Person Localization using a Wearable Camera towards Enhancing Social Interactions for Individuals with Visual Impairment

Lakshmi Gade, Sreekar Krishna and Sethuraman Panchanathan
Center for Cognitive Ubiquitous Computing (CUbiC)
School of Computing and Informatics
Arizona State University, Tempe AZ 85281, USA
Ph: 001 (732) 890 6271

lgade@asu.edu

ABSTRACT

Individuals with visual impairments are at a loss when it comes to everyday social interactions as majority (65%) of these interactions happen through visual non-verbal media. Recently, efforts have been made towards development of an assistive technology called the *Social Interaction Assistant* [14] which enables access to such useful cues so as to compensate for the lack of vision and other visual impairments. There have been studies which enumerate the important needs of such individuals when they interact in social situations. Along with feedback about their own social behavior, these studies indicate that individuals with visual disabilities are interested in a number of cues related to the people in their surroundings. In this paper, we discuss the importance of person localization while building a human-centric assistive technology which addresses the essential needs of the visually impaired users. Next, we describe the challenges that arise when a wearable camera setup is used as an input source in order to perform person localization. Finally, we present a computer vision based algorithm adapted to handle the issues that are inherent when such a wearable camera setup is used and demonstrate its performance on a number of example sequences.

Categories and Subject Descriptors

I.4.8 [**Image Processing and Computer Vision**]: Scene Analysis – *Tracking*

K.4.2 [**Computers and Society**] Social Issues – *Assistive technologies for persons with disabilities*

I.2.10 [**Artificial Intelligence**]: Vision and Scene Understanding – *Video Analysis*

General Terms

Algorithms, Performance, Experimentation, Human Factors, Theory

Keywords

Computer Vision, Object Tracking, Person Tracking, Social Interactions, Mobile Camera, Wearable Camera

1. INTRODUCTION

Recently, a new paradigm of computing termed as Human-Centered Multimedia Computing (HCMC) has emerged out of the concepts and fundamentals of Human Centered Computing (HCC) [29]. HCMC focuses on the creation of multimedia solutions that enrich everyday lifestyles of individuals through the effective use of multimedia technologies. One important aspect of HCMC, as explained by [30] focuses on deriving inspirations from human disabilities and deficits towards novel multimedia computing techniques. An important example of the same, discussed in detail in [14], is the concept of an Embodied Social Interaction Assistant which aims at developing an assistive technology aid for enhancing social interactions between individuals. A detailed evolution process of this project can be traced through the publications [14]-[17], [22]-[23], [29]-[30] in chronological order. This paper attempts at providing a solution to one persistent problem of tracking people through the primary sensing element, a wearable camera, of the conceptual Social Interaction Assistant. Following this section, we provide a brief overview of the social interaction assistant before getting into the particular issue of person localization that is of primary focus for this article.

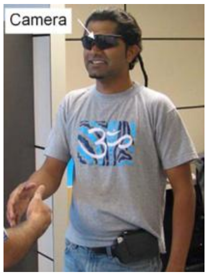


Figure 1. Visual Input Sensor for the Social Interaction Assistant

1.1 Social Interaction Assistant

Social interactions are a vital component of everyday living and encompass all forms of interpersonal communication between individuals and groups in social situations [14]. Apart from the explicit communication through words (also termed as verbal communication), a significant portion (≈ 65% [14]) of social interactions are influenced by non-verbal communication cues such as eye contact, facial expressions, hand gestures, body

MSIADU'09, October 23, 2009, Beijing, China.

posture, etc. The lack of access to such informative visual cues often inhibits individuals with visual impairments and blindness from effectively participating in day-to-day social interactions. The unique purpose of the Social Interaction Assistant is to bridge this *communication gap* between the users who are visually impaired and their sighted counterparts [14].

As shown in Figure 1, the Social Interaction Assistant makes use of a digital camera mounted on a pair of glasses to capture the visual scene in front of its users in an unobtrusive manner. The video stream thus captured is processed for the important social cues using a portable processing element connected through a USB port.

[14] introduces a systematic requirements analysis for an effective Social Interaction Assistant for aiding individuals with visual impairment and blindness. Based on an online survey (with inputs from 27 people, of whom 16 were blind, 9 had low vision, and 2 were sighted specialists in the area of visual impairment), the article provides a rank ordered list of important visual cues related to social interaction that are considered important by the target population. Most of the needs identified through this survey display the importance of extracting these following characteristics of individuals in the scene, namely, Number and location of the individuals, Facial expressions, Identity, Appearance, Eye Gaze direction, and Pose and Gestures. A brief glance through this list reveals the commonality of these issues with some of the important research questions being tackled by the computer vision and pattern recognition community.

Many advances have been made in order to extract information related to humans from videos. But, when the mobile setup of the Social Interaction Assistant is considered with real world data captured in unconstrained settings, a new dimension of complexity is added to these problems. As most of these cues are related to people in the surroundings of the user, it is essential to localize the individuals in the input video stream prior to processing for social interaction cues. The problem of person localization in general is very broad in its scope and wide varieties of challenges such as variations in articulation, scale, clothing, partial appearances, occlusions, etc make this a complex problem. Narrowing the focus, this paper targets person localization in real world video sequences captured from the wearable camera of the Social Interaction Assistant. Specifically, we focus on the task of localizing a person who is approaching the user to initiate a social interaction or just conversation. In this context, the problem of person localization can be constrained to the cases where the person of interest is facing the user.

