

# Visual Tracking Using Color Cameras and Time-of-Flight Range Imaging Sensors

Leila Sabeti, Ehsan Parvizi, Q.M. Jonathan Wu

Department of Electrical and Computer Engineering, University of Windsor, Canada

Email: {sabeti, parvizi, jwu}@uwindsor.ca

**Abstract**—This work proposes two particle filter-based visual trackers — one using output images from a color camera and the other using images from a time-of-flight range imaging sensor. These proposed trackers were compared in order to identify the advantages and drawbacks of utilizing output images from the color camera as opposed to output from the time-of-flight range imaging sensor for the most efficient visual tracking. This paper is also unique in its novel mixture of efficient methods to produce two stable and reliable human trackers using the two cameras.

**Index Terms**—visual tracking, color camera, time-of-flight sensor, illumination invariant, particle filter

## I. INTRODUCTION

Achieving reliable visual tracking in complex and real-world environments remains an ongoing research problem. Visual tracking has applications in areas such as intelligent transportation, video surveillance, human-computer interaction, medical diagnostics and video compression. The use of depth information to enhance object tracking is growing due to the availability of third dimension information — the distances between the objects and the sensor.

Stereo sensors have been most commonly employed to determine the depth map of the scene by calculating disparities from images captured by two cameras separated by a baseline. This process of stereo matching to obtain a depth-map tends to be computationally intense, and the results are not adequately accurate. In addition, passive stereo sensors require the presence of sufficient ambient illumination so that they can produce good quality shots. These limitations have motivated the development of active depth sensors such as laser range scanners and time-of-flight (TOF) sensors [1]. Overall, TOF sensors have significant advantages over laser range scanners, including higher accuracy, existence of vertical as well as horizontal scanning capability, pixel-level measurement quality and relatively compact weight and size [2]. Some of the works employing TOF sensors include [3], [4].

Object tracking using color sensors has been more popular than visual tracking using TOF range sensors because color illustrates an object's details, allows faster processing and is invariant to the object pattern's geometric differences.

Visual tracking can be classified as low-level and high-level approaches. In a low-level approach, the image is segmented or classified in order to localize the blob or object without an initial hypothesis. The high-level

approach performs object association from one frame to the next by generating an object hypothesis and then evaluating the likelihood of a set of given hypotheses for each frame, based on the most recent measurement. The particle filter [5] is one of the most successful object tracking methods for solving nonlinear cases in which noise may be non-additive and non-Gaussian. It is also able to represent simultaneous alternative hypotheses. Several researchers [6]–[9] have adopted the particle filter as a recursive Bayesian filter. Real world experimental evidence, such as in [6], [9], represent the superiority of the particle filter over mean shift, Kalman filter and the extended Kalman filter (EKF).

We have combined both low-level and high-level approaches to construct effective tracking hypotheses. This paper contributes the novel mixture of efficient methods to produce two stable and reliable human trackers using a color camera and a TOF range sensor. There is also a discussion in Section VI that clarifies the advantages and drawbacks of employing each of these sensors for effective visual tracking.

The organization of this paper is as follows. A brief explanation of the particle filter is provided in Section II. The proposed color-based and depth-based trackers are described in Sections III and IV, respectively. Section V presents the experimental results, which is followed by a discussion in Section VI. Finally, Section VII concludes the paper.

## II. THE PARTICLE FILTER

To solve nonlinear cases in which noise may be non-additive or non-Gaussian, we have exploited the particle filter (PF) based on sequential Monte Carlo simulations. The PF utilizes a set of random samples (also called particles) to estimate posterior distribution. When particles are properly placed, weighted and propagated, posterior distribution can be estimated sequentially over a period of time. Equation (1) approximates the posterior distribution:

$$p(x_t | y_{1:t}) \approx \sum_{i=1}^N w_t^i \delta(x_t - x_t^i) \quad (1)$$

where,  $\delta$  is the Dirac delta function. Particles are drawn from  $p(x_t | y_{1:t})$ , and the approximation approaches the true posterior when  $N$  is sufficiently large. Since the posterior distribution  $p(\cdot)$  is unknown, it is approximated

in particle filtering by a set of properly weighted particles drawn from a known proposal distribution,  $q(x_{0:t} | y_{1:t})$ .

