

# Recognizing Personal Location from Video

Hisashi Aoki, Bernt Schiele and Alex Pentland  
MIT Media Laboratory  
20 Ames St., Cambridge, MA 02139  
{aoki,bernt,sandy}@media.mit.edu

## Abstract

*An important function for wearable computers is the recognition of places and locations. This paper proposes an image sequence matching technique for the recognition of previously visited places. Similar in spirit to single word recognition in speech recognition, a dynamic programming algorithm is proposed for the calculation of dissimilarities of video sequences. Such video sequences represent not only the place itself but also the approaching trajectory. The algorithm uses a simple and robust representation of a single frame without compromising the discrimination between different places. Preliminary experimental results demonstrate the discrimination and the robustness of the approach with respect to the angle of the approaching trajectory.*

## 1. Introduction

Computers and cameras are becoming small enough to wear. Computers are being equipped with cameras, microphones and other sensors, which can be used to interpret what people are doing, where they are and with whom they are interacting in order to support and assist them. To attain this goal, computers should not only model and understand human actions but they should also perceive and model the environment of the human. This is important for understanding the *context* of the human, including where he is, what he is doing, how and when the system can interact with the human and what information to present. Furthermore, the knowledge of the location of a human is an important link between the virtual and the physical world. Such links can be easy to understand (i.e. the association of the computer-stored calendar with the physical wrist watch) and can be used in multiple ways. The computer might help to navigate in unfamiliar or unknown places. An important reminder might be shown on personal display when the user is in a particular place such as the supervisor's office or the supermarket [1,2,3].

As pointed out, a desired functionality of a wearable system is the modeling and detection of places where events take place. Outdoors, a Global Positioning System (GPS)

may track positions and movements. Indoors, active tags may take this role [4,5,6]. Besides the fact that GPS is not available inside buildings, the dependency on infrastructure is obviously a disadvantage.

In [6], visual tags have been used in order to identify places. We exploited object recognition techniques in order to link the physical and the virtual world [7]. These techniques analyze single images and therefore rely on the quality of single images in order to recognize objects and places accurately. For recognition, our system not only uses the visual appearance of the place itself but an image sequence of the entire trajectory used to approach the place.

Recognizing places is a common problem in robotics [8] (generally called localization). However, most approaches cannot be used directly in a wearable system due to the uncontrolled camera movements. Furthermore, robotics systems typically emphasize accuracy of the model (on which they often rely) over the online and incremental character of the system. A wearable system on the other hand should be online and incremental, in order to minimize the a priori hand modeling of the environment and to maximize the adaptability of the system to unknown environments.

Here we propose the recognition of places by matching image sequences. The proposed method analyzes frame image features of video sequences of approaches to particular places with a personal camera. Histograms of simple image features are extracted from each frame. By finding similar transitional patterns of the feature histogram in the place dictionary, similar places can be identified. The method compares entire or parts of sequences, without the need for time-consuming image segmentation. The next section describes the frame features and the matching of image sequences by dynamic programming. Section 3 gives preliminary experimental results.

## 2. Image sequence recognition

Places are characterized by their visual appearance. However, in real life, there is also a limited number of trajectories for approaching places i.e. as imposed by the street layout or the topology of a building's interior. This paper proposes to represent and recognize a place not by a

single image but rather by an image sequence including images of the place as well as the approaching trajectory. Since the place recognition does not rely on a single image but rather on an entire collection of images each frame can be represented in a simple way. This mainly increases the robustness of the algorithm to changes in the environment, and noise in the data due to uncontrolled camera movements. The price for the increase in robustness is the dependency of the technique to the approaching trajectory (including camera movement and approaching direction). However, as the experimental results demonstrate the robustness of the approach enables the recognition of trajectories even in the presence of major trajectory changes.

## 2.1 Frame feature vector

The proposed method uses a chromatic histogram as a frame feature. The histogram is made from hue values, calculated by taking the arctangent of  $(C_b, C_r)$  with  $(Y, C_b, C_r)$  defined as:

$$\begin{pmatrix} Y \\ C_b \\ C_r \end{pmatrix} = \begin{pmatrix} +0.2125 & +0.7154 & +0.0721 \\ -0.1150 & -0.3850 & +0.5000 \\ +0.5000 & -0.4540 & -0.0460 \end{pmatrix} \begin{pmatrix} R \\ G \\ B \end{pmatrix}$$

$$0 \leq Y \leq 1 \quad -0.5 \leq C_b, C_r \leq +0.5 \quad 0 \leq R, G, B \leq 1$$

Hue values are calculated by:

$$H = \arctan(C_b, C_r) / \pi \quad -1 \leq H \leq 1$$

Hue is relatively unaffected by brightness changes [9,10]. Therefore, hue histograms are robust to such changes.