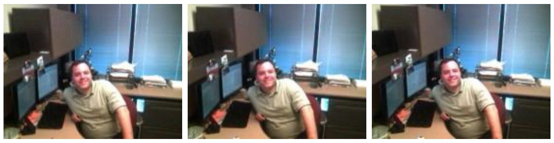


Figure 2. Person of interest at a short distance from camera

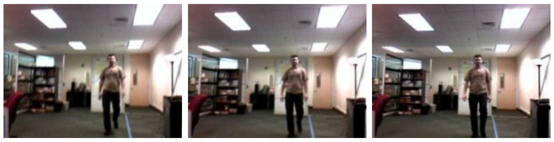


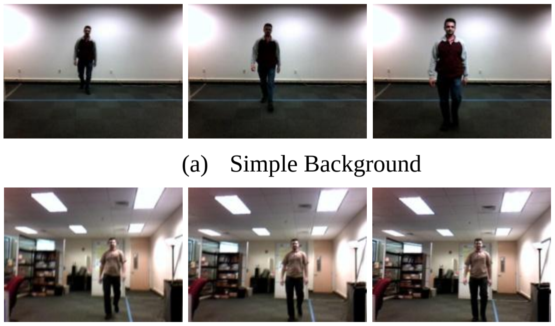
Figure 3. Person of interest at a large distance from camera

When such a person of interest is in close proximity, his/her presence can be detected by analyzing the incoming video stream for facial features (Figure 2). But when such a person is approaching the user from a distance, the size of the facial region in the video appears to be extremely small. In this case, relying on facial features alone would not suffice and there is a need to analyze the data for full body features (Figure 3). In this work, we have concentrated on improving the effectiveness of the Social Interaction Assistant by applying computer vision techniques to robustly localize people using full body features. Following section discusses some of the critical issues that are evident when performing person localization from a wearable camera setup of the Social Interaction Assistant.

1.2 Challenges in Person Localization from a wearable camera

As is the case with most computer vision based algorithms, the nature of the data has a significant influence on the performance of localization techniques. A number of factors associated with the background, object, camera/object motion, etc. determine the complexity of the problem. Following is a descriptive discussion of the imminent challenges that we encountered while processing the data using the Social Interaction Assistant.

1.2.1 Background Properties



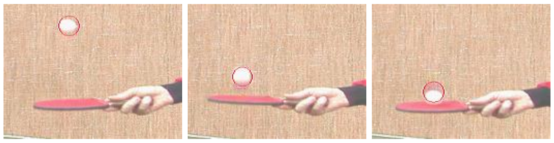
(a) Simple Background

(b) Complex Background

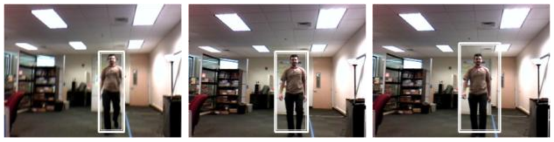
Figure 4. Background Properties

When the Social Interaction Assistant is used in natural settings, it is highly possible that there are objects in the background which move, thus causing the background to be dynamic. Also, there are bound to be regions in the background whose image features are highly similar to that of the person, thus leading to a cluttered background. Due to these factors, the problem of distinguishing the person of interest from the background becomes highly challenging in this context. Figure 4 (a) and (b) illustrate the contrast in the data due to the nature of the background

1.2.2 Object Properties



(a) Rigid, Homogeneous Object

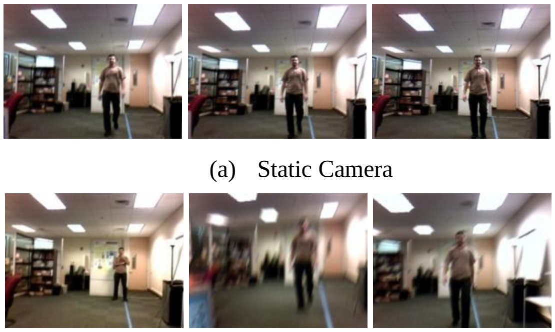


(b) Non-Rigid, Deformable, Non-Homogeneous Object

Figure 5. Object Properties

As we are interested in person localization, it can be clearly seen that the object is non-rigid in nature as there are appearance changes that occur throughout the sequence of images. Further, significant scale changes and deformities in the structure can also be observed. Also, when analyzing video frames of persons approaching the user, the basic image features in various sub-regions of the object vary vastly. For example, the image features from the facial region are considerably different from that of the torso region. Tracking detected persons from one frame to another will require individualized tracking of each region to maintain confidence. This non-homogeneity of the object poses to be a major hurdle while applying localization algorithms and has not been studied much in the literature. Figure 5(a) shows the simplicity of the data when these problems are not present, while Figure 5(b) highlights complex data formulations in a typical interaction scenario.

1.2.3 Object/Camera Motion



(a) Static Camera

(b) Mobile Camera

Figure 6. Object/Camera Motion

Traditionally, most computer vision applications use a static camera where strong assumptions of motion continuity and temporal redundancy can be made. But in our problem, as it is very natural for users to move their head continuously, the mobile nature of the platform causes abrupt motion in the image space. This is similar to the problem of working with low frame rate videos or the cases where the object exhibits abrupt movements. Recently, there has been an increase of interest in dealing with this issue in computer vision research [18][21][33][35]. Some important applications which are required to meet real-time constraints, such as teleconferencing over low bandwidth networks, and cameras on low-power embedded systems, along with those which deal with abrupt object and camera motion like sports applications are becoming common place [21]. Though solutions have been suggested, person localization through low frame rate moving cameras still remains an active research topic.

1.2.4 Other Important Factors Affecting Effective Person Localization

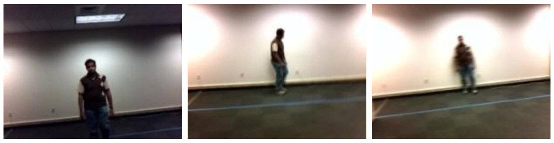


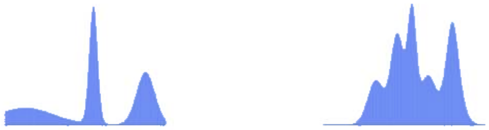
Figure 7. Changing Illumination, Pose Change and Blur

As the Social Interaction Assistant is intended to be used in uncontrolled environments, changing illumination conditions need to be taken into account. Further, partial occlusions, self occlusions, in-plane and out-of-plane rotations, pose changes, blur and various other factors can complicate the nature of the data (Figure 7).

Given the nature of this problem, in this paper we focus on the problem of robust localization of a single person approaching a user of the Social Interaction Assistant using full-body features. Issues arising due to cluttered background along with object and camera motion have been handled towards providing robustness. In the following section we discuss some of the important related work in the computer vision literature. The conceptual framework used in person localization is presented in Section 3. The details of the proposed algorithm are discussed in Section 4. Section 5 presents the results and discussions of the performance of our algorithm on videos collected from the Social Interaction Assistant. Finally, some possible directions of future work have been outlined followed by the conclusion.

