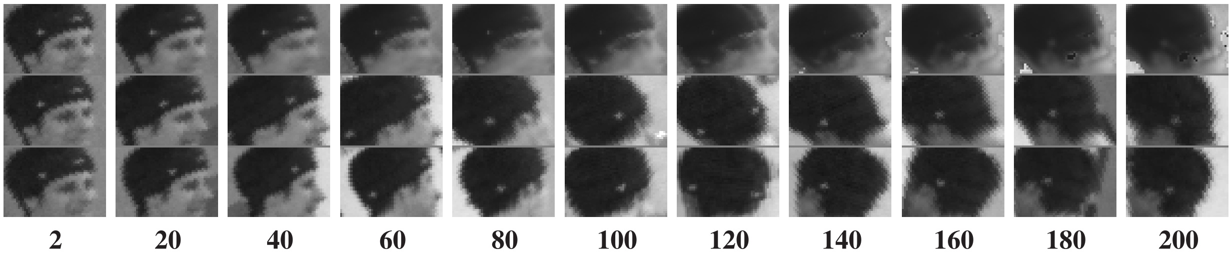

Figure 6. **Head Tracking WS Components.** Every 20^{th} frame from camera 3. When every pixel has associated a $\mathcal{WSL}$ model, the stable and wandering components can be visualized in an intuitive way as they evolve over time. The top row shows the mean of the stable component, while the middle row shows the mean of the wandering component. The input image is shown on the bottom row.

3.2. Implementation Details

In our experiments we used image brightness as the observation signal, which has been normalized between zero and one. Inference was performed using 250 particles and 5 layers of annealing. We initialize the tracker with the true pose in the first frame and the appearance model initializes based on that pose.

We will now go over the essential parameters for $\mathcal{WSL}$. The $\mathcal{WSL}$ mixing proportions are set to $\mathbf{m_1} = (0.35, 0.55, 0.15)$; the algorithm is not sensitive to the exact choice of mixing values. The moments $M_1^{(j)}$ are initialized such that $\mu_{s,1}$ equals the initial data observation d_1 and $\sigma_{s,1} = 0.075$. We also use this value as a lower threshold on the variance of $\mathcal{S}$. We make $\sigma_w = 1.5\sigma_{s,1}$ and downweigh the E_w term in (8) by $\epsilon = 1/20$ to give preference to stable structure. The adaptiveness of the stable component is influenced by the half-life of the exponential decay, set to $n_s = 30$ frames.

Occasional restarts of the appearance model are necessary when observations are persistently unstable. These situations can be detected when the mixing probability of the stable component m_s falls under a specified threshold (0.1 in our experiments). We note that this is done per pixel and therefore no useful information is lost in stable regions. Restarts are done by setting the model parameters to the initial values and centering the stable component at the current data observation.

The $\mathcal{WSL}$ model did not consider the case of missing data. Attempting to read an image outside of its boundaries or knowing that the observed data is incorrect (the occlusion case) are pertinent examples. We chose to freeze the appearance model in this case. Alternatives include restarting the model or assigning full ownership to the lost component, since the wandering and stable components cannot be responsible for generating an imprecise observation.

The size of the overall appearance model can be very large since the number of 1D $\mathcal{WSL}$ models in $\mathcal{A}_t$ equals the number of grid pixels on a limb times the number of limbs used times the number of camera views times the number of filter responses. In our experiments we kept only one appearance model for the entire posterior at the previous time based on the appearance of the expected pose in the particle set.

There is a trade-off between frame rate and accuracy, but for out setup the processing time is about 6 minutes per frame using a Matlab implementation on a standard PC.

4. Conclusions and Future Work

To address the problem of potentially inaccurate background segmentations, we have proposed a robust, adaptive, appearance model, *RoAM*, based on the Wandering-Stable-Lost ($\mathcal{WSL}$) framework. We have extended $\mathcal{WSL}$ to the case of articulated body parts and to a particle filter based tracking framework. We have demonstrated the approach on a challenging sequence and successfully enhanced the performance of a silhouette tracker by augmenting it with an adaptive appearance model. In particular, the results suggest that when background subtraction is unreliable, an adaptive appearance model for the limbs stabilizes the tracking results substantially.

