A New Multi-View Articulated Human Motion Tracking Algorithm With Improved Silhouette Extraction and View Adaptive Fusion

Zhong Liu[1,2], K. T. Ng[1], S. C. Chan[1], Xiao-Wei Song[2]
[1] Department of Electrical and Electronic Engineering, The University of Hong Kong, Hong Kong, P. R. China
[2]Department of Electric and Information Engineering, Zhongyuan University of Technology, Zhengzhou, Henan, P. R. China
Email: earlybird86@yahoo.cn, {ktng, scchan}@eee.hku.hk, sxw-tju@163.com

***Abstract*—This paper proposes a new articulated human motion tracking and pose estimation algorithm using an improved silhouette extraction method with view adaptive fusion. It is developed around the baseline algorithm in HumanEva, which uses the Annealed Particle Filter (APF). Shadow detection and removal and a level-set method are employed to achieve better silhouette extraction. An adaptive view fusion approach is also proposed to improve the matching between the human 3D model and the observations. Experimental results show that the proposed approach has considerably better performance than the baseline algorithm in the HumanEva dataset, due to better shadow handling and data fusion of multiple views.**

I. INTRODUCTION

Motion capture of realistic humans has broad applications in different areas such as human-computer interaction, analysis for human body in sports and computer animation. However, it is a challenging problem in computer vision as tracking of articulated human body is difficult due to the high dimensionality, occlusion, clutter, loss of depth and the nature of the kinematic structure of the body. All these result in a non-Gaussian and multi-modal distribution [1, 2] during modeling and estimation. Generally, 3D motion tracking algorithms can be roughly classified into two categories: discriminative-based approach and generative-based approach [3]. The former approach estimates the body poses directly from training data. For the latter framework, an articulated 3D model is used for tracking. The model is projected onto an image plane and likelihood function is computed to evaluate the quality of the match. The use of human body models greatly simplify the pose estimation problem and improve its robustness and accuracy. Popular human body models used for tracking include the truncated cylinders used in HumanEva [4], more accurate models such as SCAPE [5], and the statistical model [6] that accurately models the surface deformation as a function of pose and the physique of the subject. Multi-view based approaches [7, 8] are also commonly used in generative-based 3D body trackers to reduce ambiguity that usually happens in single view videos, as more image measurements from multiple cameras can be utilized to obtain better tracking results. However, imperfections such as shadow and noise in silhouettes can still degrade significantly the tracking performance.

In this paper, an improved silhouette extraction algorithm with shadow removal and level set-based silhouette refinement is proposed. Moreover, a view adaptive fusion strategy is proposed for multi-view articulated human tracking based on Annealed Particle Filtering (APF). The dataset of HumanEva [9] and the baseline algorithm are used for evaluation of our proposed algorithm. The dataset includes different subjects performing a continuous sequence of actions and the videos captured are synchronized with ground-truth 3D motion. Error metrics for computing 2D and 3D pose errors are also formulated. Our algorithm is developed around the baseline algorithm in HumanEva for 3D articulated tracking which uses a Bayesian framework with optimization in the form of Sequential Importance Resampling and APF. Experimental results show that the proposed approach has considerably better performance than the baseline algorithm in the HumanEva dataset, due to better shadow handling and data fusion of multiple views.

The organization of the paper is as follows: in Section II, the articulated human motion and pose estimation algorithm is briefly reviewed. The proposed foreground extraction and the view adaptive fusion method for multi-view tracking based on APF will be introduced. In Section III, experimental results are shown in order to verify the effectiveness of the proposed algorithms. Section IV concludes the paper.

II. PROPOSED TRACKING APPROACH

In [9, 10], the baseline algorithm is based on the Bayesian estimation framework. Given a sequence of image observation $\mathbf{y}_{1:t} \equiv (\mathbf{y}_1, \ldots, \mathbf{y}_t)$, the aim is to estimate the posterior probability density function $p(\mathbf{x}_t|\mathbf{y}_{1:t})$ for the state $\mathbf{x}_t$ of the human body at time instant t. For Markov processes, the state at time t is only dependent on the previous state and the observation is only dependent on the current state. Using the Bayes formula, one can recursively update the posterior probability as follows:

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

Since the multivariate integral is difficult to be evaluated analytically, Monte Carlo method such as particle filter is frequently employed in such filtering problems. The baseline algorithm also utilizes an articulated 3D model represented by truncated cylinders to track human motion. The human body model has two sets of parameters that describe the pose and shape of the body respectively. Moreover, a 3D kinematic tree with Euler angles is used to model the skeleton of human

This work was supported in parts by Hong Kong Research Grant Council (RGC), the Hong Kong Innovation and Technology Fund (ITF) and the National Natural Science Foundation of China (No.60902063).

body. We now describe the various system components in our proposed algorithm.