2. RELATED WORK

Historically, two distinct approaches have been used by researchers for searching and localizing objects in videos. On one hand, there are detection algorithms which focus on locating an object using specific, spatial features in every frame. For example, haar-based rectangular features [36] and histograms of oriented gradients [9]. On the other hand, there are tracking algorithms which trail an object using generic image features once it is located by exploiting the temporal redundancy in videos. For example, color histograms [25] and edge orientation histograms [41].



(a) Detection based features (b) Tracking based features

Figure 8. Typical likelihoods using Tracking and Detection based features

Irrespective of whether detection or tracking is involved, reduced representations through image features are used to derive the probability that an object is present at a particular location in the image. Popularly termed as *likelihood*, this probability provides a weighting measure of confidence in the presence of an object at a particular location. If an image is represented as a 1-d signal, then the typical likelihood values at various points on it can be depicted as shown in Figure 8. Figure 8(a) shows a likelihood distribution using features generally used in detection algorithms while Figure 8(b) shows a likelihood distribution when tracking based features are used. Note how the detection likelihood is a sharp graph providing an easy opportunity to set thresholds when compared to the tracking likelihood. This is due fact that detection algorithms generally build object models by analyzing a predetermined set of training data. As the images in the video are searched for such well learned, structure-aware features, the likelihood values give a strong indication of the presence or absence of the object (Figure 8(a)). Downside of which is the difficulty in handling object deformability, non-rigidity and other such issues.

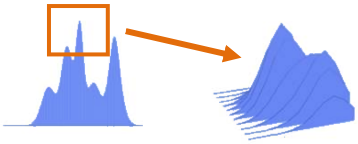


Figure 9. Temporal redundancy in Tracking Algorithms

On the contrary, tracking algorithms invariably use simple, structure-ignorant low-level image features. Due to this, the likelihood values alone cannot indicate the presence of the object with confidence. Therefore, these algorithms exploit the temporal redundancy in videos hoping that the object deformation and motion is not abrupt between frames. Once an object is located, it is followed efficiently by searching for appropriate simple image features in a reduced search space (Figure 9). Due to the use of simple, low level image features for tracking, these algorithms can easily handle non-rigidity, deformability, etc. to a large extent while clutter still remains a major issue.

Following is a brief discussion on various detection and tracking algorithms targeted towards the problem of person localization.

2.1 Detection Algorithms

As mentioned previously, detection algorithms exploit the specific, distinctive features of an object type and apply learning algorithms to detect a general class of objects. They use information related to the relative feature positions, invariant structural features, characteristic patterns and appearances [42] to locate objects within the gallery image. But, when the object is complex, like a person, it becomes difficult for these algorithms to achieve generality thereby failing even under minute non-rigidity. A number of factors such as variations in articulation, pose, clothing, scale and partial occlusions make this problem very challenging at the image processing level.

When assumptions about the background cannot be made, learning algorithms which take advantage of the relative positions of body parts are used to build classifiers. The kind of low-level features generally used in this context are gradient strengths and gradient orientations [9][11][45], edgelets [38], entropy[24] and haar-like features. Some of the well-known higher level descriptors are histogram of oriented gradients [9] and covariance features [32]. Efforts have been made to make these descriptors scale invariant as well [11].

In order to allow the above mentioned detectors to work within reasonable real-time assumptions, researchers have resorted to two kinds of approaches. One category include *part-based approaches* such as Implicit Shape Models [1] and constellation models [43] which place emphasis on detecting parts of the object followed by aggregating the part-based knowledge to infer the presence of the object, while the other category of algorithms tries to search for relevant descriptors for the whole object in a cascaded manner [36][45].

Shape-based Chamfer matching [2] is a popular technique used in multiple ways for person detection as the silhouette gives a strong indication of the presence of a person. In recent times, Chamfer matching has been used extensively by the person detection and localization community. It has been applied with hierarchically arranged templates to obtain the initial candidate detection blocks so that they can be analyzed further by techniques such as segmentation, neural networks, etc [4][44]. It has also been used as a validation tool to overcome ambiguities in detection results obtained by the Implicit Shape Model technique [20].

2.2 Tracking Algorithms

Assuming that there is temporal object redundancy in the incoming videos, many algorithms have been proposed to track objects over frames and build confidence as they go. Generally they make the simplifying assumption that the properties of the object depend only on its properties in the previous frame, i.e. the evolution of the object is a Markovian process of first order. Based on these assumptions, a number of deterministic as well as stochastic algorithms have been developed.

Deterministic algorithms usually apply iterative approaches to find the best estimate of the object in a particular image in the video sequence [41]. Optimal solutions based on various similarity measures between the object template (built from a single object image or a class of object images) and regions in the current image, such as sum of squared differences (SSD), histogram-based distances, distances in eigenspace and other projected spaces and conformity to particular object models, have been explored [41]. MeanShift [8] is a popular, efficient optimization-based tracking algorithm which has been widely used.

Stochastic algorithms use the state space approach of modeling dynamic systems and formulate tracking as a problem of probabilistic state estimation using noisy measurements [1]. In the context of visual object tracking, it is the problem of probabilistically estimating the object’s properties such as its location, scale and orientation by efficiently looking for appropriate image features of the object. Most of these stochastic algorithms perform Bayesian filtering at each step for tracking, i.e. they predict the probable state distribution based on all the available information and then update their estimate according to the new observations. Kalman filtering [37] is one such algorithm which fixes the type of the underlying system to be linear with Gaussian noise distributions and analytically gives an optimal estimate based on this assumption. As most tracking scenarios do not fit into this linear-gaussian model and as analytic solutions for non-linear, non-gaussian systems are not feasible, approximations to the underlying distribution are widely used from both parametric and non-parametric perspective. Some of the popular parametric algorithms include Extended Kalman Filtering (EKF) [37] and Unscented Kalman Filtering (UKF) [13], while non-parametric algorithms include the plethora of different particle filters that have been proposed in recent times [43].