Considering that the system is a first order Markov process and observations are conditionally independent given the states, particle weights can be computed from the following iterative equation:

$$w_t^i = w_{t-1}^i \frac{p(y_t | x_t^i) p(x_t^i | x_{t-1}^i)}{q(x_t^i | x_{t-1}^i, y_t)} \quad (2)$$

where,  $p(y_t | x_t^i)$  is the sensor's likelihood distribution,  $p(x_t^i | x_{t-1}^i)$  is the transition density and  $q(x_t^i | x_{t-1}^i, y_t)$  is the proposal distribution. In the conventional PF — CONDENSATION [5] — transition probability is adopted as the proposal distribution [10]. Particles drawn from the transition density do not consider the most recent observations, making their contributions to the posterior estimation negligible. As a result, background clutter can distract the conventional PF. Our work takes into account the latest measurements using an auxiliary sensor while assessing importance sampling.

### III. COLOR TRACKING

We have implemented an efficient, color-based human face tracker that uses a CCD color sensor. The human head is approximated by an ellipse,  $X = \{x_t^c, y_t^c, l_m, l_M\}$  where,  $(x_t^c, y_t^c)$  are the center coordinates, and  $l_m$  and  $l_M$  are the lengths of the minor and major axes. To estimate the head's state, particle weights are measured using the three distributions of the equation (2). The following sections explain how these distributions are defined in this work using low-level and high-level frame information to produce precise weight computations.

#### A. The proposal distribution

Our system's proposal distribution combines high and low-level approaches in the sense that particles are drawn from the transition prior distribution by using the CONDENSATION algorithm. In addition, samples are propagated through a Gaussian distribution obtained from low-level information. The low-level process used is a fast skin color classifier [11] — a technique that selects the color that appears most in the face of the model as the reference color, so that pixels similar in color to the reference color are selected as candidates.

This method takes a skin patch from the target person to obtain that region's maximum appearing color. This task can be done manually or by using face detection techniques in the first frame. Matrix  $\mathbf{S}$  is defined as the color-level matrix for the skin patch from the target person's face, expressed as

$$\mathbf{S}_c = \begin{bmatrix} c_{11} & c_{12} & \dots & c_{1j} \\ c_{21} & c_{22} & \dots & c_{2j} \\ \vdots & & & \\ c_{i1} & c_{i2} & \dots & c_{ij} \end{bmatrix} \quad (3)$$

where,  $c_{ij} \in \{r_{ij}, g_{ij}, b_{ij}\}$  is the color of each pixel in the skin patch. Let  $\mathbf{I}_c$  denote the color-level matrix for the whole image, expressed as

$$\mathbf{I}_c = \begin{bmatrix} c'_{11} & c'_{12} & \dots & c'_{1n} \\ c'_{21} & c'_{22} & \dots & c'_{2n} \\ \vdots & & & \\ c'_{m1} & c'_{m2} & \dots & c'_{mn} \end{bmatrix} \quad (4)$$

where,  $m \geq i$  and  $n \geq j$ . Now, if  $r_0, g_0$  and  $b_0$  are the red, blue and green values of the most repeated color in the skin patch, then subtracting  $r_0, g_0$  and  $b_0$  from all the RGB values in the image will provide

$$\delta r_{ij} = |r_{ij} - r_0|, \quad (5)$$

$$\delta g_{ij} = |g_{ij} - g_0|, \quad (6)$$

and

$$\delta b_{ij} = |b_{ij} - b_0| \quad (7)$$

where,  $i \in \{1, 2, \dots, m\}$  and  $j \in \{1, 2, \dots, n\}$ . Now we define  $\delta$  as

$$\delta_{ij} = \max(\delta r_{ij}, \delta g_{ij}, \delta b_{ij}) \quad (8)$$

, and the maximum RGB value in the image,  $\eta_{ij}$ , as

$$\eta_{ij} = \max(r_{ij}, g_{ij}, b_{ij}) \quad (9)$$

where,  $i \in \{1, 2, \dots, m\}$  and  $j \in \{1, 2, \dots, n\}$ . If all the colors are normalized with values between 0 and 1, then  $\eta_{ij}$  usually has a value of 1.  $\Delta_{ij}$  is a boundary value for each pixel, determined by

$$\Delta_{ij} = \eta_{ij} - \delta_{ij} \quad (10)$$

A  $\Delta_{ij}$  greater than a specific color threshold implies that the maximum difference has fallen in the range and the color of the selected pixel belongs to the skin distribution. The threshold value is obtained from the receiver operating characteristic (ROC) curve — which is tested for various sets of images, including CVL face database, with a set of 798 images containing 114 people.