Each frame is represented by an  $N$ -bin hue histogram, which can be plotted as a point in an  $N$ -dimensional space. Each sequence therefore corresponds to a trajectory in the  $N$ -dimensional space.

When the histogram of frame  $j$  in sequence  $i$  is denoted as the  $N$ -dimensional vector  $\mathbf{f}_{ij}$ , a  $N \times M$  matrix denotes the trajectory  $\mathbf{T}_i$ :

$$\mathbf{T}_i = (\mathbf{f}_{i1}, \mathbf{f}_{i2}, \dots, \mathbf{f}_{iM})$$

with  $M$ , equal to the number of frames in sequence  $i$ .

## 2.2 Trajectory distance

In order to calculate the similarity between two trajectories  $\mathbf{T}_k$  and  $\mathbf{T}_l$ , the distance value  $\mathbf{D}_{kl}$  between each vector in  $\mathbf{T}_k$  and  $\mathbf{T}_l$  is used:

$$\mathbf{D}_{kl} = \begin{pmatrix} d(\mathbf{f}_{k1}, \mathbf{f}_{l1}) & d(\mathbf{f}_{k2}, \mathbf{f}_{l1}) & \dots & d(\mathbf{f}_{kM_k}, \mathbf{f}_{l1}) \\ d(\mathbf{f}_{k1}, \mathbf{f}_{l2}) & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ d(\mathbf{f}_{k1}, \mathbf{f}_{lM_l}) & \dots & \dots & d(\mathbf{f}_{kM_k}, \mathbf{f}_{lM_l}) \end{pmatrix}$$

where  $d(\mathbf{a}, \mathbf{b})$  is a distance between two vectors  $\mathbf{a}$  and  $\mathbf{b}$ . The following discusses different examples of  $\mathbf{D}$  with  $M_k$  and  $M_l$  the length of  $\mathbf{T}_k$  and  $\mathbf{T}_l$  respectively. The diagonal elements are 0 when  $k$  and  $l$  are identical sequences.

However, even if two trajectories taken to approach the same place are identical, the approaching speeds may be different. Therefore, the image sequence matching algorithm should use non-linear time warping similar to the dynamic programming algorithm used for single spoken word recognition [11].

The sequence of correspondences in the matrix  $\mathbf{D}_{kl}$  is denoted as  $P_{kl}$ , the global minimal path. The total distance along  $P_{kl}$  represents the distance between two trajectories. The global trajectory distance is calculated as:

$$T_{kl} = \sum_{(i,j) \in P_{kl}} (D_{kl})_{ij} = \sum_{(i,j) \in P_{kl}} d(\mathbf{f}_{ki}, \mathbf{f}_{lj})$$

If sequences  $k$  and  $l$  are the same,  $T_{kl}$  should be 0 and  $\mathbf{D}_{kl}$  is diagonal.  $T_{kl}$  will have small value if sequences  $k$  and  $l$  are similar. A problem, similar to image sequence matching as described above is single word recognition in the context of speech recognition. Dynamic programming algorithms have been employed for this problem allowing non-linear time warping between the input sequence (here the incoming video-stream) and a dictionary of words (here a place dictionary of approaching trajectories)[11,12].

Assuming that the starting position of the image sequence is known, the employed dynamic programming algorithm calculates incrementally the global minimal path to reach each column of the sequence. Once the final position is reached the global minimal path can be backtracked to the start position (see [11,12] for further details).

## 2.3 Place detection

In order to detect places, a place dictionary of trajectories to known places is constructed. When a new trajectory  $\mathbf{T}_l$  is recorded, the system calculates trajectory distance  $T_{kl}$  for all sequences  $k$  in the dictionary and the trajectory  $k$  is chosen with smallest  $T_{kl}$ .

## 3. Experiments

For this experiment, video sequences have been recorded approaching three different indoor places (A,B,C) in the same room. For each place, sequences are recorded along four different approaching angles. Three segments,

one from each of the three places, are used as the place dictionary. Also, video sequences have been taken approaching two different outdoor places (D,E). For each place, sequences are recorded along seven different approaching angles. Two segments, one from each of the two places, are used as the place dictionary. Outdoor video sequences were taken approximately between 3pm and 4pm. To test robustness of the proposed algorithm, outdoor sequences which don't have many colors, i.e., when white walls and pavements cover large parts of the image are chosen. Fig. 1 shows the places and the approximate approaching trajectories selected randomly.