Sequential monte-carlo based Particle Filtering techniques have gained a lot of attention recently. These techniques approximate the state distribution of the tracked object using a finite set of weighted samples using various features of the system. For visual object tracking, a number of features have been used to build different kinds of observation models, each of which have their own advantages and disadvantages. Color histograms [25], contours [12], appearance models, intensity gradients [5], region covariance [28], texture, edge-orientation histograms, haar-like rectangular features [41] and various combinations of these features fused in different ways [39][40][41] are a few examples. Apart from the kind of observation models used, this technique allows for variations in the filtering process itself. A lot of work has gone into adapting this algorithm to better perform in the context of visual object tracking.

While both the areas of detection and tracking have been explored extensively, there is an impending need to address some of the issues faced by low frame rate visual tracking of objects. Especially in the case of Social Interaction Assistant, person localization in low frame rate video is of utmost importance. In our work, we have attempted to modify the color histogram comparison based particle filtering algorithm to handle the complexities that occur when the camera is mobile on the Social Interaction Assistant.

3. CONCEPTUAL FRAMEWORK

As discussed in the previous section, detection and tracking offer distinctive advantages with a related set of associated problems. In our case, due to the movement of the camera, tracking algorithms tend to drift away due to the lack of temporal redundancy, while detection algorithms fail to localize individuals in front of the user in most of the frames due to the complex nature of the object. Though there are clear advantages in applying these techniques individually, the strengths of both these approaches need to be combined in order to tackle the challenges posed by the complex setting of the Social Interaction Assistant. In the past, a few researchers have approached the problem of tracking in low frame rate or abrupt videos by mixing the vivid advantages of detection and tracking. In particular, techniques have been proposed to improve the performance of standard particle filtering algorithm by utilizing independent object detectors [21][27]. Unfortunately, in our case, detection algorithms are not guaranteed to provide reliable results on most of the frames. Thus, such approaches need to be improved so as to incorporate the deterministic nature of the detection algorithms while avoiding explicit use of pre-trained detectors. In our experience, the Social Interaction Assistant offers a weak temporal redundancy in most cases. We exploit this information trickle between frames to get an approximate estimate of the object location.

Due to flexibility in the design and the breadth of their scope, particle filtering algorithms provide a good platform to address the issues arising due to complex data. These algorithms give an estimate of an object's position by discretely building the underlying distribution which determines the object's properties. But, real-time constraints impose limits on the number of particles and the strength of the observation models that can be used. This generally causes the final estimate to be noisy when conventional particle filtering approaches are applied. Unless the choice of the particles and the observation models fit the underlying data well, the estimate is likely to drift away as the tracking progresses. Also, most conventional observation models do not account for issues such as clutter and non-homogeneity of objects and thus lead to incorrect weighting of particles. Even in the presence of these problems, if the motion of the object can be modeled correctly, good performance can be achieved through particle filtering. But in our application, due to the relative motion of the camera, the object (or person) exhibits abrupt motion changes in the image plane and so it is not possible to model the motion accurately. Due to these limitations, the design of conventional particle filtering algorithms need to be altered to handle the challenges inherent in the data obtained using the Social Interaction Assistant.

To mitigate the problems faced in the use of the Social Interaction Assistant, we propose a new particle filtering framework that gets an initial estimate of the person's location by spreading particles over a reasonably large area and then successively corrects the position though a deterministic search in a reduced search space. Termed as Structured Mode Searching Particle Filter (SMSPF), the algorithm uses color histogram comparison in the particle filtering framework at each step to get an initial estimate which is then corrected by applying a structured search based on gradient features and chamfer matching. The details of this algorithm are described in the next section.

4. STRUCTURED MODE SEARCHING PARTICLE FILTER

The goal of this particle filtering framework is focused on tracking a single person under the following circumstances, namely

- Image region with the person is non-rigid and non-homogeneous
- Image region with the person exhibits significant scale changes
- Image region with the person exhibits abrupt motions of small magnitude in the image space due to the movement of the camera.
- Background is cluttered

It is assumed that an independent person detection algorithm will initialize the SMSPF algorithm in the complete system.

4.1 Particle Filtering Framework

Particle filters discretely achieve Bayesian filtering at each step. As the person of interest can exhibit abrupt motion changes in the image space, it is extremely difficult to model the placement of the person in the current image based on the previous information. Essentially, the state of each particle's position becomes independent of its state in the previous step. Thus, the prior distribution can be considered to be a uniform random distribution over the support region of the image.

$$p(x_t^i \mid x_{t-1}^i) = p(x_t^i) \qquad (1)$$

As it is essential for particle filtering algorithm to choose a good set of particles, it would be useful to pick a good portion of them near the estimate in the previous step. By approximating this previous estimate to be equivalent to a measurement of the image region with the person in the current step, the proposal distribution of each particle can be chosen to be dependent only on this measurement

$$q(x_t^i \mid x_{t-1}^i z_t) = q(x_t^i \mid z_t) \qquad (2)$$

Though the propagation of information through particles is lost by making such an assumption, it gives a better sampling of the underlying system. We employ a large variance Gaussian with its mean centered at the previous estimate for successive frame particle propagation. By using such a set of particles, a larger area is covered, thus accounting for abrupt motion changes and a good portion of them are picked near the previous estimate, thus exploiting the weak temporal redundancy. With the above alone, due to the small number of particles which are spread widely across the image, typical observations give an approximate location of the person. When such an estimate partially overlaps with the desired object, the best match occurs between the intersection of the estimate and the object region as shown in Figure 10. But, it is not trivial to detect this partial presence of the object due to the presence of background clutter. Many of the image properties between the background and the object might match thereby leading to bad particle locations. To counter this problem, we propose to use efficient image feature representation of the desired object through integral histograms [11] and employ an efficient search around the estimate to accurately localize the object.

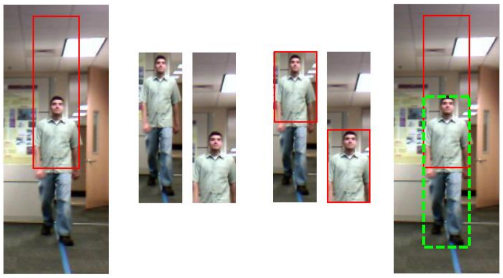


Figure 10. Structured Search

4.2 Structured Search

As the estimate obtained using widely spread particles gives the approximate location of the object, the search for the image block with a person in it can be restricted to a region around it. We have employed a grid-based approach to discretely search for the object of interest (a person) instead of checking at every pixel. By dividing the estimate into an *m x n* grid and sliding a window along the bins of the grid as shown in Figure 11, the search space can be restricted to a region close to the estimate. By finding the location which gives the best match with the person template, we can localize the person in the video sequence with better accuracy.