2.1 Annealed Particle Filter

Due to the good performance of the APF used in the baseline algorithm, it is also adopted in this work. It proceeds from layer *M* to layer 1 at each time instant, which updates the probability density over the state parameters. Each layer undergoes two steps:

1) Resample and propagate with a Gaussian diffusion model using the state density at layer $m+1$, which is represented by a set of particles with associated normalized weights $S_{t,m+1} \equiv \{\mathbf{x}_{t,m+1}^{(i)}, \pi_{t,m+1}^{(i)}\}_{i=1}^{N}$:

$$\{\mathbf{x}_{t,m}^{(i)}\}_{i=1}^{N} \sim \sum_{j=1}^{N} \pi_{t,m+1}^{(j)} \mathcal{N}\left(\mathbf{x}_{t,m+1}^{(j)}, \alpha^{M-m}\boldsymbol{\Sigma}\right), \quad (2)$$

where diffusion covariance matrix $\boldsymbol{\Sigma}$ controls the range of the search at each layer with a larger $\boldsymbol{\Sigma}$ indicating a wider sampling. From layer *M* to layer 1, $\boldsymbol{\Sigma}$ is scaled by a parameter α to drive the particles to the mode of the posterior distribution.

2) Weight each particle with an annealed function:

$$\pi_{t,m}^{(i)} = \frac{\mathrm{p}\left(\mathbf{y}_t \middle| \mathbf{x}_{t,m}^{(i)}\right)^{\beta^m}}{\sum_{j=1}^{N} \mathrm{p}\left(\mathbf{y}_t \middle| \mathbf{x}_{t,m}^{(j)}\right)^{\beta^m}}, i \in \{1, \ldots, N\}, \quad (3)$$

where β^m is the annealing rate optimized so that approximately half the particles get selected to the next layer. The resulting particle set $S_{t,m} \equiv \{\mathbf{x}_{t,m}^{(i)}, \pi_{t,m}^{(i)}\}_{i=1}^{N}$ is used to initialize the state density at next layer by re-applying (2).

The expected as well as the maximum a posteriori state at frame t is computed from the particle set $S_{t,1}$ at the bottom layer:

$$\mathbf{x}_t = \sum_{i=1}^{N} \pi_{t,1}^{(i)} \mathbf{x}_{t,1}^{(i)}. \quad (4)$$

2.2 Improved Silhouette Estimation with Shadow Removal and Level Set Tracking

For each particle in the posterior representation, its likelihood represents how well the projection fits the observed image(s). Many image features could be used such as color distribution, templates and optical flow constrains. As summarized in [4], most common approaches are based on silhouette and edges.

In the baseline algorithm in HumanEva, each pixel (i.e. *pix*) is modeled as a mixture of Gaussian distributions with the mean μ and the covariance ρ [9, 10]. It assumes that the distribution over foreground is uniform. However, the shadowed area caused by the moving object could also be classified as a part of the foreground. In the baseline algorithm of [9], the risk factor δ is added as follows:

$$\mathcal{N}(pix, \mu, \rho) < \frac{1}{256 \times 256 \times 256\delta}. \quad (5)$$

When δ is increased, the shadowed area can be removed from the foreground. However, it also increases the risk that a part of the true foreground may be removed along with the shadowed area.

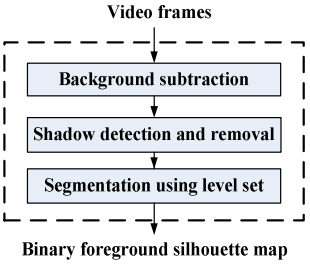

Figure 1. Block diagram of the foreground silhouette extraction pipeline.

