Activity maps for location-aware computing

D. Demirdjian, K. Tollmar, K. Koile, N. Checka and T. Darrell
MIT Artificial Intelligence Laboratory, Cambridge MA 02139
demirdji@ai.mit.edu

Abstract

Location-based context is important for many applications. Previous systems offered only coarse room-level features or used manually specified room regions to determine fine-scale features. We propose a location context mechanism based on activity maps, which define regions of similar context based on observations of 3-D patterns of location and motion in an environment. We describe an algorithm for obtaining activity maps using the spatio-temporal clustering of visual tracking data. We show how the recovered maps correspond to regions for common tasks in the environment and describe their use in some applications.

1 Introduction

Location can provide context for many important applications. When specifying a device such as a printer or display, users would typically like a system to know which device is closest or most easily viewed. A user may request that email and other messages arrive at the place where the user physically is, and that notification be consistent with the task occurring in the space. Users may wish to have music or other media follow them as they move in the environment using the most appropriate display resources. For each of these tasks, location context information is important. [15]

Simply considering the instantaneous 3-D location of users is useful, but alone is insufficient as context information. Applications have to generalize context information from previous experience, and an application writer would like to access categorical context information, such as what activity a user is performing. In addition, other features such as motion and shape (configuration) of the user are often important to distinguish activity: contrast a person walking past a desk with a person sitting at that desk.

While location cues alone can't fully determine what objects or tasks are being used in a particular activity, we have found that activities are correlated with location cues. By looking for patterns in these location cues, we can infer activity behavior. We attempt to find an "activity map", which divides a physical space based on observed location features (location, motion, shape, ...) into regions corresponding to activities or sets of activities.

Previous approaches have partitioned space based on simple proximity or relied on user specified maps for regions. In contrast, we argue that location regions should be learned from observed activity, including motion and shape cues as well as position. Regions can overlap in space, since motion or shape can indicate a different activity.

In this paper we describe an algorithm for computing location context based on 3-D person tracking techniques and the use of automatically generated activity maps. Our system is robust to many of the issues that often plague computer vision systems, such as dynamic illumination or fast motions. We form activity regions using a spatio-temporal clustering method and use the resulting regions to define an activity map. This map is used at run time to contextualize user preferences, e.g., allowing "location-sticky" settings for messaging, environmental controls, and/or media delivery.

In the following sections, we review related previous work. We describe our real-time 3-D person tracker. Then we introduce our activity map representation and its use with location-context cues. A map generation algorithm is then presented and map results in a "smart office" environment are shown. Finally, we show a prototype application for location-sticky services using our activity map-based algorithm. We conclude with a discussion of experiments in progress and possible future extensions to the system.

2 Previous Work

Context cues for ubiquitous and pervasive computing have been a topic of increasing interest recently, e.g. [12]. Many systems that provide indoor location awareness and/or location context cues have been proposed, including schemes based on active badges [13], passive receivers [14] and wireless networking systems. Many of these technologies require specific hardware to function and could not be used by a person without an attached device and transmitter. In contrast, our goal is to provide location awareness

by means of computer vision techniques so that users are not required to wear special purpose devices or to explicitly provide a map of their environment.

Computer vision-based methods have been the subject of much research in passive tracking in the past decade, and systems are becoming reliable and cheap enough to deploy in office and home research testbeds. Early vision-based systems for tracking people indoors relied on simple monocular color cues to separate the person from the background and were designed for interaction with games or virtual environments [6, 18, 16, 9, 8]. However tracking using monocular vision methods is difficult when there is significant dynamic illumination from video monitors, video projectors, and changing levels of outdoor illumination (passing clouds, etc.).

To track people and objects visually despite dynamic illumination, researchers have turned to methods that use extended regions of the spectrum and/or multi-view geometry. Multiview and/or stereo methods are a popular way to overcome illumination dependence in indoor tracking. The Easyliving system used a set of stereo range cameras to track people as they moved in an environment [4]. Similar systems were developed by [2]. Systems to estimate stereo despite sparse background surfaces were developed in [5].