If this search is performed based on scale-invariant features, then it can be extended to identify scale changes as well. In order to achieve search over scale, the estimate and the sliding window need to be divided into different number of bins. If the search is performed using smaller number of bins as compared to the estimate, then shrinking of the object can be identified while searching with higher number of bins can account for dilation of the object. For example, if a *(m-1) x (n-1)* grid is used with the sliding window while a *m x n* grid is used with the estimate, then the best match will find a shrink in the object size. Similarly if a *m x n* grid sliding window is used with a *(m-1) x (n-1)* estimate grid, then dilations can be detected. It can be seen that this search is characterized by the number of bins *m x n* into which the sliding window and the estimate are divided. Based on the nature of the problem, the number of bins and the amount of sweep across scale and space can be adjusted to suit the problem. Currently, these parameters are being set manually. This structured search framework can be extended to include online algorithms which can adapt the number of grid bins based on the evolution of the object.

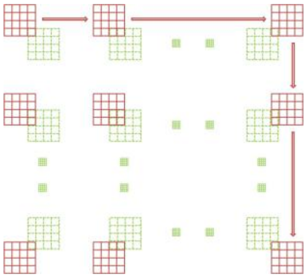


Figure 11. Sliding window of the Structured Search (Green: Estimate; Red: Sliding window).

4.3 Problems with Search across Scale

In the ideal case, we should be able to obtain the best match across space and scale by simple feature comparison with the person template. But, when global scale invariant features such as histograms are used for comparison, there are problems associated with search across scale. These are discussed in detail in [7]. One major problem is that such features cannot distinguish between the complete object and similar-featured sub-regions. In fact, such features tend to give a better match with smaller sub-regions as they contain lesser noise. [7] proposes a technique to robustly track *blobs* across scale while handling this issue. But, this cannot be applied directly in our case as our object of interest is not a blob and instead is a non-homogeneous object moving against a cluttered background. Further, it is also possible that the image region with the person combined with a portion of the cluttered background results in a higher match score with the template. In our approach, we have followed the underlying principle of enforcing the need to enclose the boundary of the object by a good match as in [7]. To deal with the complex nature of the data, we have introduced a chamfer-matching based technique to robustly search for persons across scale.

4.4 Chamfer Matching in Structured Search

Chamfer matching gives a measure of confidence that indicates the presence of the person based on silhouette information. We incorporate this confidence into the structured search discussed above in order to detect the precise location of the person around the particle filter estimate. An edge map of the image under consideration is first obtained which is then divided into (mxn) windows similar to the structured search and an elliptical ring mask is then applied to each of these windows as shown in Figure 12. This mask is applied so as to eliminate the edges that arise due to clothing and background thereby emphasizing the boundary edges which are likely to appear in the ring region if a window is precisely placed on the object perimeter. A distance transformed image of the window is then obtained using the masked edges. A distance transform is nothing but a mapping which represents the distance of each pixel to the nearest edge (See Figure 12).

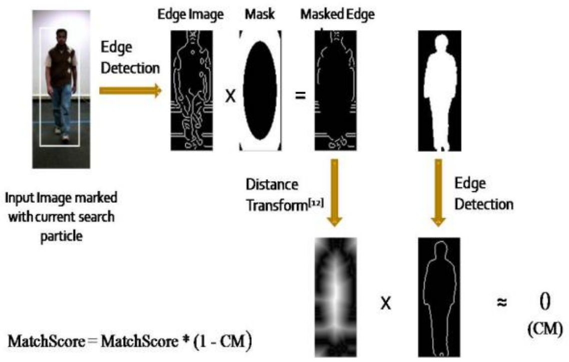


Figure 12. Incorporating Chamfer Matching into Structured Search

By applying chamfer matching with a generic person contour resized to the current particle filter estimate size, a confidence number in locating the desired object within the image region can be obtained. A value close to 0 indicates a strong confidence of the presence of a person and vice versa. As 1 is the maximum value that can be obtained by the chamfer match, this measure can be incorporated into the match score of the structured search using the following equation

$$MatchScore = MatchScore * (1 - ChamferMatch) \qquad (3)$$

By incorporating the knowledge about the contour of the person into the structured search, robust localization of the person can be achieved by searching across space and scale.

As in [25], we have employed this technique using HSV color histogram comparison to get likelihoods at each of the particle locations. Since intensity is separated from color in this color space, it is reasonably insensitive to illumination changes. The histograms are compared using the well-known Bhattacharyya Similarity Coefficient which guarantees near optimality and scale invariance. For person detection and tracking, gradient based features have been widely used and their applicability has been strongly established by various algorithms like Histogram of Oriented Gradients (HoGs) [9]. Here we have used Edge Orientation Histogram (EOH) based structured search built using integral histograms in order to perform the structured search in near real-time with Chamfer Matching embedded efficiently into the searching framework

5. EXPERIMENTS AND RESULTS

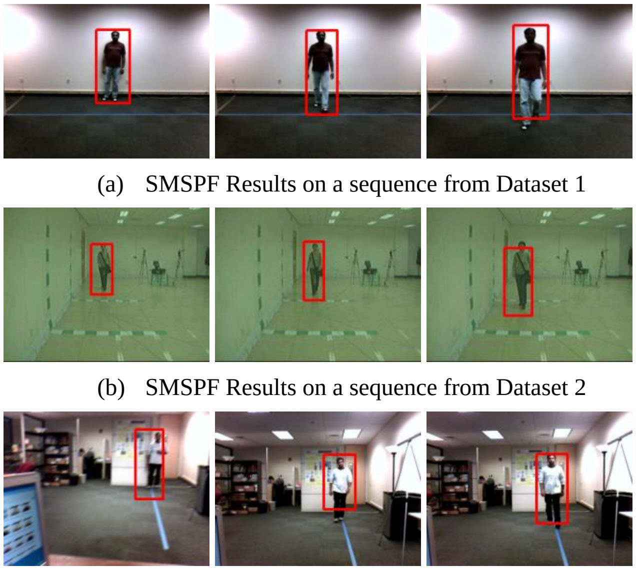
5.1 DataSets

The performance of the structured mode searching particle filter (SMSPF) has been tested using three datasets where a single person faces the camera while approaching it. There are significant scale changes in each of these sequences. Further, non-rigidity and deformability of the person region can also be clearly observed. Different scenarios with varying degrees of complexity of the background and camera movement have been considered. Following is a brief description of these datasets.