We found that the quality of the foreground silhouette will affect significantly the results of the tracking algorithm. This motivated us to propose an improved method to extract the silhouette. Fig. 1 shows the block diagram of the proposed foreground silhouette extraction algorithm. It consists of background subtraction for initial silhouette extraction, shadow detection and level set tracking to obtain a better binary foreground silhouette for subsequent tracking. Firstly, we omitted the δ in (5) to extract the foreground information. Then, a deterministic non-model-based method for shadow detection and removal based on HSV color space is employed. It has been shown to be very robust to noise and its computational complexity is relatively low [11]. As we shall see later, it significantly improves the extraction of the silhouettes. The decision statistic is:

$$SP(\mathrm{x},\mathrm{y}) = \begin{cases} 1, \ \omega \le \frac{\mathrm{I}_k^{\mathrm{V}}(x,y)}{\mathrm{B}_k^{\mathrm{V}}} \le \varepsilon \wedge \left(\mathrm{I}_k^{\mathrm{S}}(x,y) - \mathrm{B}_k^{\mathrm{S}}(x,y)\right) \\ \quad \le \tau_{\mathrm{S}} \wedge \left|\mathrm{I}_k^{\mathrm{H}}(x,y) - \mathrm{B}_k^{\mathrm{H}}(x,y)\right| \le \tau_{\mathrm{H}} \\ 0, \ otherwise, \end{cases}, \quad (6)$$

where $\mathrm{I}_k(x,y)$ and $\mathrm{B}_k(x,y)$ are the pixel values at coordinate (x,y) in the image of frame k and in the background model respectively. The symbol ε is used to prevent the shadow identification for those points where the background is slightly changed by noise. The symbol of ω takes into account the light intensity.

Finally, we employ level set method to refine the silhouette. The method in [12] is adopted since it allows any region-based segmentation energy to be re-formulated in a local way, which is capable of segmenting objects with heterogeneous feature profiles that would be difficult to capture correctly using a standard global method. To get more accurate silhouette, we use the silhouette derived from the previous two steps as an initialization to the method.

2.3 Bi-directional Silhouette-based Likelihood Function

After extracting the silhouette, the bi-directional silhouette-based likelihood function [4] is used to measure the matching between the body model and the image observations. This function constrains the body to lie inside the image silhouette and vice versa. It solves the problem when the model does not fully cover the image silhouette. Since multiple views are used for tracking, it is necessary to combine appropriately the likelihoods computed from multiple views. In [4], the following combining likelihood is used:

$$-\log \mathrm{p}\left(\mathbf{y}_t\middle|\mathbf{x}_{t,m}^{(i)}\right)=\frac{1}{K}\frac{1}{|L|}\sum_{k=1}^{K}\sum_{l\in L}-\log \mathrm{p}^{l}\left(\mathbf{y}_t{}^{(k)}\middle|\mathbf{x}_{t,m}^{(i)}\right), \quad (7)$$

where K is the number of cameras, and L is a set of likelihood functions. In (7), the likelihood of each view is treated equally. However, as each view carries different information of the body parts, it is desirable to employ a view adaptive method for fusing the likelihoods, which we shall describe below.

2.4 APF with View Adaptive Fusion

Due to incorrect estimated pose, we found that the matching scores for different views are quite different. As an illustration, let's consider some example tracking results using the baseline algorithm for Frame 848 as shown at the top of Fig. 4. The match between the estimated articulated human body model with the first and third views seem not as good (especially we can see the left forearm). If more weights are given to those views with larger matching errors in the formulation of (7), then better results can be obtained.

Therefore, we propose a view adaptive fusion mechanism based on APF by putting different weights in fusing the likelihood of each view in (7) according to their matching errors. Specifically, the model corresponding to the maximum a posteriori poses $\mathbf{x}_t$ derived at the end of each time instant is projected back to the image plane of each view. Based on the matching errors of the model in different views, the performance of the previous APF estimate is evaluated by calculating the normalized view weighting parameter $\vartheta_{k,t}$ at the time instant t :

$$\vartheta_{k,t}=\frac{\mathrm{E}_{k,t}}{\sum_{\mathrm{k=1}}^{\mathrm{K}}\mathrm{E}_{k,t}}, \text{ where } \mathrm{E}_{k,t}\ \propto -\log\ \mathrm{p}\left(\mathbf{y}_t{}^{(k)}\middle|\mathbf{x}_t\right). \quad (8)$$