Our work suggests a number of future directions. Here we have only explored appearance based on image brightness which may be sensitive to lighting changes. We have proposed other possible features but determining which image features are most appropriate requires more research. We have used no dynamics in our tracker and we expect a prior model of human motion to likely improve robustness further.

Acknowledgments: We thank L. Sigal, H. Haussecker and T. P.-C. Chen for helpful discussions. This work was partially supported by Intel Research and by NSF grant IIS-0535075.

References

[1] A. Balan, L. Sigal and M. Black. A Quantitative Evaluation of Video-based 3D Person Tracking. *IEEE Workshop on VS-PETS*, pp. 349–356, 2005.

[2] M. Black and A. Jepson. EigenTracking: Robust Matching and Tracking of Articulated Objects Using a View-Based Representation. *ECCV* , (1):329–342, 1996.

[3] C. Bregler and J. Malik. Tracking People with Twists and Exponential Maps. *CVPR*, pp. 8–15, 1998.

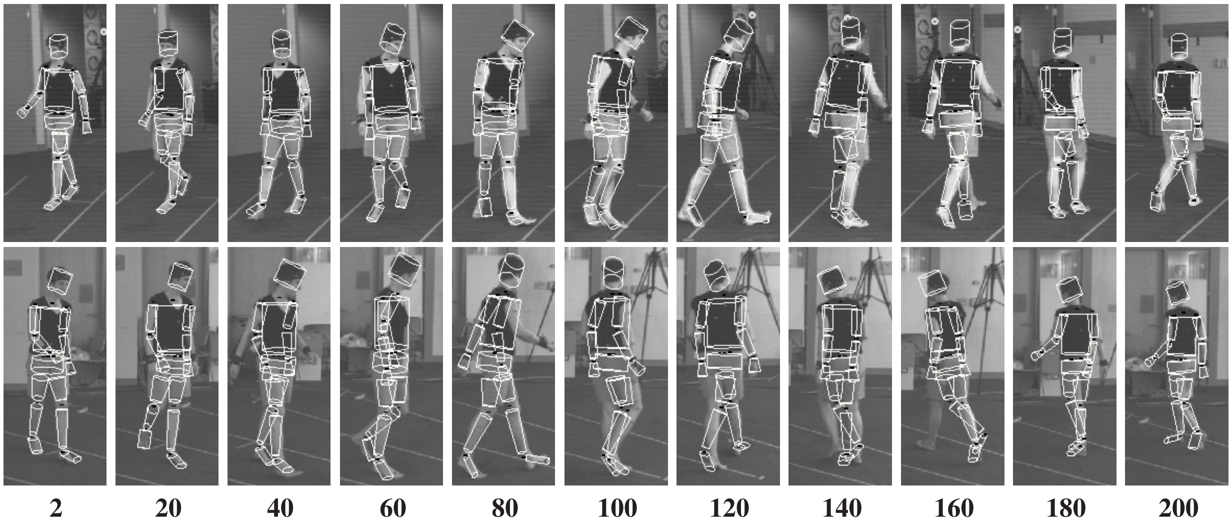

Figure 7. **Tracking Results using Foreground Silhouettes Likelihood.** Only two of the views shown out of 4. Only foreground silhouettes are used to perform tracking. The legs swap identities and the arms are often stuck next to the torso.