(a) DataSet 1 (Collected at CUbiC[1]) : Plain Background; Static Camera; 320x240 resolution

(b) DataSet 2 (CASIA[2] Gait Dataset B with subject approaching the camera [4]) : Slightly cluttered Background; Static Camera; 320x240 resolution

(c) DataSet 3 (Collected at CUbiC) : Cluttered Background; Mobile Camera; 320x240 resolution



(a) SMSPF Results on a sequence from Dataset 1

(b) SMSPF Results on a sequence from Dataset 2

(c) SMSPF Results on a sequence from Dataset 3

Figure 13. SMSPF Results

Figure 13 shows the sample results on each of the datasets used.

5.2 Evaluation Metrics

In order to test the robustness of this algorithm and the applicability in complex situations, its performance has been compared with the Color Particle Filtering algorithm [25]. Assuming that a detection algorithm can detect persons in at least some frames, the image region containing the person in each of the test sequences has been manually set. The following two criteria have been used to evaluate their performance [3].

- Area Overlap (A0)
- Distance between Centroids (DC)

Manually labeled rectangular regions around the person in the image have been used as the ground truth for this purpose. Suppose *gTruth$_i$* is the ground truth in the *ith* frame and *track$_i$* is the rectangular region output by a tracking algorithm, then the area overlap criterion is defined as follows

$$AO(gTruth_i, track_i) = \frac{Area(gTruth_i \cap track_i)}{Area(gTruth_i \cup track_i)} \qquad (4)$$

The average area overlap can be computed for each data sequence as

$$AvgAOR = \frac{1}{N}\sum_{i=1}^{N} AO_i \qquad (5)$$

[1] Center for Cognitive Ubiquitous Computing

[2] Portions of the research in this paper use the CASIA Gait Database collected by Institute of Automation, Chinese Academy of Sciences

Similar to [3], we use Object Tracking Error (OTE) which is the average distance between the centroid of the ground truth bounding box and the centroid of the result given by a tracking algorithm

$$OTE = \frac{1}{N}\sum_{i=1}^{N}\sqrt{(Centroid_{gTruth_i} - Centroid_{track_i})} \qquad (7)$$

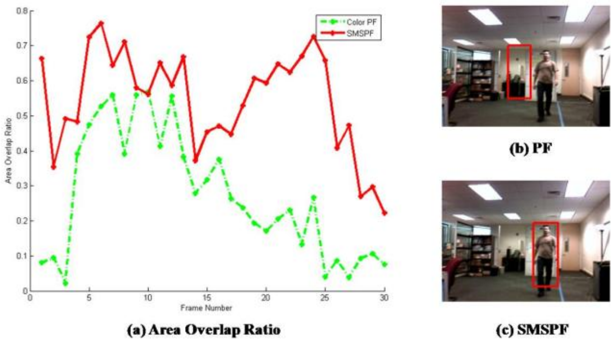
In order to evaluate the performance of these algorithms using a single metric which encodes information from both area overlap and the distance between centroids, we have used a measure termed as the Tracking Evaluation Measure (TEM) which is the harmonic mean of the average area overlap fraction (AvgAOR) and a non-linear mapping of the Object tracking error (OTE).

$$TEM = 2*\frac{AvgAOR*e^{-kOTE}}{AvgAOR+e^{-kOTE}} \qquad (8)$$

where k is a constant which exponentially penalizes the cases where the distance between centroids is large.

5.3 Results

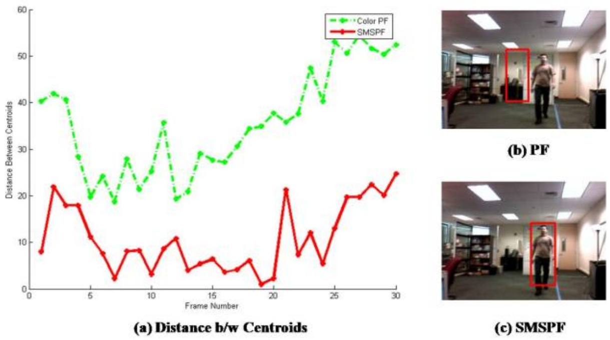
Particle Filtering has been widely used to handle complex scenarios by maintaining multiple hypotheses. As mentioned in [21], in order to handle abrupt motion changes, it is essential that the particles are widely spread while tracking. Following this principle, we have compared the performance of color particle filter (PF) [25] and the structured mode searching particle filter (SMSPF) by using a 2-D Gaussian with large variance as the system model. The position of the person and its scale have been included in the state vector. In order to compensate for the computational cost of structured search, only 50 particles were used for the SMSPF algorithm while 100 particles were used for the PF algorithm. A 10x10 grid with a sweep of 8 steps along the spatial dimension and 3 steps along the scale dimension were incorporated in the structured search.



(a) Area Overlap Ratio (b) PF (c) SMSPF

Figure 14. AO (Dotted Line: Color PF; Solid Line: SMSPF)

Figure 14 and Figure 15 illustrate the comparison of the area overlap ratio and the distance between centroids at each frame of an example sequence. The sample frames are shown beside the tracking results. From Figure 14(a), it is evident that the SMSPF algorithm (red) shows a significant improvement over the color particle filter algorithm (green). Here, the area overlap ratio using SMSPF is much closer to 1 in most of the frames while the color particle filter drifts away causing this measure to be closer to 0. The distance between centroids measure also indicates a greater precision of the SMSPF algorithm as seen in Figure 15(a) where the distance between centroids using color particle filter is much higher than that with SMSPF(≈0).