Large matching error in a view will result in a lower likelihood and the parameter $\vartheta_{k,t}$ will increase the weighting of the corresponding view by using:

$$-\log \mathrm{p}\left(\mathbf{y}_t\middle|\mathbf{x}_{t,m}^{(i)}\right)=\frac{1}{|L|}\sum_{k=1}^{K}\vartheta_{k,t-1}\sum_{l\in L}-\log \mathrm{p}^{l}\left(\mathbf{y}_t{}^{(k)}\middle|\mathbf{x}_{t,m}^{(i)}\right), \quad (9)$$

which aims to improve the matching of the model to this view in the next time instant. The algorithm is summarized in Table I. More sophisticated view adaptive fusion mechanism can also be employed.

TABLE I THE VIEW ADAPTIVE FUSION ALGORITHM BASED ON APF

Initialization $\vartheta_{k,t-1}$
for each: time instance t
  Initialization $S_{t,M+1}=S_{t-1,1}$
  for m = M down to 1 **do**
    1. Draw N particles from the weighted set $S_{t,m+1}\equiv\{\mathbf{x}_{t,m+1}^{(i)},\pi_{t,m+1}^{(i)}\}_{i=1}^{N}$ with a Gaussian diffusion model and with replacement using Monte Carlo sampling from the state probability density at previous layer m+1 using
    $\{\mathbf{x}_{t,m}^{(i)}\}_{i=1}^{N}\sim\sum_{j=1}^{N}\pi_{t,m+1}^{(j)}\mathcal{N}\left(\mathbf{x}_{t,m+1}^{(j)},\alpha^{M-m}\boldsymbol{\Sigma}\right)$.
    2. **for each** particle i
      Calculate the negative log-likelihood for the predicted particles:
      $-\log \mathrm{p}\left(\mathbf{y}_t\middle|\mathbf{x}_{t,m}^{(i)}\right)=$
      $\frac{1}{|L|}\sum_{k=1}^{K}\vartheta_{k,t-1}\sum_{l\in L}-\log \mathrm{p}^{l}\left(\mathbf{y}_t{}^{(k)}\middle|\mathbf{x}_{t,m}^{(i)}\right)$.
    end for each
    3. Calculate the normalized weights:
    $\pi_{t,m}^{(i)}=\frac{\mathrm{p}\left(\mathbf{y}_t\middle|\mathbf{x}_{t,m}^{(i)}\right)^{\beta^m}}{\sum_{j=1}^{N}\mathrm{p}\left(\mathbf{y}_t\middle|\mathbf{x}_{t,m}^{(j)}\right)^{\beta^m}}$, $i\in\{1,\ldots,N\}$.
  end for
  The maximum a posteriori poses at frame t is calculated using:
  $\mathbf{x}_t=\sum_{i=1}^{N}\pi_{t,1}^{(i)}\mathbf{x}_{t,1}^{(i)}$.
  Calculate the normalized view weighting parameter $\vartheta_{k,t}$ at the time instant t:
  $\vartheta_{k,t}=\frac{E_{k,t}}{\sum_{k=1}^{K}E_{k,t}}$, where $E_{k,t}\propto-\log \mathrm{p}\left(\mathbf{y}_t{}^{(k)}\middle|\mathbf{x}_t\right)$.
end for each