The first central moment of the skin pixels located in the image provides the mean value of the Gaussian distribution used as a part of the proposal distribution:

$$E[x] = \frac{\sum_x \sum_y x \mathbf{I}(x, y)}{\sum_x \sum_y \mathbf{I}(x, y)} \quad (11)$$

$$E[y] = \frac{\sum_x \sum_y y \mathbf{I}(x, y)}{\sum_x \sum_y \mathbf{I}(x, y)} \quad (12)$$

and, the second central moment provides the covariance

$$Cov(x, y) = \begin{bmatrix} \sigma_{xx} & \sigma_{xy} \\ \sigma_{xy} & \sigma_{yy} \end{bmatrix} \quad (13)$$

where,

$$\sigma_{xx} = \sum_x \sum_y (x - E[x])^2 \mathbf{I}(x, y) \quad (14)$$

$$\sigma_{yy} = \sum_x \sum_y (y - E[y])^2 \mathbf{I}(x, y) \quad (15)$$

$$\sigma_{xy} = \sum_x \sum_y (x - E[x]) (y - E[y]) \mathbf{I}(x, y) \quad (16)$$

Particles are propagated through the transition prior distribution as well as by the Gaussian distribution obtained from the low-level distribution mentioned above.

### B. The likelihood distribution

Since color-based systems are sensitive to gradual or drastic illumination variations, an illumination invariant method based on color ranks [12] was adopted in this system's high-level process to create a reliable likelihood distribution for the proposed tracker. This method compares the color orderings of the object pixels rather than comparing colors because while an object's colors can vary over time due to changes in illumination, visual angle and camera parameters, the color orderings of an object's pixels are highly preserved. In other words, if the color level of a pixel  $p_1$  is higher than that of  $p_2$  from the model object in the first image  $\mathbf{I}$ , the same fact is true for the corresponding pixels  $p'_1$  and  $p'_2$  in the second image  $\mathbf{I}_2$ .

Color ranks are computed in a manner similar to the normalized cumulative color histograms. The color rank measure of pixel  $p$  for image  $\mathbf{I}$ , is expressed as:

$$r^k[\mathbf{I}](p) = \frac{\sum_{l=0}^{c^k(p)} H^k[\mathbf{I}](l)}{\sum_{l=0}^{L-1} H^k[\mathbf{I}](l)}, \quad k = R, G, B \quad (17)$$

where  $L$  is the number of levels used to quantize the color components generally set to 256, and  $H^k[\mathbf{I}](l)$  is the histogram count of pixels in the image  $\mathbf{I}$  that contain the R,G or B level of  $l$ .

The likelihood distribution [13] is defined as:

$$p(y_t | x_t^i) = 1 - d^i(s_T[\mathbf{I}_K], s_M[\mathbf{I}_K]) \quad (18)$$

where  $d^i$  is the distance measure between the spatial ranks information of the target and the object model for each particle  $i$ . To obtain the similarity measure, color ranks are computed for each particle and compared to the object model estimated in the previous frame. To compute this distance measure, the color rank values are first quantized into  $N$  sections,

$$\zeta = 0, \frac{1}{N}, \frac{2}{N}, \dots, 1 \quad N \leq L \quad (19)$$

The total number of RGB levels in the image (usually 256) is denoted by  $L$ . Note that the rank measure of a pixel  $p$  for color  $k$  in image  $\mathbf{I}$  has a value between 0 and 1:

$$r^k[\mathbf{I}](p) \in [0, 1] \quad k = R, G, B \quad (20)$$

For each image, the rank measurements closest to the quantized values are computed. Hence, for  $n \in [0, N-1]$ :

$$l'_1 = \arg \min_{l=0, \dots, L-1} \|r^k[\mathbf{I}_1](l) - \zeta(n)\| \quad (21)$$

$$q^k[\mathbf{I}_1](n) = r^k[\mathbf{I}_1](l'_1) \quad (22)$$

and,

$$l'_2 = \arg \min_{l=0, \dots, L-1} \|r^k[\mathbf{I}_2](l) - \zeta(n)\| \quad (23)$$

$$q^k[\mathbf{I}_2](n) = r^k[\mathbf{I}_2](l'_2) \quad (24)$$

The summation of all the rank measures at each segmented section is obtained as:

$$a^k[\mathbf{I}_1](n) = \sum_{\{l_1 | r^k[\mathbf{I}_1](l_1) \in \mathbf{U}_1\}} r^k[\mathbf{I}_1](l_1) \quad (25)$$

$$a^k[\mathbf{I}_2](n) = \sum_{\{l_2 | r^k[\mathbf{I}_2](l_2) \in \mathbf{U}_2\}} r^k[\mathbf{I}_2](l_2) \quad (26)$$