(a) Distance b/w Centroids (b) PF (c) SMSPF

Figure 15. DC (Dotted Line: Color PF; Solid Line: SMSPF)

Figure 16, Figure 17 and Figure 18 show the Tracking Evaluation Measure (TEM) for Datasets 1, 2 and 3. In majority of the cases, the SMSPF algorithm outperforms the color particle filtering algorithm with a higher TEM score.

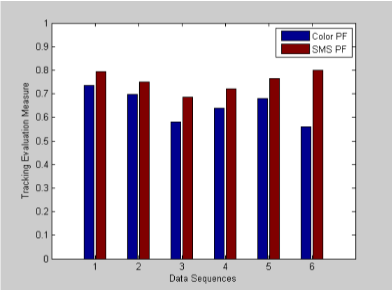


Figure 16. Evaluation Measure for DataSet 1

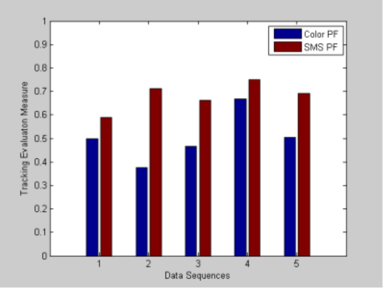


Figure 17. Evaluation Measure for DataSet 2

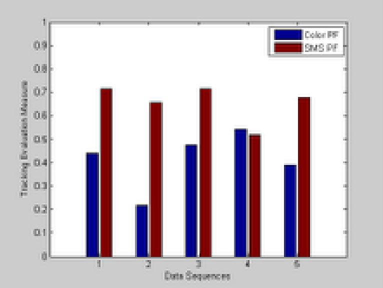


Figure 18. Evaluation Measure for DataSet 3

The results presented as a comparison between Color PF and SMSPF shows that incorporating a *deterministic* structured search into the *stochastic* particle filtering framework improves the person tracking performance in complex scenarios. The SMSPF algorithm strikes a balance between *specificity* and *generality* offered by detection and tracking algorithms as discussed in Section 2. It uses specific structure-aware features in the search in order to handle non-homogeneity of the object and the cluttered nature of the background. On the other hand, generality is maintained by using simple, global features in the particle filtering framework so as to handle non-rigidity and deformability of the object. The clear advantage of using the structured search can be observed on the complex Dataset 3 which encompasses most of the challenges generally encountered while using the Social Interaction Assistant.

6. FUTURE WORK

As a first step towards achieving robust person localization in the Social Interaction Assistant platform, we have currently considered the cases where the movement of the camera is small. The generic structured search proposed in this work can be adapted to handle drastic abrupt motions of the camera as well. One way to handle such cases is to use a very small set of particles spread over a large region in conjunction with the structured search at each particle region. Also, improving the efficiency of the observation models would computationally ease such near-exhaustive searches. Further, in this work, we used a generic person silhouette in our chamfer matching step to validate the positions in the structured search. Better validation can be obtained by using person dependent silhouettes and better boundary masks which accurately capture the relevant structure of the person's body. The current implementation has been focused only towards people facing the camera. This can be readily extended to handle other cases by effectively selecting the relevant silhouettes based on the application.

7. CONCLUSION

Person localization in videos captured from a wearable camera involves tracking non-rigid, deformable, non-homogeneous image regions which exhibit random motion patterns in cluttered backgrounds. By incorporating ideas of specificity associated with deterministic detection algorithms along with the generality of stochastic tracking algorithms, we have presented a particle filtering technique which effectively localizes individuals across a range of space and scale once a person is detected. This technique is useful in achieving person localization in videos captured using any mobile camera platform where there is low temporal redundancy between frames. Our immediate application being the wearable Social Interaction Assistant, which aims to enhance the everyday social interaction experience of the visually impaired, we have been able to achieve near real-time person localization.

8. REFERENCES

[1] Arulampalam, S. , Maskell, S. , Gordon, N. , Clapp, T. 2002. A Tutorial on Particle Filters for On-line Non-linear/Non-Gaussian Bayesian Tracking. IEEE Transactions on Signal Processing, vol. 50, 174.188.

[2] Barrow, H.G., Tenenbaum, J.M., Bolles, R.C., Wolf, H.C. 1977. Parametric correspondence and chamfer matching: Two new techniques for image matching. In proceedings of the 5th International Joint Conference on Artificial Intelligence. Cambridge, MA, 659-663

[3] Bashir, F. , Porikli, F. 2006. Performance evaluation of object detection and tracking systems. In IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS2006) .

[4] Bertozzi, M. et al. 2003. Shape-based pedestrian detection and localization. In Proceedings of IEEE International Conference on Intelligent Transportation Systems, vol. 1, 328–333.

[5] Birchfield, S. 1998. Elliptical Head Tracking Using Intensity Gradients and Color Histograms. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 232.

[6] CASIA, CASIA Gait Database, http://www.sinobiometrics.com

[7] Collin, R. 2003. Mean-Shift blob tracking through scale space. In proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol 2, 234-240.

[8] Comaniciu, D. , Ramesh, V. , Meer, P. 2000. Real-time tracking of non-rigid objects using mean shift. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, vol 2, 142-149 .

[9] Dalal, N., Triggs, B. 2005. Histograms of Oriented Gradients for Human Detection. In Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 893, 88.

[10] Fergus, R., Perona, P., Zisserman, A.2003. Object class recognition by unsupervised scale-invariant learning. In proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. vol 2, 264-271

[11] He, N., Cao, J., Song, L. 2008. Scale Space Histogram of Oriented Gradients for Human Detection. International Symposium on Information Science and Engineering, 167-170.

[12] Isard, M. , Blake, A. 1998. Condensation-conditional density propagation for visual tracking. International Journal of Computer Vision, vol. 29, 5-28.

[13] Julier, S. , Uhlmann, J.1997. A New Extension of the Kalman Filter to Nonlinear Systems. In Proceedings of *SPIE*, vol. 3068, 182-193.

[14] Krishna, S., Colbry, C., Black, J., Balasubramanian, V., Panchanathan, S. 2008. A Systematic Requirements Analysis and Development of an Assistive Device to Enhance the Social Interactions of People Who are Blind or Visually Impaired, Marseille, France.