Much work has been done in the area of learning models of activity from vision. [10] has shown how to learn characteristic motion maps which represent non-parametric distributions of pedestrians trajectories. [3] introduced an entropic estimation algorithm that yields a concise and computationally lightweight HMM (Hidden Markov Model) of office activity.

In this paper we present a method for automatically estimating activity zones based on observed user behaviors. We use simple position, motion, and shape features, but our work can be extended to include higher order features including object and multi-person interaction.

3 3-D Person Tracker

This section introduces the 3-D person tracker developed in our group. Our tracker uses multiple stereo cameras that observe a particular space and provide such information as the number of persons in the space, as well as a set of data (location, height) attached to each person. By storing tracking data, the system also provides an history of location features of every person in the space.

Our tracking system performs dense, fast range-based tracking with modest computational complexity and is presented next (further details on the tracking system can be found in [5]).

When tracking multiple people, we have found that rendering an orthographic vertical projection of detected foreground pixels is a useful representation (see Figure 1). A

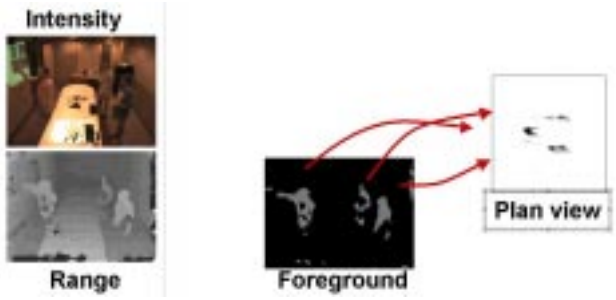

Figure 1. Intensity, disparity, foreground and foreground projection images.

"plan view" image facilitates correspondence in time since only 2D search is required. Previous systems would segment foreground data into regions prior to projecting into a plan-view, followed by region-level tracking and integration, potentially leading to sub-optimal segmentation and/or object fragmentation. Instead, we develop a technique that altogether avoids any early segmentation of foreground data. We merge the plan-view images from each view and estimate over time a set of trajectories that best represents the integrated foreground density. Trajectory estimation is performed by finding connected components in a spatio-temporal filtered volume.

To estimate the trajectory of objects over time, we combine information from multiple stereo views. The true extent of an individual object in a given image is generally difficult to identify. An optimal trajectory segmentation should consider the assignment of an individual pixel to all possible trajectories estimated over time. Systems which perform an early segmentation and grouping of foreground data before trajectory estimation preclude this possibility.

We adopt a late-segmentation strategy that finds the best trajectory in an integrated spatio-temporal representation by combining foreground pixels from each view. By assuming that objects move on a ground plane, a "plan-view assumption" allows us to completely model instantaneous foreground information as a 2-D orthographic density projection[1, 11]). Over time, we compute a 3-D spatio-temporal plan-view volume.

We project (x_j, y_j, d_j) from each foreground point $\vec{p}_j$ into world coordinates (U_j, V_j, W_j). (See Figure 2.) U, V are chosen to be orthogonal axes on the ground plane, and W normal to the ground plane. We then compute the spatio-temporal plan view volume (Figure 2), with

$$P(u, v, t) = \sum_{\{\vec{p}_j \,|\, U_j = u, V_j = v, t_j = t\}} 1$$

Each independently moving object in the scene generates a continuous volume in the spatio-temporal plan view

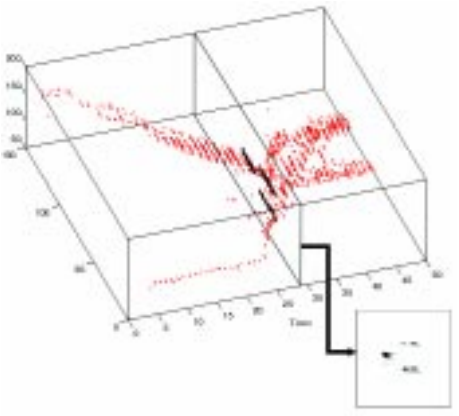

Figure 2. Spatio-temporal representation of projected foreground points.

volume $P(u, v, t)$. When the trajectories of moving objects do not overlap, the trajectory estimation is easy and computed by connected-component analysis in $P(u, v, t)$ (each component is then a trajectory).

