Towards Fish-Eye Camera Based In-Home Activity Assessment

Erhan Bas, Deniz Erdogmus, Umut Ozertem, Misha Pavel
Department of Biomedical Engineering, Oregon Health and Science University, Portland, OR, USA

***Abstract*—Indoors localization, activity classification, and behavioral modeling are increasingly important for surveillance applications including independent living and remote health monitoring. In this paper, we study the suitability of fish-eye cameras (high-resolution CCD sensors with very-wide-angle lenses) for the purpose of monitoring people in indoors environments. The results indicate that these sensors are very useful for automatic activity monitoring and people tracking. We identify practical and mathematical problems related to information extraction from these video sequences and identify future directions to solve these issues.**

I. INTRODUCTION

Assessment and classification of individuals' activities is an important component in many monitoring and assessment applications ranging from security to healthcare [1-4]. In many of these systems it is necessary to distinguish 'normal" activities from those that deviate from the expected patterns. The application areas range from security to care for elders and chronically ill. In fact, care for elders – one of the rising economic and social challenges – represents a particularly important application area. One of the labor-intensive aspects of caring for elders is the necessity of continuously monitoring their behavior and asserting that everything is ok. The ability to assess "ok-ness" is particularly important for elders who live in their own homes. An automatic activity monitoring system would provide the necessary information and would detect potential problems and adverse events such as falls. Such system would have many other applications including security assurance. In security systems, it is critical to detect abnormal or illegal activities. Another example involves context-aware computing where individuals' locations, gestures and activities are used to select the most appropriate information and applications, (see e.g., [5,6]). Such context aware interfaces are useful components of cognitive aids for elders and cognitively impaired individuals.

A fundamental component in any of these applications is the ability to detect the locations and movements of the patients as well as the caregivers. There have been numerous attempts to estimate and track the location of individual as well as his gestures and gross movements for the purpose of pervasive healthcare (e.g., [7]). Most of prior attempts have been based on various wireless systems using the strength of signal to estimate distances. The results of these attempts generally yield limited accuracy that is typically not sufficient for classification of activities. In addition, these RSSI-based methods require each of the participants to wear a device or a tag [8]. The requirement of wearing a device is at odds with the notion of unobtrusive assessment and is generally useless for tracking visitors in an elder's dwelling.

EB, DE, UO, MP are with the Oregon Health and Science University (e-mails: base@csee.ogi.edu, derdogmus@ieee.org, ozertemu@csee.ogi.edu, pavel@bme.ogi.edu).

The approach to indoor tracking described in this paper is based on imaging sensors, namely a CCD cameras operating in the visible spectrum. The sensor used in the location tracking system is equipped with a wide-angle lens and is mounted on or near the ceiling of the room to be monitored. The video frames are processed locally to extract location information. An individual's location estimation is performed in two steps. First, the areas with significant motion signal are detected using background subtraction and thresholding. The background subtraction used in this project is based on modeling each pixel with a Gaussian mixture. Second, the location estimation and tracking is computed using filtering procedures based on state-estimation techniques (also known as Kalman filtering).

We note in passing that although privacy may be an issue for some participants, there are two mitigating factors. First, the image-based monitoring system may help the participants maintain independence. Second, the participants are assured that no images are stored or transmitted outside the local client PC; only location and movement estimates, together with activity classification are extracted from the images and used for further processing. The video frames are discarded.

2. MOVING OBJECT DETECTION

In the presented work, we employ the background-foreground separation approach to segmenting moving bodies from video sequences. A K-component Gaussian Mixture Model (GMM) is updated on-line to characterize each of the 3 color features (RGB in this case, because empirical evidence suggests Gaussian clusters fit well in this coordinate system [9]) for each pixel. The premise of this approach is that a moving object obstructing the background scene at a pixel will introduce a color change, thus will be assessed to be a low-likelihood outcome of the particular GMM model associated with the color distribution of that pixel. In-home environments, especially in the case of a fisheye camera mounted on the ceiling and looking vertically down, could be especially expected to have relatively tight clusters of color feature vectors for steady background objects (steady means stationary on the order of a few seconds, adjustable by an algorithm parameter). Consequently, if the current color vector of a pixel is far from the Gaussian component centers, it will have a low likelihood under the background model for that pixel and will be classified as foreground and vice versa. Specifically, a pixel color value that is less than 2.5 standard deviations from the mean of any of the K Gaussian components is decided to belong to the background. If a match occurs, then

that mixture (weight of that particular component, mean and covariance) is updated with the new pixel color value; if no match occurs then a new mixture model is created with the mean at that most recent pixel value and an initially low-weight, high variance value is imposed (this accounts for the possibility of a new stationary object, such as a new piece of furniture in our application, introduced to the environment to merge into the background over time). The least probable (smallest weighted) mixture component is eliminated by the introduction of this new component to prevent the GMM from exhibiting a growing number of components.