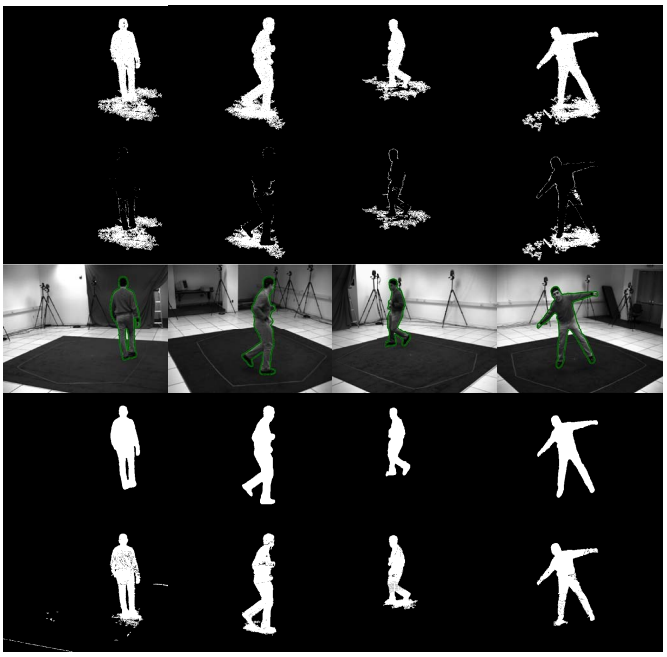

Figure 2. (First row) Results of background subtraction; (second row) results of shadow detection; (third row) results of level set; (fourth row) foreground binary silhouette; (bottom row) results of baseline algorithm.

III. EXPERIMENTAL RESULTS

Our method was tested on the S2 sequence of the HumanEva-II dataset [9, 10]. In this sequence, a subject starts by walking along an elliptical path followed by a jogging phase and then a balancing phase. Our approach was tested using frames from 400 to 600 in the jogging phase which is difficult to track for fast motion and frames from 792 to 1000 in the balancing phase. The configuration of the experiment is as follows. We use 5 layers, 200 particles per layer and the action model of G-AJL [4], which are the same as the base configuration in the baseline algorithm. Four views are used in the tracking. In order to compare with the baseline algorithm, we use the pose tracked by the baseline algorithm to initialize the proposed algorithm.

Fig. 2 shows example results in each step of foreground extraction. The final results are compared with those of the baseline algorithm. It can be shown that the shadow of the object has been removed and the silhouettes extracted are more accurate. The error between the estimated pose and the ground truth pose is expressed as the average Euclidean distance between each virtual marker [4]. The performances of the baseline algorithm, the baseline algorithm with improved silhouette and the view adaptive fusion algorithm based on

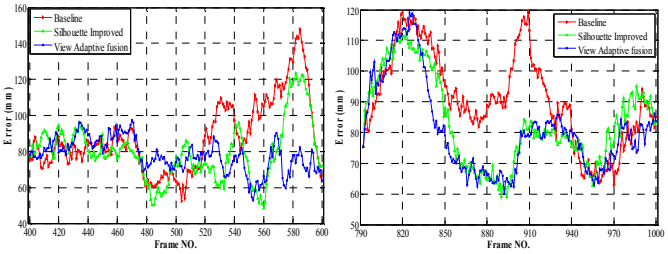

Figure 3. The performance of the proposed algorithm; (left) the absolute error of jogging sequence, (right) the absolute error of the balancing sequence.

APF are plotted in Fig. 3. Because the foreground silhouettes of the multiple views have been benefited by the improved silhouette, the tracking results outperform the baseline algorithm on the whole. When the view adaptive fusion algorithm is used, the overall performance is further improved.

Fig. 4 presents example tracking results in the framework of multi-view articulated human tracking. The tracking results have been improved by improving the silhouette, especially for the area of feet (e.g., the left foot of the human) which is influenced by the shadow. After we have employed the view adaptive fusion method, the views with larger matching errors have been given more weights and better tracking results are obtained (e.g., the first and second views in the frame 848). Meanwhile, it is worthy to note that, as more accurate 3D pose is estimated, all the views have lower errors than those of the baseline algorithm. For comparison, the tracking results of view2 and view4 of frame 566 and frames 571 to 573 from the jogging sequence are shown in Fig. 5. With the view adaptive fusion approach, more accurate tracking results can be achieved. For instance, the legs at the front and behind can also be tracked correctly (the left side in blue and the right side in yellow).

IV. CONCLUSION

A new articulated human motion tracking and pose estimation algorithm using an improved silhouette extraction method with view adaptive fusion has been presented. Shadow detection and removal, and level-set methods are employed to achieve better silhouette extraction. An adaptive view fusion approach is also proposed to improve the model matching. Experimental results show that the proposed approach has considerably better performance than the baseline algorithm in the HumanEva dataset.