When the trajectories of moving objects overlap (*e.g.* crossing of two people), the volume associated with these trajectories in $P(u, v, t)$ also overlap and make the extraction of trajectories more difficult. In order to overcome this, a graph is built from a piece-wise connected-component analysis of $P(u, v, t)$. Nodes correspond here to trajectory crossing and branches to non-ambiguous trajectories between two crossings. A color histogram is then estimated for each branch of the graph (using all images associated with this branch). Trajectories are estimated by finding in the graph the paths consisting of branches having the most similar color histograms. This may be done instantaneously using a greedy search strategy or using the slower but optimal dynamic programming technique described in [5].

This stereo-based tracking system runs at about 12 Hz on a standard computer (Pentium 4, 1.7GHz).

4 Activity Maps

Activity zones are represented in what we call an activity map. This map is the key to our system's ability to provide context information to applications in an intelligent environment. The zones represent regions of a physical space in which observed activity features —location, motion (represented as velocity), shape (represented as height)—have similar values. Ideally, each zone corresponds to a region in which a person is likely to be engaged in similar activities. A relatively still person sitting at a particular location, for example, may be reading, writing, or typing. While knowing that a desk or book is near the person will allow us to more accurately infer actual activity, simply knowing that the person is in a work environment and located in a particular zone in a particular way provides valuable context information for an application program in an intelligent environment.

Our system generates an activity map by clustering spatio-temporal data gathered using our 3-D person tracker. Later the activity map is used to determine what the location context is for that user. As the person enters an activity zone, for example, notification is sent to application programs running in the environment; the applications then react accordingly. Notifications may be sent when the person has been in an activity zone for a certain amount of time or when he exits an activity zone.

An activity map may be thus used with observed real-time features to provide location context for applications in a pervasive computing environment.

4.1 Automatic Estimation of An Activity Map

The person tracker provides a history of 3-D information of every person in the observed space. The 3-D information consists of (x, y, h) where x,y is the coordinates of the person in the ground plane and h is the relative height of the person with respect to the floor.

Since tracking data are time-stamped, the instantaneous velocity (v_x, v_y, v_h) can be derived. We determine a persons features' from the history of spatio-temporal tracking data of a person. We characterize a person at location (x, y) are: $f(x, y) = (h, v, v_{\text{lt}})$ where $v = \sqrt{v_x^2 + v_y^2}$ is the instant ground plane velocity norm and v_{lt}, the average ground plane velocity norm over a certain amount of time. By using the features $f(x, y)$, we can capture the configuration (sitting, standing) and motion of a person over both short and long period of time.

4.2 Segmentation Algorithm

By tracking people in a space for a long period of time, a dense set of observed location features $f_i(x, y)$ can be gathered. We define an activity zone as a connected region where observed location features $f_i(x, y)$ have similar values. An activity zone $\mathcal{Z}_k$ is defined by a connected region R_k in the 2-D space defined by (x, y) and a characteristic feature $F_k = (h, v, v_{\text{lt}})$ representing the typical activity in this area. As different activities may happen at the same location (x, y), activity zones may overlap as well.

Estimating the activity maps involves segmenting observed features $f_i(x, y)$ into activity zones $\mathcal{Z}_k$. In order to perform the segmentation, we use a 2-step approach:

Step 1 Classification of features $f_i(x, y)$ where each class corresponds to a specific activity with characteristic

feature F_k. In order to group features $f_i(x, y)$, we perform an unsupervised classification using a standard k-means algorithm. This algorithm classifies features $f_i(x, y)$ into N classes, where each class has a mean feature F_k. This step does not take into account the location (x, y) of the features.