The above-mentioned background adaptation procedure is important to account also for small changes such as brightness variations besides the new entries to the background. For this purpose, an online update algorithm is used. The probability of observing a certain pixel value for a channel (a vector for R-G-B channels, or a scalar value for a single Gray level channel) after t frames is given as

$$p(\mathbf{c}_t) = \sum_{k=1}^{K} w_{k,t} G(\mathbf{c}_t; \boldsymbol{\mu}_{k,t} \boldsymbol{\Sigma}_{k,t}) \quad (1)$$

where $w_{k,t}$ is the weight of the k^{th} component to the Gaussian mixture, $\boldsymbol{\mu}_{k,t}$ is the mean vector of this component and $\boldsymbol{\Sigma}_{k,t}$ is the covariance matrix computed from the pixel value history (we assume the covariance to be diagonal with σ_k^{i2} in its I^{th} diagonal entry). In our implementation, pixel colors are represented in the RGB space and a 3-dimensional color vector is modeled by the GMM in (1). Components of the GMM are ranked in descending order of weights. For background modeling, the first (most weighted) Gaussian component is used and for foreground segmentation two different foreground calculation schemes are implemented: (i) the usual way to calculate foreground, using the information of mixture models, (ii) difference between current frame and most-likely background image.

In (i), for each color channel, the following condition is checked to identify foreground pixels. Let $\mathbf{c}$ be the color vector of a pixel. If the sum of the weights of Gaussian components for which $\mathbf{c}$-$\boldsymbol{\mu}_k$ is within 2.5 standard deviations (for each respective component) is less than a threshold γ (0.7 in our implementation), that is

$$| c^i - \mu_k^i | \leq 2.5\sigma_k^i, i = 1,2,3 \quad (2)$$

then that pixel is marked as foreground.

In (ii), subtracting the current frame from the most-likely background image (obtained approximately by assigning the mean of the largest weight Gaussian as the color of each pixel, denoted by BKG) yields an RGB difference image that, when compared with a preset threshold ν in each color channel (set to 2^5 for 2^8-level channels), segments the frame into background and foreground objects (by intersecting the binary segmentation result of each channel). Then, the foreground objects are detected and labeled using connected component analysis.

To reduce shadow effects, foreground pixels are found as in (ii), and shadow regions are determined by comparing foreground regions with their respective BKG pixels converted to HSV coordinates. This approach is based on the premise that shadow results in significant change in H and has relatively little effect on V and S [10]. A sample shadow estimate removed foreground detection result can be seen in Figure 1.

Figure 1. A typical output frame showing the results of foreground object detection based on the GMM background model with shadow removal (right) and estimated object location based on particle filter with background subtraction (BGS) result (left). On the right, black cluster is the moving object and white cluster is the estimated shadow.

Figure 2. The view from fish-eye camera (left) and object localization from foreground segmentation (right). On the right, the segmented foreground object is assumed to be a standing person whose feet are centered at a small radius and an angle determined by the mode of the angular density of pixels in the cluster. The mode enables robustness to the inclusion of attached objects such as the chair that the person is moving in this frame.

III. OBJECT LOCALIZATION AND TRACKING

A fish-eye camera mounted on the ceiling directed vertically to image a room naturally induces a polar coordinate system (see Figure 2), consequently, for dynamic tracking of a moving object in a state estimation framework, the measurement equation becomes nonlinear (Cartesian to polar coordinate frame transformation). Upon identifying a cluster of pixels corresponding to a foreground object (after shadow removal), these pixel locations are converted to polar coordinates (r,θ) assuming the center of the frame to be the origin and the horizontal-right direction is zero-phase with counterclockwise angle measurement (clockwise in images due to y-axis being flipped). Based on extensive observation of standing and walking people, it is deduced that the body of the person occupies a cluster of pixels that form a radial main axis, and the edge of a foot becomes the minimum-radius pixel in this cluster. To calculate the minimum-radius, first object orientation is estimated. Orientation of a moving object from the center of the image is calculated by finding the mode of the angle distribution, $\theta_{\max}$ (estimated with 1^o resolution using a 1^o-fixed-width histogram estimate of the angle distribution of the foreground cluster). Then we project all foreground pixels to the line segment along $\theta_{\max}$, and determine the distribution of radii with a histogram (using 1-pixel-wide fixed bins). The smallest radius, $r_{\min}$, at which the histogram attains 10% of its peak value is estimated to be the foot location. The use of such a threshold on the density eliminates the effect of