[15] Krishna, S., Little, G., Black, J., Panchanathan S. 2005. A wearable face recognition system for individuals with visual impairments. In Proceedings of the 7th international ACM SIGACCESS conference on Computers and accessibility, 106-113.

[16] Krishna, S., Little, G., Black, J., Panchanathan, S.2005. iCARE interaction assistant: a wearable face recognition system for individuals with visual impairments. ASSETS 2005, 216-217

[17] Krishna, S., Panchanathan, S. 2009. On the detection of stereotypic body mannerisms using embodied motion

sensors. IEEE Transactions on Systems, Man, Cybernetics-Part C, vol. Submitted, April 2009.

[18] Kwon, J., Lee K.M.2008. Tracking of Abrupt Motion Using Wang-Landau Monte Carlo Estimation. In proceedings of the 10th European Conference on Computer Vision, Part 1, 387-400.

[19] Leibe, B., Leonardis, A., Schiele, B. 2004.Combined object categorization and segmentation with an implicit shape model. In ECCV workshop on statistical learning in computer vision, 17–32.

[20] Leibe, B., Seemann, E., Schiele, B. 2005. Pedestrian Detection in Crowded Scenes. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 1.

[21] Li, Y. , Ai, H. , Yamashita, T. , Lao, S. , Kawade, M. 2007. Tracking in Low Frame Rate Video: A Cascade Particle Filter with Discriminative Observers of Different Lifespans. In Proceedings of International Conference on Computer Vision and Pattern Recognition, 1-8.

[22] McDaniel, T., Krishna, S., Balasubramanian, V., Colbry, D., Panchanathan, S. 2008. Using a haptic belt to convey non-verbal communication cues during social interactions to individuals who are blind. Haptic, Audio and Visual Environments and Games, 2008. HAVE 2008, IEEE International Workshop on, 2008.

[23] McDaniel, T., Krishna, S.,Colbry, D., Panchanathan, S., 2009,. Using tactile rhythm to convey interpersonal distances to individuals who are blind. In Proceedings of the 27th international conference extended abstracts on Human factors in computing systems, Boston, MA, USA: ACM, 2009, pp. 4669-4674.

[24] Meng, L., Li, L., Mei, S., We, W.2008. Directional Entropy Feature for Human Detection. In Proceedings of the 19th International Conference on Pattern Recognition. ICPR 2008.

[25] Nummiaro, K., Koller-Meier, E. , Gool, L.V. 2002. A colorbased particle filter.

[26] Ogale, N.A. 2006. A survey of techniques for human detection from video. Survey, University of Maryland.

[27] Okuma, K. , Taleghani, A. , Freitas, N.D. , Little, J.J. Lowe, D.G. 2004. A Boosted Particle Filter: Multitarget Detection and Tracking. *Computer Vision - ECG 2004*, 28-39.

[28] Palaio, H., Batista, J. 2008. A region covariance embedded in a particle filter for multi-objects tracking. VS08.

[29] Panchanathan, S.; Krishna, S.; Black, J.A.; Balasubramanian, V., 2008, Human Centered Multimedia Computing: A New Paradigm for the Design of Assistive and Rehabilitative Environments, International Conference on Signal Processing, Communications and Networking,. ICSCN '08.

[30] Panchanathan, S., Krishna, S., McDaniel, T., Balasubramanian, V. 2008. Enriched human-centered multimedia computing through inspirations from disabilities and deficit-centered computing solutions. In Proceeding of the 3rd ACM international workshop on Human-centered computing, Vancouver, British Columbia, Canada: ACM, 35-42.

[31] Porikli, F. 2005. Integral histogram: A fast way to extract histograms in Cartesian spaces. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, vol 1, 829–836.

[32] Porikli, F., Meer, P. , Tuzel, O. 2007. Human detection via classification on riemannian manifolds. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 1–8.

[33] Porikli, F. , Tuzel., O. 2005. Object Tracking in Low-Frame-Rate Video. SPIE Image and Video Communications and Processing, vol. 5685, 72-79.

[34] Porikli, F. , Tuzel, O., Meer, P. 2006. Covariance tracking using model update based means on Riemannian manifolds. In Proceedings of IEEE Conference. On Computer Vision and Pattern Recognition.

[35] Philomin, V., Duraiswami, R., Davis, L.2000. Qausi-Random Sampling for Condensation. In proceedings of the 6th European Conference on Computer Vision, Part II, 134-149.

[36] Viola, P. , Jones, M.J. 2004. Robust Real-Time Face Detection. International Journal of Computer Vision, vol 57, 137-154.

[37] Welch, G. , Bishop, G. 2001. An introduction to the kalman filter. Technical Report, University of North Carolina at Chapel Hill.

[38] Wu, B., Nevatia, R. 2005. Detection of multiple, partially occluded humans in a single image by Bayesian combination of edgelet part detectors. In Proceedings of the Tenth IEEE International Conference on Computer Vision, vol 1, 90-97.

[39] Wu, Y. , Huang, T.S. 2004. Robust Visual Tracking by Integrating Multiple Cues Based on Co-Inference Learning. International Journal on Computer Vision, vol. 58, 55-71.

[40] Xu, X. , Li, B. 2005. Head Tracking Using Particle Filter with Intensity Gradient and Color Histogram. IEEE International Conference on Multimedia and Expo, 888-891.

[41] Yang, C., Duraiswami, R. , Davis, L. 2005. Fast Multiple Object Tracking via a Hierarchical Particle Filter. In Proceedings of the Tenth IEEE International Conference on Computer Vision, vol 1.

[42] Yang, M. et al. 2002. Detecting faces in images: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, 34–58.

[43] Yilmaz, A., Javed, O., Shah, M. 2006. Object Tracking:Survey. In ACM Computing Surveys (CSUR), vol 38.

[44] Zhao, L. , Davis, L. 2005. Closely Coupled Object Detection and Segmentation. In Proceeding s of International Conference on Computer Vision, 454-461.

[45] Zhu, Q. , Yeh, M., Cheng, K., Avidan, S. 2006. Fast Human Detection Using a Cascade of Histograms of Oriented Gradients. In Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1491-1498.