Step 2 Estimation of connected regions R_k by grouping points (x, y) corresponding to the same activity. This step consists of finding connected components in each of the classes from step 1. For each class F_k, a feature $f_i(x, y)$ is selected. A region is then grown from the seed (x, y) by searching for points (x', y') such that there is a feature $f_j(x', y')$ in class F_k and the distance between (x, y) and (x', y') is close. When a region cannot be grown further, all features used for the region growing are removed from F_k. If F_k is not empty, a new seed (x, y) is picked as a seed to grow a new region.

At the end of Step 2, regions corresponding to small numbers of location features are removed (these regions correspond either to non frequent person's behaviors or to errors from the person tracker). The remaining regions define the activity map.

4.3 Person's Activity Zone Detection

Using an activity map and run time data from the person tracker, the estimation of a person's activity is performed as follow.

Let (x, y) be the location and $f = (h, v, v_{\text{lt}})$, the location feature of a person estimated by the tracker. The corresponding activity zone $\mathcal{Z}$ is found by first finding the regions R_k close to location (x, y). This gives a subset of activity zones $\{\mathcal{Z}_k\}$. The correct activity zone $\mathcal{Z}$ is found as the one from the subset $\{\mathcal{Z}_k\}$ whose feature F_k is the closest to the person's location feature f.

5 Experiments and Applications

5.1 Experiments

We describe two experiments in which our system automatically generated activity maps for different environments, a one-person office and a two-person office. Each office is equipped with a single stereo camera mounted on the wall in a standard surveillance camera configuration. For each experiment, tracking data was recorded over a long period of time and activity maps were estimated off-line using the approach previously described (due to the high number of data, the segmentation algorithm takes several minutes to run). In all of the experiments, the initial number of classes N for step 1 was set to $N = 10$, and at the end of step 2, regions corresponding to small numbers of location features were removed.

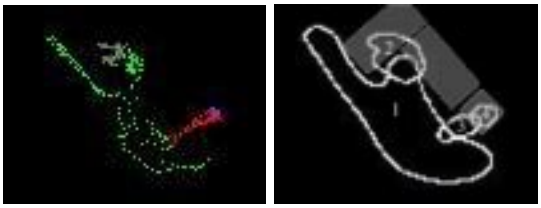

Figure 3. A one-person office.

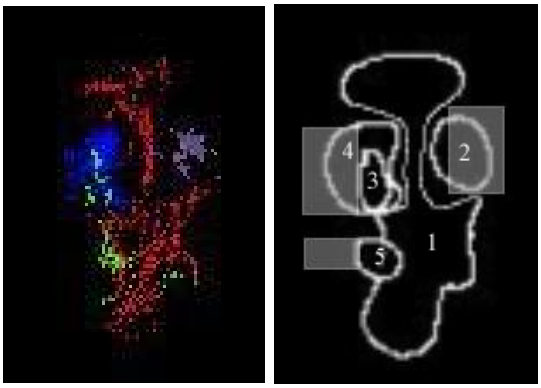

Figure 4. A two-person office.

Results are shown Figures 3 and 4. In each experiment, the automatically generated activity maps segment the space into zones related to structures in the environment (chairs, desk, file cabinet, corridors...). (In the next section, we discuss experiments in using these zones.)

In the case of the one-person office (Figure 3), the estimated activity map contains 4 zones. Zone 1 corresponds to the "walking context", zone 2 corresponds to the "working context" (desk), zones 3 and 4 correspond to the "resting context" (chair on the bottom right of the picture). Zone 3 could be associated to the transition between zone 1 and zone 4 (chair). The location features (velocity, height) corresponding to the different zones are not shown here but we observed that they correspond to expected values: regular standing heights in zones 1 and 3, low heights in zones 2 and 4. Velocities were large in zone 1, medium in zone 3 and small in zones 2 and 4.

The activity map estimated for the two-person office (Figure 4) contains 5 zones. Zone 1 corresponds to the "walking context", zone 2 corresponds to the "working context" (desk) of user A and zones 3 and 4 correspond to the "working context" (desk) of user B (zone 3 is included in zone 4 and corresponds to smaller velocities). Zone 5 corresponds to the file cabinet.