The summations of equations (25) and (26) cover all color levels  $l$ , whose color rank measures  $r^k[\mathbf{I}](l)$  belong to  $\mathbf{U}_1$  and  $\mathbf{U}_2$ , respectively.  $\mathbf{U}_1$  and  $\mathbf{U}_2$  include all the rank measures inside the section  $n$  where,

$$\mathbf{U}_1 = [q^k[\mathbf{I}_1](n), q^k[\mathbf{I}_1](n+1)] \quad (27)$$

and

$$\mathbf{U}_2 = [q^k[\mathbf{I}_2](n), q^k[\mathbf{I}_2](n+1)] \quad (28)$$

Finally, the distance measure is computed as:

$$d^i(a[\mathbf{I}_1], a[\mathbf{I}_2]) = \sum_k \sqrt{\sum_{n=0}^{N-1} (a^k[\mathbf{I}_1](n) - a^k[\mathbf{I}_2](n))^2} \quad (29)$$

for  $k = R, G, B$  and  $i = 1, 2, \dots, N$ .

### C. Measuring particle weights

Weights are computed through equation (2) by selecting the proper proposal distribution to propagate samples, as well as using an illumination invariant likelihood distribution. A first order dynamic model with adaptive noise is employed to obtain the transition density:

$$\mathbf{x}_t = \mathbf{A}\mathbf{x}_{t-1} + \mathbf{K}_{t-1}\mathbf{N}_{t-1} \quad (30)$$

where,  $\mathbf{A}$  represents the deterministic component of this model, and  $\mathbf{N}_{t-1}$  is a multivariate Gaussian random variable. The constant velocity model with  $\mathbf{A} = \mathbf{I}$  is considered here for the nature of random head motion.  $\mathbf{K}$  is related to the head's velocity in the sense that it increases when the head is moving with a higher velocity — causing an increase in the variance of the process noise. Consequently, samples are propagated over a larger area in the state space to increase the efficiency of head localization for faster head motions. Samples' weights are then normalized so that:

$$\sum_{i=1}^N w_t^i = 1 \quad (31)$$

Finally, the head's location is estimated by computing the expected value of the weights while color information is updated at each frame only if the estimation confidence is high.

#### IV. DEPTH TRACKING

In addition to the color-based tracker, we have implemented an efficient, depth-based human tracking algorithm using a TOF range sensor<sup>1</sup>. The tracker was tested for human tracking by approximating a human body with a scalable rectangle.

##### A. The proposal distribution

This system's proposal distribution combines high and low-level approaches in the sense that particles are drawn from the transition prior distribution by using the CONDENSATION algorithm. In addition, samples are propagated through a Gaussian distribution obtained from low-level information. An efficient segmentation method was selected as the bottom-up process to detect objects of interest [14]. In this method, the depth density function of each frame is evaluated to determine the objects' concentrations in the scene using an adaptive selection of range dividers based on the distribution. The corresponding depth distribution  $p(z)$  is estimated using a kernel applied to the depth histogram. The local maxima and minima vectors of  $p(z)$ , i.e.,  $L_{max}$  and  $L_{min}$  are determined from equations (32) and (33), respectively. The  $i$ -th element of  $L_{max}$  is expressed as:

$$L_{max}(i) = \arg \max p(z) . \quad (32)$$

Then,

$$L_{min}(j) = \arg \min_{z \in R_j} p(z) , \quad (33)$$

$$R_j = \{L_{max}(j) < z < L_{max}(j + 1)\} \quad (34)$$

where,  $j = 1, 2, \dots, M - 1$ , and  $M$  is the total number of the local maxima.

The range dividers  $S(\cdot)$  can be determined from

$$S(k) = \begin{cases} Z_{min} & k = 1 \\ L_{min}(k - 1) & 2 \leq k \leq M \\ Z_{max} & k = M + 1 \end{cases} \quad (35)$$

where,  $(Z_{min}, Z_{max})$  is the dynamic range of the pixel values. These dividers are used to partition the scene into different depth divisions  $D_l$  to separate adjacent and overlapping objects, denoted by:

$$D_l(x, y) = \begin{cases} 1 & S(l) \leq \mathbf{I}(x, y) < S(l + 1) \\ 0 & \text{otherwise} \end{cases} , \quad (36)$$

where  $l = 1, 2, \dots, M$ , and  $\mathbf{I} : \mathbb{R}^2 \rightarrow (Z_{min}, Z_{max})$  refers to the depth image. Each division holds a number of objects that are detected using connected component analysis in that particular depth domain. The object blobs resulting from this process are then used to extract the final depth pixel values of each of the detected objects. These objects are further narrowed down to human candidates based on their aspect ratios and relative sizes to depth means.