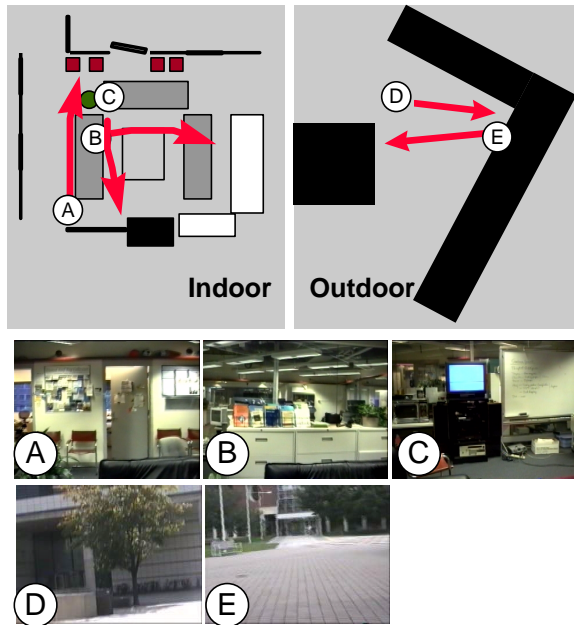


Fig. 1 Places and approaches

The video sequences are taken by a video camera and subsampled to 5 frames per second for indoor sequences, and 1 frame per second for outdoor sequences. Each indoor sequence is about 10 seconds long, each outdoor sequences is about 50 seconds long. About 50 frames are used for each indoor and outdoor sequences. 16 bin chromatic histograms are calculated as frame feature  $\mathbf{f}$  ( $N=16$ ), and  $\chi^2$  distance is used in calculating the distance between two  $\mathbf{f}$ 's. In this experiment, distance is calculated as [12];

$$\chi^2(\mathbf{A}, \mathbf{B}) = \sum_i \frac{(a_i - b_i)^2}{a_i + b_i}$$

We'd like to point out that the  $\chi^2$  statistics is not metric since the triangular inequality does not hold. Even though this violates an intrinsic assumption of the dynamic programming approach introduced above, the  $\chi^2$  statistic

is used since it has been shown to attain best discrimination for recognition purposes compared to several other distances[13].

Trajectory distances  $T$  are shown on Table 1. All nine indoor test sequences are detected as the correct places by choosing the global minimal trajectory distance over the place dictionary. Test sequence 3 for place D and test sequence 1 for place E are misdetected. However, there is no confusion between indoors and outdoors. These distances have been obtained for trajectories with differences of the approaching angle up to  $\pm 60^\circ$ . These preliminary results indicate the robustness of the technique to variations of the trajectory in speed as well as in approaching angle. No attempt has been made to control camera movements during recording.

dictionary->		A	B	C	D	E
A	test 1	1.00	8.10	23.6	17.5	21.9
	test 2	1.22	6.46	23.0	14.5	17.8
	test 3	1.77	9.35	24.4	19.7	24.2
B	test 1	6.30	2.03	21.4	10.3	13.7
	test 2	6.99	2.27	22.0	9.02	11.4
	test 3	7.04	3.59	23.9	10.6	15.7
C	test 1	14.8	9.87	5.30	15.7	22.5
	test 2	23.3	12.3	5.57	15.6	26.1
	test 3	27.8	23.2	9.24	19.4	26.3
D	test 1	22.3	15.1	19.6	2.11	8.33
	test 2	21.8	19.4	17.0	5.60	9.90
	test 3	40.0	24.9	24.8	8.87	9.83
	test 4	29.2	17.2	21.9	5.49	8.68
	test 5	28.0	14.9	24.5	3.18	8.19
	test 6	24.9	12.7	22.7	4.57	7.66
E	test 1	31.2	19.6	24.8	10.0	5.94
	test 2	24.5	17.1	26.7	7.84	3.53
	test 3	36.5	22.5	27.9	9.25	6.15
	test 4	51.8	29.7	28.9	16.7	16.7
	test 5	47.3	27.2	27.7	13.5	13.4
	test 6	38.0	22.7	24.8	13.2	8.73

Table 1 Trajectory distances

Fig. 2 shows the distance matrices  $\mathbf{D}$  and the global minimal path. Components are indicated by brightness, and global minimal paths are indicated as black and white trails. In the figure, small numbers are distributed around diagonals in the pairs of test sequence and the correct dictionary, while there are large numbers around diagonals in the false pairs.

#### 4. Conclusion

We introduce a place detection method by computer vision in the context of wearable computing. Instead of using single frames for the representation of places we propose to match image sequences of the approach to places. A dynamic programming algorithm has been introduced in order to deal with non-linear time variations between

trajectories. Preliminary experimental results demonstrate that a simple and robust representation for single frames is sufficient in order to discriminate between different places, even if the presence of large variation of the approaching trajectory such as angle and speed.