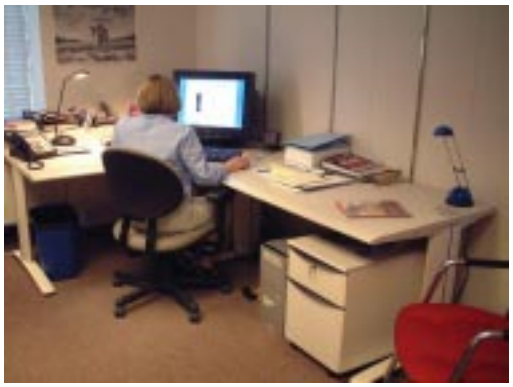

Figure 5. The light and the computer screen are turned on as Sara is sitting in zone 2.

Figure 6. Another light is turned on as James sits in zone 4. Information is displayed on the wall between zone 2 and zone 4.

5.2 Evaluation in Prototypical Applications

In addition to experimenting with the automatic generation of activity zones, we have begun testing the use of our system in an "intelligent" environment. We created two simple scenarios that illustrate a context-aware environment, and used the scenarios to implement a prototype application that we then evaluated. The scenarios are informed by previous work on activity zones, in particular private and public zones [17].

Scenario A: In the morning Sara arrives at work and enters her office. The room lights turn on automatically and the computer screen starts up when she sits down by her desk. While organizing her day and reading her emails, she is listening to morning news on the radio.

Scenario B: James is walking by Sara's office. Seeing Sara working on her computer reminds him about a presentation that they are to give next week. James opens the door and greets her. Sara swivels her chair around and welcomes him. The volume of her radio goes down and after some small talk they decide that they would like to look over the last presentation that they gave on the topic. James sits down in the chair next to Sara's desk and the ambient light in the room increases. Sara asks the room to display the presentation information so that both she and James can see it, and the presentation slides appear on the wall display between them. They then start to work on their presentation. After a short time, the calendar system reminds Sara about the weekly staff meeting, and it also informs her that she has one voice mail that was recorded during her meeting.

Our prototype application focuses on three tasks from the above scenarios: control of light, audio, and display of information in an office.

We used our person tracking system to generate an activity map, added preferred light and audio settings to particular activity zones, then gathered context information for people working in the offices.

In Figure 5 and Figure 6 show two scenes from our prototype system in use. Figure 5 illustrates the light having been turned on when someone is working in zone 2. Figure 6 illustrates the light having been turned on when a visitor is sitting in zone 4. Figure 6 also shows the automatic choice of display (computer screen or projector) between a person sitting in zone 2 (at the desk) and a person sitting in zone 4 (in the chair).

Our preliminary experiments revealed four primary lessons. First, our system does a good job at automatically partitioning a space into zones; a person does not have to specify bounds or characteristics of activities that take place in the zones. Second, the zones provide fine-grained enough partitions of space for certain applications. We can still get "intelligent" behavior in an environment without providing more specific information about either the physical environment (e.g. identifying furniture locations) or a person's actual activities (e.g. reading). Third, our system does a good job at triggering relevant applications by matching a person's location to a particular activity zone without requiring that they wear sensors. Fourth, the sum total of changes in environment state (light and music) and information display state proved useful even in our preliminary user studies.

6 Discussion

In this paper, we show how location context can be obtained with a purely passive observation system. Our system can see location regions that are much smaller than the usual room-level location abstraction without requiring that users wear special purpose devices. Location regions are defined by user activity, and are automatically estimated by observing user behavior.

Our system tracks groups of users with a multi-view stereo trajectory estimation method, then automatically generates an activity map. Our experiments on different environments show that our system is able to generate activity maps that give an improved understanding of a person's context over previous approaches to sub-room location modeling which required that users explicitly define physical regions [4]. The improvement is due to our system's ability to use fine-scaled features that include position, motion and height, making the identification of a person's context more accurate than one based on position along.