shadow pixels misclassified as foreground. The height of the person being typically larger than the width, the mode is also found to be a reliable estimate of the centerline axis of the body when the person is in the scene from head-to-toe. The mode also is a more robust estimator for angle compared to the mean because a person moving an inanimate object might be perceived as one object by the foreground segmentation algorithm. These observations are illustrated in Figure 2.

The position estimate $(r_{\min}, \theta_{\max})$ obtained from each frame is utilized as the measurement equation of a dynamic state-model of the moving object to be tracked with a particle filter. The particle filter technique is a sequential Monte Carlo approach to recursive Bayesian state estimation in dynamic systems and its computationally convenient and efficient algorithms apply to linear and nonlinear dynamic systems corrupted by Gaussian and non-Gaussian noise distributions [11]. State dynamics and the measurement equation of our system model is given as

$$\begin{aligned} \mathbf{s}_t &= \mathbf{F}\mathbf{s}_{t-1} + \mathbf{a}_{t-1} \\ \mathbf{z}_t &= \mathbf{g}(\mathbf{s}_t) + \mathbf{u}_t \end{aligned} \tag{3}$$

where the state transition is a linear Newton dynamic model with a spherical white random Gaussian acceleration $\mathbf{a}_t$. The measurement noise $\mathbf{u}_t$ is modeled to be state-dependent additive white Gaussian process. The state vector consists of the two-dimensional Cartesian position and velocity vectors in the image plane, and the measurement vector consists of the position vector in polar coordinates:

$$\begin{aligned} \mathbf{s} &= \begin{bmatrix} p_x & p_y & v_x & v_y \end{bmatrix}^T \\ \mathbf{z}_t &= \begin{bmatrix} p_r & p_\theta \end{bmatrix}^T \end{aligned} \tag{4}$$

In particular, we model $\mathbf{a}$ as a 4-dimensional Gaussian with zero-mean and zero-variance in the first two dimensions corresponding to position, and with 0-mean and $\sigma_a^2\mathbf{I}$ covariance in the latter two dimensions corresponding to velocity (we set $\sigma_a = 1/(2T^2)$ pixel/s^2), where $1/T$ is the frame rate. The state transition matrix and the Cartesian-to-polar measurement functions are:

$$\mathbf{F} = \begin{bmatrix} 1 & 0 & T & 0 \\ 0 & 1 & 0 & T \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \qquad \mathbf{g}(\mathbf{s}) = \begin{bmatrix} \sqrt{p_x^2 + p_y^2} \\ \arctan(p_y / p_x) \end{bmatrix} \tag{5}$$

The measurement noise $\mathbf{u}_t$ is modeled as a state-dependent Gaussian based on experimental data collected as described next. Ground truth position data of a subject walking on semicircles centered at the origin and lines passing through the origin have been acquired by manually marking the location of the midpoint of the feet of the subject. The semicircular walking trajectories consisted of radii equal to 18, 36, 54, and 72 inches (1inch=2.54cm). The linear trajectories were placed at 0°, 90°, 135°, and 180° all merging at the origin. The following assumptions were made: (i) the radial and angular measurement errors (noise) of the frame-based estimates described in Section 2 are independent, (ii) both radial and angular noise distributions are independent of angular position, but not radial position, (iii) radial error also depends on radial velocity but not tangential velocity. Experimental data was fitted with kernel density estimates (KDE) and then simplified to radius-dependent Gaussians – it was observed that the assumptions were reasonably valid given the manually obtained ground-truth data. For instance, the radius dependency of the radial error is summarized by the KDE shown in Figure 3. An analytical function was used to approximate each of the mean and standard deviation of the Gaussian radius error models illustrated in this figure. Similarly, radius-dependent standard deviation was modeled for angular measurement error; the angle measurement error was found to be zero-mean regardless of position and speed. Finally, it has been observed that (assumption (iii) above) if a radial velocity component exists (subject walking directly to or away from the origin), then depending on the radial speed, the bias in radius measurement error must be compensated (due to step-size). This was included in the model as a linear additive bias in the form $\alpha\mathbf{v}_r$, where $\mathbf{v}_r$ is the radial velocity vector (calculated from the state using: $\mathbf{v}_r = \mathbf{p}\mathbf{p}^T\mathbf{v}/\mathbf{p}^T\mathbf{p}$). The noise model is continuously updated during particle filter estimation iterations based on the current state estimates and the models developed as described above.