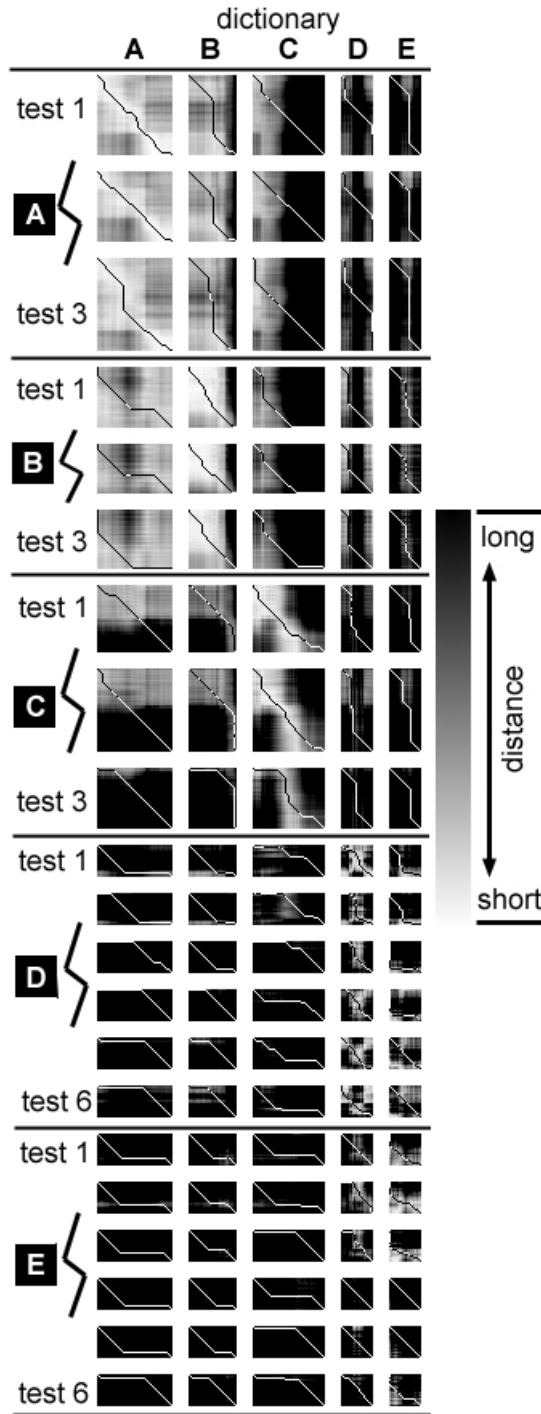


Fig. 2 Distance matrices and global minimal path  
The method is currently being applied to larger datasets

and will be ported to PC-based wearable computers developed at the MIT Media Laboratory. Future work includes the automatic and incremental learning of place dictionaries in order to construct a topological map of the user's environment.

## References

- [1] M. Weiser. *The computer of the twenty-first century*. Scientific American, 1991.
- [2] S. Kakez, C. Vania, and P. Bisson. *Virtually documented environment*. In Proceedings of the First Intl. Symposium on Wearable Computers ISWC97.
- [3] J. Rekimoto and K. Nagao. *Agent augmented reality: a software agent meets the world*. In Proceedings of Second Intl. Conference on Multiagent Systems (ICMAS-96).
- [4] R. Want and A. Hopper. *Active badges and personal interactive computing objects*. IEEE Trans. on Consumer Electronics, 38(1):10-20, Feb. 1992.
- [5] J. Orwant. *For want of a bit the user was lost: Cheap user modeling*. IBM Systems Journal, 35(3), 1996.
- [6] T. Starner, S. Mann, B. Rhodes, J. Healey, D. Kirsh, R.W. Picard, and A. Pentland. *Augmented reality through wearable computing*. Presence 6(4): 386-398, 1997.
- [7] B. Schiele and J. Crowley. *Probabilistic object recognition using multidimensional receptive field histogram*. Intl. Conference on Pattern Recognition (ICPR96) B:50-54, 1996.
- [8] J. Borenstein and H.R. Everett and L. Feng. *Where am I? Sensors and methods for autonomous mobile positioning*. Tech. Report of University of Maryland, 1996.
- [9] D.H. Ballard and C.M. Brown. *Computer Vision*. published by Prentice Hall. 1982.
- [10] H. Aoki, S. Shimotsuji and O. Hori. *A shot classification method of selecting effective key-frames for video browsing*. In Proceedings of ACM Multimedia 96: 1-10. 1996.
- [11] H. Sakoe and S. Chiba. *Dynamic programming algorithm for spoken word recognition*. Readings in Speech Recognition: 159-165. 1990.
- [12] W.H. Press and S.A Teukolsky and W.T. Vetterling and B.P. Flannery. *Numerical recipes in C*. Published by Cambridge University Press. 1992.
- [13] B. Schiele. *Object Recognition using Multidimensional Receptive Field Histograms*. Doctoral thesis, I.N.P.Grenoble. 1997.