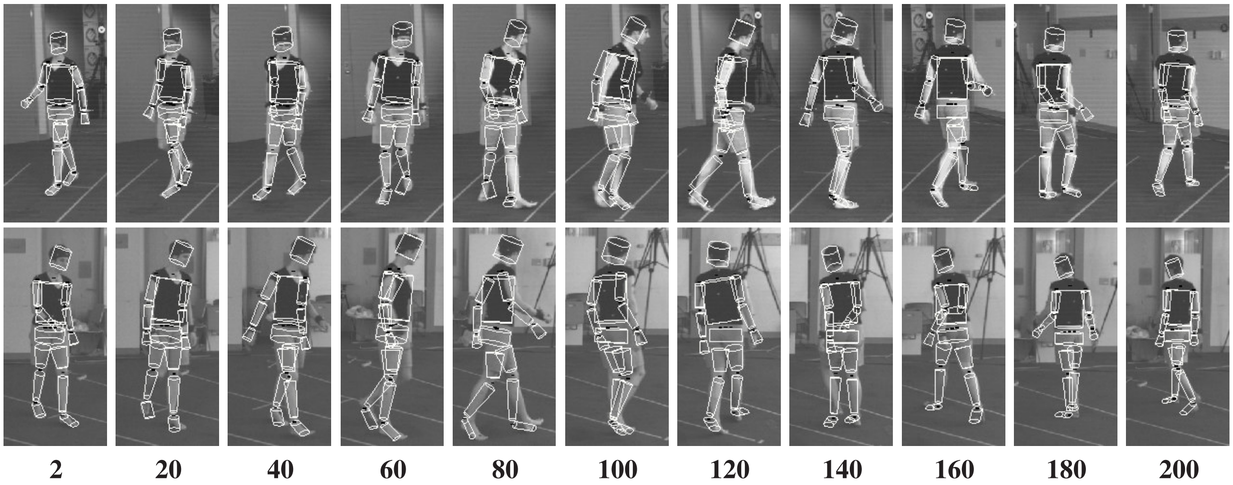

Figure 8. **Tracking Results using A Mixture Likelihood.** Appearance and foreground silhouettes are combined into a mixture likelihood with $\lambda = 0.2$. It disambiguates the left leg from the right leg and follows the arms more often.

[4] T.-J. Cham and J. Rehg. A Multiple Hypothesis Approach to Figure Tracking. *CVPR*, (1):239–245, 1999.

[5] J. Deutscher and I. Reid. Articulated Body Motion Capture by Stochastic Search. *IJCV*, 61(2):185–205, 2004.

[6] M. Isard and A. Blake. Condensation – Conditional Density Propagation for Visual Tracking. *IJCV*, 29(1):5–28, 1998.

[7] A. Jepson, D. Fleet, and T. El-Maraghi. Robust Online Appearance Models for Visual Tracking. *PAMI*, 25(10):1296–1311, 2003.

[8] S. Ju, M. Black, and Y. Yacoob. Cardboard people: A Parameterized Model of Articulated Motion. *International Conference on Automatic Face and Gesture Recognition*, pp. 38–44, 1996.

[9] E. Poon and D. Fleet. Hybrid Monte Carlo Filtering: Edge-based People Tracking. *IEEE Workshop on Motion and Video Computing*, pp. 151–158, 2002.

[10] D. Ramanan, D. Forsyth, and A. Zisserman. Strike a Pose: Tracking People by Finding Stylized Poses. *CVPR* , (1):271–278, 2005.

[11] T. Roberts, S. McKenna, and I. Ricketts. Adaptive Learning of Statistical Appearance Models for 3D Human Tracking. *BMVC*, pp. 333–342, 2002.

[12] D. Ross, J. Lim, and M.-H. Yang. Adaptive Probabilistic Visual Tracking with Incremental Subspace Update. *ECCV*, (2):470-482, 2004.

[13] H. Sidenbladh and M. Black. Learning Image Statistics for Bayesian Tracking. *ICCV* , (2):709–716, 2001.

[14] H. Sidenbladh, F. De la Torre, and M. Black. A Framework for Modeling the Appearance of 3D Articulated Figures. *IEEE Int. Conf. on Automatic Face and Gesture Recognition*, pp. 368–375, 2000.

[15] H. Sidenbladh, M. Black, and D. Fleet. Stochastic Tracking of 3D Human Figures Using 2D Image Motion. *ECCV*, (2):702–718, 2000.

[16] C. Sminchisescu and B. Triggs. Estimating Articulated Human Motion with Covariance Scaled Sampling. *IJRR*, 22(6):371–391, 2003.