<sup>1</sup><http://www.mesa-imaging.ch/prodviews.php>

Next, each segmented object is characterized by its depth histogram as well as its horizontal and vertical distributions — which are exclusive for every object in the scene and form unique signatures for the purpose of tracking. The resulting tracked object can not only be located with respect to the horizontal and vertical axes of the camera, but its distance from the imaging sensor is available at all times, making it a 3-D tracking experience.

##### B. The likelihood distribution

Depth histogram is used to build the likelihood distribution of the PF in equation (2) as

$$p(y_t | x_t^i) = 1 - d^i(H_T[\mathbf{I}], H_M[\mathbf{I}]) \quad (37)$$

$d^i$  is the distance measure between the depth histogram information of the target and the object model for each particle  $i$ , and is determined as:

$$d^i(H[\mathbf{I}_1], H[\mathbf{I}_2]) = \sqrt{\sum_{l=0}^{L-1} (H[\mathbf{I}_1](l) - H[\mathbf{I}_2](l))^2} \quad (38)$$

where  $i = 1, 2, \dots, N$ ,  $L$  is the number of levels used to quantize the depth values generally set to 256, and  $H[\mathbf{I}](l)$  is the histogram count of pixels in the image  $\mathbf{I}$  that contain the depth level of  $l$ .

##### C. Measuring particle weights

A first order dynamic model with adaptive noise is employed to obtain the transition density:

$$\mathbf{x}_t = \mathbf{A}\mathbf{x}_{t-1} + \mathbf{K}_{t-1}\mathbf{N}_{t-1} \quad (39)$$

where,  $\mathbf{A}$  represents the deterministic component of this model, and  $\mathbf{N}_{t-1}$  is a multivariate Gaussian random variable. Samples are taken from the proposal distribution containing the segmented object, as well as from the predicted area defined by the transition density. The constant velocity model with  $\mathbf{A} = \mathbf{I}$  is chosen for the nature of random human walking.  $\mathbf{K}$  is related to the object's velocity in the sense that it increases when the person is moving with higher velocity — causing an increase in the variance of the process noise. Consequently, for faster human motions samples are propagated over a larger area in the state space to increase the efficiency of object localization. Samples' weights are then normalized using equation (31).

In the proposed depth-based tracker, the person's location is estimated by computing the expected value of the weights while the histograms are also updated due to the occurrence of object deformation during translation if estimation confidence is high.

#### V. EXPERIMENTAL RESULTS

Experiments were performed on a 2 GHz PC with 2 GB memory running Windows XP as the operating system. More than 15 different videos, each including more than 500 frames, were used in the experiment. Most of the publicly available videos used in the color-based tracker

contained real-world environments that do not consider any special constraints. The 24-bit RGB sequences had a resolution of  $240 \times 320$ .

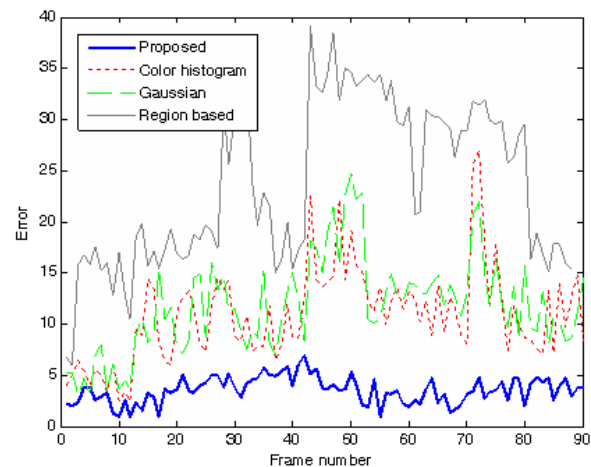
To demonstrate the efficiency of the proposed work in color-based object tracking, we have utilized both real-world data and synthetic data simulations. There are large-scale changes, pose and illumination variations, occlusions and rapid motions in the testing sequence. Background clutter is also another factor present in the sequence that causes difficulties in tracking.