There are many avenues of future work planned for our system. In addition to making the system more accurate and fast, we plan to add a statistical estimation formulation to the region estimation process. This will make the estimated regions more stable to noise in the sensing process. We also wish to include higher-level information about the tasks users are performing in the environment and the objects they are manipulating to aid in determining activity. We are developing an articulated body tracker [7] that estimates the body pose of a user (arms, torso and head positions). By using the body pose information in our approach (instead of using location only) we think that many sub-classes of activity will automatically emerge from the segmentation process. We also speculate that more complex application behavior can be achieved by augmenting the system with knowledge of objects (e.g. desk, computer) and human behavior (e.g. people generally read, write at a desk). Adding a simple object recognition system and task knowledge base is planned future research.

We have also shown that even with purely perceptual information about context, interesting and useful applications can be developed for intelligent environments. Activity maps lay a good groundwork for further exploration of richer definitions of context: these definitions might include information about sound, objects in an environment (e.g. furniture), and "typical" activities that take place using or near objects. Adding richer information to our system will enable inference about what a person is doing (e.g. reading, writing), thus enabling even more "intelligent" behavior. Finally, we intend to explore issues of privacy in an environment augmented with a person tracking system such as ours,

References

[1] D. J. Beymer and K. Konolige. Real-time tracking of multiple people using stereo. In *Frame-Rate Workshop*, 1999.

[2] D.J. Beymer. Person counting using stereo. In *HUMO00*, pages 127–134, 2000.

[3] M. Brand. Learning convise models of human activity from ambient video via a structure-inducing m-step estimator. Technical Report TR-97-25, MERL - A Mitsubishi Electric Research Laboratory, 1997.

[4] B. Brumitt, J. Krumm, B B. Meyers, and S. Shafer. Ubiquitous computing and the role of geometry. *IEEE Personal Communications*, 7-5:41–43, August 2000.

[5] T. Darrell, D. Demirdjian, N. Checka, and P. Felzenszwalb. Plan-view trajectory estimation with dense stereo background models. In *2001 International Conference on Computer Vision*, 2001.

[6] T. Darrell, P. Maes, B. Blumberg, and A. Pentland. A novel environment for situated vision and behavior, 1994.

[7] D. Demirdjian and T. Darrell. 3-d articulated pose tracking for untethered diectic reference. In *ICMI02 (to appear), Pittsburgh, Pennsylvania*, October 2002.

[8] A. Elgammal, D. Harwood, and L. S. Davis. Nonparametric background model for background subtraction. In *6th European Conference of Computer Vision*, 2000.

[9] I. Haritaoglu, D. Harwood, and L.S. Davis. W4: Real-time surveillance of people and their activities. *PAMI*, 22(8):809–830, August 2000.

[10] N. Johnson and D. Hogg. Learning the distribution of object trajectories for event recognition. *IVC*, 14(8):609–615, August 1996.

[11] J. Krumm, S. Harris, B. Meyers, B. Brummit, M. Hale, and S. Shafer. Multi-camera multi-person tracking for easyliving. In *3rd IEEE Workshop on Visual Surveillance*, 2000.

[12] T. Moran and P.Dourish. Special issue on context-aware computing. *Human-Computer Interaction*, 16, 2001.

[13] M.Weiser. The computer for the 21st century. *Scientific America*, 265(3):94–104, 1991.

[14] Nissanka B. Priyantha, Anit Chakraborty, and Hari Balakrishnan. The cricket location-support system. In *Mobile Computing and Networking*, pages 32–43, 2000.

[15] Bill Schilit Roy Want. Special issue on location-aware computing. *IEEE Computer*, 34(8), 2001.

[16] C. Stauffer and W.E.L. Grimson. Learning patterns of activity using real-time tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8), 2000.

[17] K. Tollmar and S. Junestrand. Video mediated communication for domestic environments -architectural and technological design. In *Proc. Lecture Notes in C. S.*, 1999.

[18] C.R. Wren, A. Azarbayejani, T.J. Darrell, and A.P. Pentland. Pfinder: Real-time tracking of the human body. *PAMI*, 19(7):780–785, July 1997.