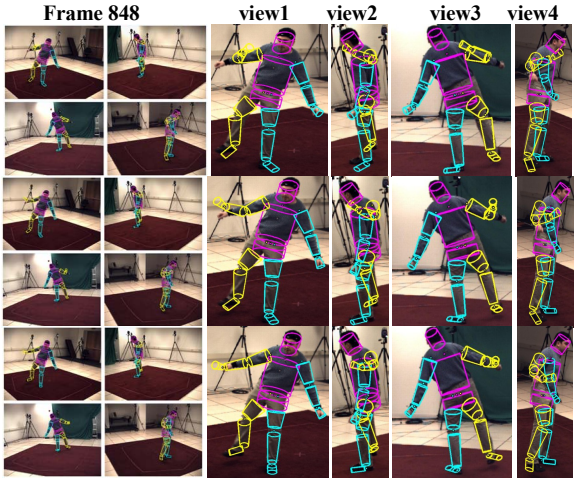

Figure 4. Results of Frame 848: (top) baseline algorithm, error 101 mm; (middle) baseline algorithm with improved silhouette, error 93 mm; (bottom) the view adaptive fusion algorithm based on APF, error 80mm.

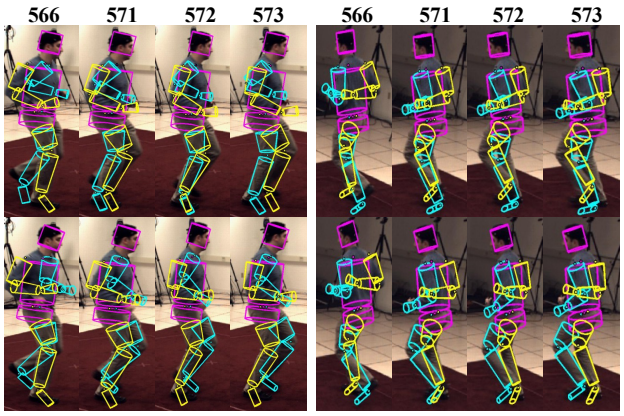

Figure 5. Upper are the results using the baseline algorithm with improved silhouette, (left) view2, (right) view4; lower are the results of corresponding views using the view adaptive fusion algorithm based on APF.

REFERENCES

[1] T. J. Cham and J. M. Rehg, "A multiple hypothesis approach to figure tracking," in Computer Vision and Pattern Recognition, volume 2, pp. 239–245, 1999.

[2] S. Wachter and H. H. Nagel, "Tracking persons in monocular image sequences," Computer Vision and Image Understanding, vol. 74, no. 3, pp. 174–192, June 1999.

[3] F. Guo and G. Qian, "Multi-view tracking of articulated human motion in silhouette and pose manifolds," in Proc. of ICASSP, pp. 1781–1784, April 2009.

[4] L. Sigal, A. Balan, and M. Black, "Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion," International Journal of Computer Vision, vol. 87, no. 1-2, pp. 4–27, 2010.

[5] D. Anguelov, P. Srinivasan, D. Koller,S. Thrun, J. Rodgers, and J. Davis, "Scape: shape completion and animation of people," ACM Trans. on Graphics, vol. 24, no. 3, pp. 408–416, 2005.

[6] N.Hasler, C.Stoll, M.Sunkel, B.Rosenhahm, and H.P.Seidel, "A statical model of human pose and body shape," Computer Graphics Forum (Proc. Eurographics 2009), Munich, Germany, March 2009.

[7] S. L. Dockstader and A. M. Tekalp, "Multiple camera tracking ofinteracting and occluded human motion," Proc. IEEE, vol. 89, no. 10, pp. 1441–1455, 2001.

[8] A. Mittal and L. S. Davis, "M2Tracker: a multi-view approach tosegmenting and tracking people in a cluttered scene," IJCV, vol. 51, no. 3, pp. 189–203, 2003.

[9] HumanEva: http://vision.cs.brown.edu/humaneva/

[10] L. Sigal and M. J. Black, "Humaneva: Synchronized video andmotion capture dataset for evaluation of articulated human motion," Technical Report CS-06-08, Brown University, Department ofComputer Science, Providence, RI, September 2006.

[11] A. Prati, I. Mikic, M. M. Trivedi, and R. Cucchiara, "Detecting moving shadows: Algorithm and Evaluation," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 25, No. 7, July 2003.

[12] S. Lankton and A. Tannenbaum, "Localizing Region-Based Active Contours," IEEE Transactions on Image Processing, vol. 17, no. 11, pp. 2029–2039, 2008.