In the first experiment, synthetic sequences were created by superimposing a face image of  $50 \times 63$  pixels moving with a nonlinear dynamic equation on a cluttered image background. The brightness of the face was also changed several times during the translation to create a more complex scene. Fig. 1 illustrates how the face was superimposed on the cluttered scene. Note that the face locations in Fig. 1.a are not the results of simulation and are only given as an example. The particle-based driver's face tracker using the proposed observation model was then compared to similar trackers using popular color-based measurement methods. These methods include region-based, parametric (Gaussian-based) and non-parametric (histogram-based) skin distribution modeling methods [15]. Fig. 1.b shows the comparison results. Error is based on the absolute value of the Euclidean distance between the estimated and true face locations in pixels. The number of frames is 90 and the number of particles is 100. Maximum sample propagation area is the 40-pixel proximity of the previous estimated location of the head. This efficient combination of color information and illumination-invariant cumulative color ratios for the desired face and similar objects in the likelihood model's background imbues the proposed observation model with a minimal error value in respect to all previously mentioned observation models. The region-based observation model shows the highest error rate in the test, since this method generalizes skin distribution and has a high false positive rate that results in the higher possibility of distraction by the background clutter.

There is often a large inconsistency of performance between simulations or systems tested in the laboratory and systems tested in real-world environments due to the complexity of the latter. This study obtained video sequences of a real driving scenario to better explore the performance of the system in real-world situations. Fig. 2 considers a driving situation during the day where the brightness reflected on the driver's face changes rapidly (note the color changes in the driver's jacket and the ceiling of the vehicle). Fig. 2 demonstrates the comparison results of two particle-based trackers using different observation models — a histogram-based model and the proposed model. A color histogram with 16 bins is considered in this evaluation. After a rapid brightness change, colors on the driver's face shifted, which degrades the histogram-based system to the part of the face with less color variation while also setting it up to be easily distracted by background clutter if the background color



(a)



(b)

Figure 1. Comparison of the tracker's performance for different observation models. (a): Test bench (b): Simulation results

is close to the color of the histogram bins. In the proposed method, however, the ratio of colors is considered alongside the color information — the former being invariant to brightness changes. Fig. 3 shows a driver occluding his face with his hand. Although the colors of the hand and face are similar, the tracker is still able to distinguish and track the face. Other skin color pixels overlaid on the image demonstrate the efficiency of the skin color detector in detecting the face — only colors very close to the most appearing face color are selected. In cases of complete occlusion, estimation of the head location might not be exactly correct, but the tracker can recover from the complete occlusion due to the existence of multiple hypotheses.

Fig. 4 is an example of rapid and random head motion in different directions when the person jumps up and down. The proposed likelihood model — based on the combination of color ratio and efficient skin color modeling — assures successful tracking in situations where there are excessive orientation or scale changes in the object between two consecutive frames, with color remaining invariant to object translation and rotation. The color tracker's successful results proved the system's

ability to overcome the above challenges, especially in cases of gradual and drastic illumination variations.

An SR-3000 TOF sensor with the resolution of  $176 \times 144$  pixels was used for the depth tracker. The operating range was from 0.3 to 7.5 meters for this sensor, and the field of view (FOV) was  $47.5 \times 39.5$  degrees. The depth tracker was also successfully tested for large-scale changes such as pose and illumination variations, occlusions and rapid motions.

Fig. 5 shows the results of the tracker for pose and scale changes. Illumination variation does not affect the depth information, so the proposed tracker is completely insensitive to illumination changes. The color tracker, on the other hand, cannot perform well when the object does not appear clearly or is saturated by severe brightness.

The sequence of images shown in Fig. 6 estimate and track human body locations with estimation results depicted by a red dot. The depth tracker is evaluated under abrupt changes in the scale and poses of the objects, illustrated in Fig. 7. The depth tracker handles occlusions and rapid body translations perfectly, due to the inclusion of the target's depth information. The efficiency of the proposed depth tracker using TOF data is owing to the employment of the effective segmentation method described in Section IV, which improves the particle filter's importance sampling.

## VI. DISCUSSION

In situations where the lighting conditions are inadequate (including poor lighting and absence of light), depth information can be used to reliably detect the objects of interest. Color tracking methods fail to produce satisfactory results in the aforementioned situations because they use passive sensors that are explicitly dependent on the availability of environmental illumination. In depth tracking algorithms, segmentation can be performed easily using a depth classifier, which is more functional in cluttered scenes [16].

One of the most significant challenges in tracking scenarios is dealing with object occlusion. Depth information is not available from a single color camera, so such information from range sensors is very helpful in handling object occlusions as well as applications such as pedestrian detection for collision avoidance where information about the distance of the object — i.e., pedestrian — from the vehicle is required. Depth-based tracking is also useful because it is not sensitive to an object's rotation — unlike color-based tracking which requires multiple features such as intensity and color information to handle this sensitivity. In the case of severe background clutter, even the addition of edge information is not sufficient. Drastic illumination variations can interfere with 2-D tracking by degrading the tracked object's characteristics in intensity domains by introducing varying colors, shadows, etc. Depth sensors, however, are unaffected by changes in lighting conditions, which leaves the chosen features intact to be efficiently used in tracking.