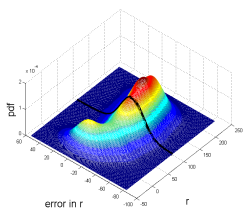

Figure 3. KDE of radius error vs radius. The black cross section is the error distribution for a particular radius.

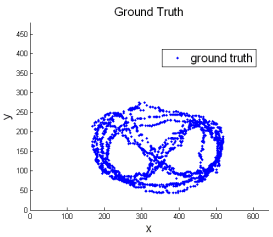

Figure 4. Ground truth of position for each frame in the test experiment.

IV. EXPERIMENTAL RESULTS

Particle filter (PF) estimation results are compared with the single-frame based background subtraction (BGS) based instantaneous estimates for each frame. The PF is implemented with 300 particles initialized to the foreground location estimate using BGS plus unit (pixel^2) variance Gaussian perturbation. At time t, position estimate of an object is given by the weighted average of the three highest weighted particles (assuming unimodal state distribution, this is meant to approximate the mode). A single frame of the object detection step is depicted in Figure 1, where red dot shows our position estimate based on the current state, whereas green dot represents the actual location of the object in this particular frame. The location of the detected object is represented with the intersection of centered circle with the radial line that makes an angle $\theta_{\max}$ with the horizontal axis as shown in Figure 2.

Ground truth of the object location over 2300 frames is given in Figure 4. Figures 5-8 show the error in p_r and p_θ over time and over p_r and p_θ consecutively. Tracking with PF reduces the error in the mean especially in p_r with some outliers. Outlier regions are mainly resulted from the system model that is used in this experiment that has some inabilities to track sudden jumps of the angle when the

subject passes near the origin.[1] As a result, at small radii, an object that moves across the 0^o line segment emerging from the origin makes a sudden jump in θ that is not a multiple of 2π (typically around π). Figure 6 shows examples of π-rad transitions that occur around frames 520, 1350, and 2000, which leads to an improper update of particle weights, causing poor estimates. Although resampling provides a partial remedy for this problem, the best solution is to modify the angular measurement noise model such that it becomes more uniform towards the origin; this correction will be implemented in the algorithm in the future. Figures 9-10 display the comparison of the error of p_r and p_Θ in BGS and PF methods with their correlation coefficient ρ. Although current model reduces the average error, the PF errors are still correlated with frame-based measurements obtained using BGS. This is due to the simplistic random walk model utilized in the dynamic model and in the future, an adaptive acceleration prediction model will be incorporated into the model allowing for a time-varying acceleration strategy and a smoother trajectory profile.

V. CONCLUSIONS

In this paper, we presented our initial experiments with fish-eye cameras for indoor motion tracking. The results indicate that these sensors are very promising for accurate assessment of complex behavior and activity patterns with the proper application of video processing and computer vision techniques. Due to the ceiling-mount setup we preferred in order to obtain a relatively simple spherical (polar) coordinate system transformation between the images in the sequence and the 3-dimensional environment, the accuracy of localization and activity/pose classification is hampered when the person/object of interest is directly under the camera, at zero-radius in the image plane. The radius dependent resolution of the image (m^2/pixel) and associated increased uncertainty with measuring and tracking position and velocity need to be taken into account for proper dynamic modeling and tracking using particle filters or equivalent technique. The results indicate that using at least two fisheye cameras (for instance mounted on the ceiling at opposite corners) will increase tracking accuracy significantly. Future work will include multi-camera tracking as well as automatic detection of events of interest, such as medication adherence and falls of the individual being monitored.

ACKNOWLEDGEMENTS

This work was supported by Intel under the OHSU BAIC project, and partially by the following grants NIA: P30-AG024978, NSF: ECS-0524835, ECS-0622239, and IIS-0713690.

REFERENCES

[1] Y. Ivanov, C. Stauffer, A. Bobick, W.E.L. Grimson, "Video Surveillance of Interactions," Proceedings of the 2nd IEEE International Workshop on Visual Surveillance (VS1999), Fort Collins, Colorado, USA, pp. 82–89, 1999.