Depth image on its own is only useful for object detection at the blob level because object details or micro-features cannot be determined based on range images alone (if not in close proximity). In this case, skin color information can be used for fast face segmentation using color images, followed by eye or mouth tracking in applications such as drowsy driver detection or face recognition.

TOF range sensors do tend to be more expensive than color cameras, however, and the resolution of the images provided by TOF sensors is often limited [16]. In this regard, they might give inaccurate information about the object's distance — putting them at a disadvantage when an application requires precise distance information, such as in the case of pedestrian detection. Another limitation of these sensors is their restricted non-ambiguity range. For a SR-3000 camera, for example, this range is limited to 7.5 m (at 20 MHz modulation). The final drawback of TOF range-imaging sensors is that they are only designed for operation in indoor lighting conditions. In outdoor applications, the TOF range imager only operates in the brightest background lighting conditions<sup>2</sup>.

## VII. CONCLUSION

This paper successfully investigates and addresses the advantages and drawbacks of employing color camera and TOF range sensor output for the vital process of image tracking. An efficient color tracker and an original depth tracker based on range data were both implemented. The results show that color sensors are more suitable for outdoor applications and purposes that require the detection of a blob's details while TOF sensors are more appropriate for object tracking in limited illumination and applications that require information about the distance between the object and the camera. Today's range sensors do not provide color information to our knowledge, but more efficient tracking in unconstrained environments is possible if the information from the two sources is combined. Such an arrangement efficiently handles common challenges like object occlusion and illumination variation while providing the detection of an object's details and exploiting useful color information.

## ACKNOWLEDGMENT

This research has been supported in part by the Canada Research Chair Program, AUTO21 NCE and the NSERC Discovery grant.

## REFERENCES

- [1] T. Oggier, F. Lustenberger, and N. Blanc, "Miniature 3d tof camera for real-time imaging," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 4021 NAI, pp. 212–216, 2006.
- [2] J. W. Weingarten, G. Gruener, and R. Siegwart, "A state-of-the-art 3d sensor for robot navigation," *2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, vol. 3, pp. 2155–2160, 2004.

<sup>2</sup><http://www.mesa-imaging.ch/publications.php>



Figure 2. Comparison between two different observation models when illumination varies drastically in two consecutive frames. top: based on the proposed likelihood model and bottom: based on color histogram



Figure 3. The proposed system's performance when face is occluded by hand

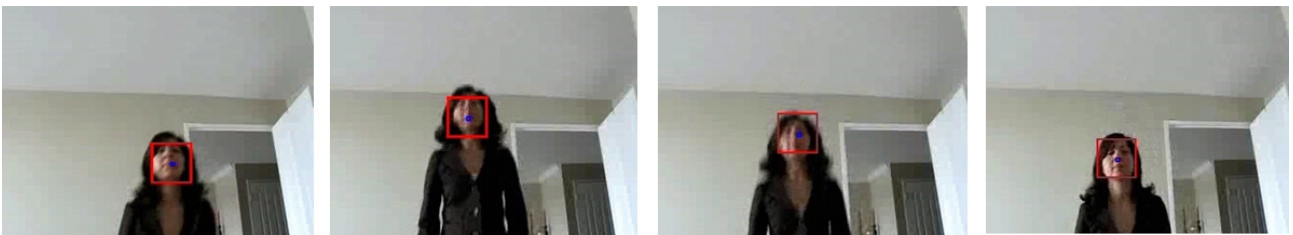


Figure 4. The proposed system's results for rapid head motion (frames 70, 76, 78, and 79)



Figure 5. Tracking results with pose and scale changes (frames 50, 61, 69, and 77)



Figure 6. Tracking results of the proposed depth tracker in case of pose changes and occlusion (frames 1, 7, 11, 18, 29, 40, 49, 57, 67, 72, 86, and 89)



Figure 7. Tracking results of the proposed depth tracker in case of occlusion, scale and pose changes (frames 50, 57, 66, 80, 92, 99, 103, 122, 127, 141, 180, and 193)