[1] Note that 2π-jumps in θ when the subject crosses the x-axis can be easily handled by using a proper angle wrapping scheme when measuring angle differences for likelihood calculations.

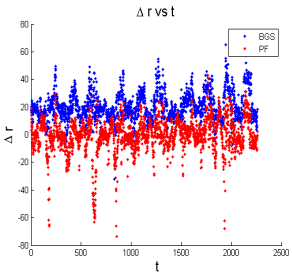

Figure 5. Error in r vs t.

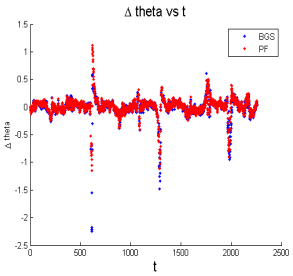

Figure 6. Error in θ vs t.

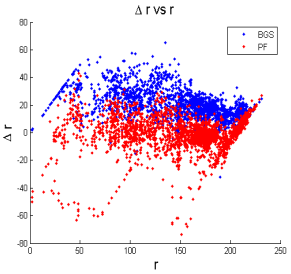

Figure 7. Error in r vs r.

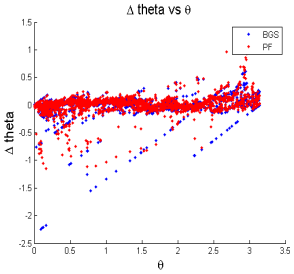

Figure 8. Error in θ vs θ.

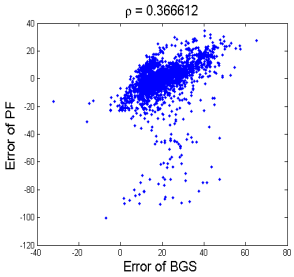

Figure 9. Comparison of error in r BGS wrt PF.

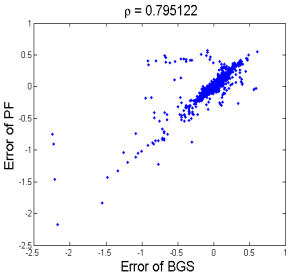

Figure 10. Comparison of error in θ BGS wrt PF.

[2] R. Cucchiara, C. Grana, M. Piccardi, A. Prati, "Detecting Moving Objects, Ghosts and Shadows in Video Streams," IEEE Transactions on PAMI, vol. 25, no. 10, pp. 1337–1342, 2003.

[3] R. Cucchiara, C. Grana, A. Prati, G. Tardini, R. Vezzani, "Using Computer Vision Techniques for Dangerous Situation Detection in Domotic Applications," Proceedings of the International Conference on Intelligent Distributed Surveillance Systems (IDSS04), London, Great Britain, pp. 1–5, 2004.

[4] J. Menendez, S.A. Velastin, "A Method for Obtaining Neural Network Training Sets in Video Sequences," Proceedings of the 3rd IEEE International Workshop on Visual Surveillance (VS2000), Dublin, Ireland, pp. 69–75, 2000.

[5] T. Moran, P. Dourish, "Introduction to the Special Issue on Context-Aware Computing," Human-Computer Interaction, vol. 16, no. 2-3, 2001.

[6] H. Balakrishnan, N.B. Priyantha, "The Cricket Indoor Location System: Experience and Status," Proceedings of the Workshop on Location Aware Computing, Seattle, WA, USA, pp. 7-9, 2003.

[7] W.R. Jih, S.Y. Cheng, J.Y.J. Hsu, "Context-Aware Access Control on Pervasive Healthcare," Proceedings of Workshop: Mobility, Agents, and Mobile Services (MAM), pp. 21 – 28, 2005.

[8] G.V. Zàruba, M. Huber, F.A. Kamangar, I. Chlamtac, "Indoor Location Tracking Using RSSI Readings from a Single Wi-Fi Access Point," Wireless Networks, vol. 13, no. 2, pp. 221-235, 2007.

[9] C. Stauffer, W.E.L. Grimson, "Adaptive Background Mixture Models For Real-time Tracking", Proceedings of CVPR, pp. 246-252, 1999.

[10] Prati, et al., "Detecting Moving Shadows: Algorithms and Evaluation," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 25, pp. 918-923, 2003.

[11] A. Doucet, N. de Freitas, N. Gordon (eds), *Sequential Monte Carlo Methods in Practice*, Springer-Verlag, 2001.