- [3] F. Xu and K. Fujimura, "Human detection using depth and gray images," *Proceedings IEEE Conference on Advanced Video and Signal Based Surveillance. AVSS 2003*, pp. 115–21, 2003.
- [4] S. B. Gokturk and C. Tomasi, "3d head tracking based on recognition and interpolation using a time-of-flight depth sensor," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. 211–217, 2004.
- [5] M. Isard and A. Blake, "Condensation - conditional density propagation for visual tracking," *International Journal of Computer Vision*, vol. 29, no. 1, pp. 5–28, 1998.
- [6] K. Nummiaro, E. Koller-Meier, and L. Van Gool, "An adaptive color-based particle filter," *Image and Vision Computing*, vol. 21, no. 1, pp. 99–110, 2003.
- [7] Y. Jin and F. Mokhtarian, "Towards robust head tracking by particles," *Proceedings - International Conference on Image Processing, ICIP*, vol. 3, pp. 864–867, 2005.
- [8] Y. Chen and Y. Rui, "Real-time speaker tracking using particle filter sensor fusion," *Proceedings of the IEEE*, vol. 92, no. 3, pp. 485–494, 2004.
- [9] F.-E. Ababsa, M. Mallem, and D. Roussel, "Comparison between particle filter approach and kalman filter-based technique for head tracking in augmented reality systems," *Proceedings - IEEE International Conference on Robotics and Automation*, vol. 2004, no. 1, pp. 1021–1026, 2004.
- [10] Y. Rui and Y. Chen, "Better proposal distributions: Object tracking using unscented particle filter," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. 786–793, 2001.
- [11] L. Sabeti and Q. M. J. Wu, "High-speed skin color segmentation for real-time human tracking," *IEEE International Conference on Systems, Man and Cybernetics*, pp. 2378–2382, 7-10 Oct. 2007.
- [12] D. Muselet and L. Macaire, "Fuzzy spatial ranks for object recognition across illumination changes," *2006 IEEE International Conference on Multimedia and Expo, ICME 2006 - Proceedings*, vol. 2006, pp. 985–988, 2006.
- [13] L. Sabeti and Q. M. J. Wu, "Illumination invariant visual tracking using particle filter," *EURASIP Journal on Image and Video Processing*, 2008, in Press.
- [14] E. Parvizi and Q. M. J. Wu, "Multiple object tracking based on adaptive depth segmentation," *Computer and Robot Vision, 2008. CRV '08. Fifth Canadian Conference on*, 2008, in Press.
- [15] V. Vezhnevets, V. Sazonov, and A. Andreeva, "A Survey on Pixel-Based Skin Color Detection Techniques," *Proc. Graphicon*, pp. 85–92, 2003.
- [16] T. Gandhi and M. M. Trivedi, "Pedestrian protection systems: Issues, survey, and challenges," *IEEE Transactions on Intelligent Transportation Systems*, vol. 8, no. 3, pp. 413–430, 2007.

**Leila Sabeti** is currently a Ph.D. candidate at the University of Windsor, Windsor, ON, Canada. She received her M.A.Sc. degree in electrical engineering from the University of Windsor in 2004.

She has published over 12 scientific papers in areas of computer vision, image processing and ASIC design. Her current research interests include image, video and signal processing and active video object localization and tracking.

Ms. Sabeti is a student member of IEEE and WIE organizations.

**Ehsan Parvizi** is currently a M.A.Sc. candidate at the University of Windsor, Windsor, ON, Canada. He received his B.A.Sc. degree in electrical engineering from the University of Tehran, Tehran, Iran, in 2003.

He has published two scientific papers in areas of computer vision and image processing. His research interests include real-time object detection and tracking, 3-D image analysis, machine learning, pattern recognition, image processing, multimedia, and biometrics.

**Q. M. Jonathan Wu** received his Ph.D. degree in electrical engineering from the University of Wales, Swansea, U.K., in 1990.

In 1995, he joined the National Research Council of Canada where he became a Senior Research Officer and Group Leader. He is currently a full Professor in the Department of Electrical and Computer Engineering, University of Windsor, Windsor, ON, Canada. He was recently named as the Canada Research Chair (CRC) in Automotive Sensors and Sensing Systems. He has published over 100 scientific papers in areas of computer vision, neural networks, fuzzy systems, robotics, micro-sensors and actuators, and integrated micro-systems. His current research interests include 3-D image analysis, active video object tracking and extraction, vision-guided robotics, sensor analysis and fusion, wireless sensor networks, and integrated micro-systems.

Dr. Wu is an Associate Editor for the IEEE Transaction on Systems, Man, and Cybernetics-Part A and is on the editorial board of the Journal of Control and Intelligent Systems.