iWalk, A Tool for Interacting with Geo-Located Data Through Movement and Gesture

Visruth Premraj
Computer Science Dept
Stony Brook University
Stony Brook, NY
visroo@gmail.com

Margaret Schedel
Music Dept
Stony Brook University
Stony Brook, NY
gem@schedel.net

Tamara L. Berg
Computer Science Dept
Stony Brook University
Stony Brook, NY
tlberg@cs.sunysb.edu

ABSTRACT

In this work, we present iWalk, a multimedia exploration tool that provides an interactive virtual environment for physically exploring geo-tagged data. This tool is flexible enough for users to easily explore their own collections, or existing collections from the web. Two interaction modalities are incorporated into our tool – movement and gesture. Movement (walking around the physical space) is used to intuitively move through the digital data space of a collection. Gesture is used for direct data manipulation; the user is able to select the mapping between gestures and interactions. In addition, we also provide functionality for exploring data that is not geo-located. Here the user defines a rough mapping between the data collection space and the physical interaction space and then operates the program as usual. We have currently tested the system on three different data sets – a large collection of geo-tagged photographs from Flickr, a geo-located sound collection, and a museum collection that is not geo-located.

Categories and Subject Descriptors

H.5.1 [**Information Interfaces and Presentations**]: Multimedia Information—*artificial, augmented and virtual realities, evaluation/methodology*; H.5.2 [**User Interfaces**]: input devices and strategies, interaction styles

General Terms

Design, Human Factors

1. INTRODUCTION

Location information and digital media are quickly becoming intertwined as more and more phones, cameras, and computers integrate GPS systems directly into their hardware. Enormous quantities of geo-located data in the form of images, videos, and sounds are now freely available on the web, and yet the average user's experience of interacting with this data remains impoverished and frustrating. For example, Flickr has hundreds of millions of user contributed geo-tagged images, but to interact with these images a user must click on a map (or provide an input location) and then slowly click through page after page of results to see all of the associated data for a location.

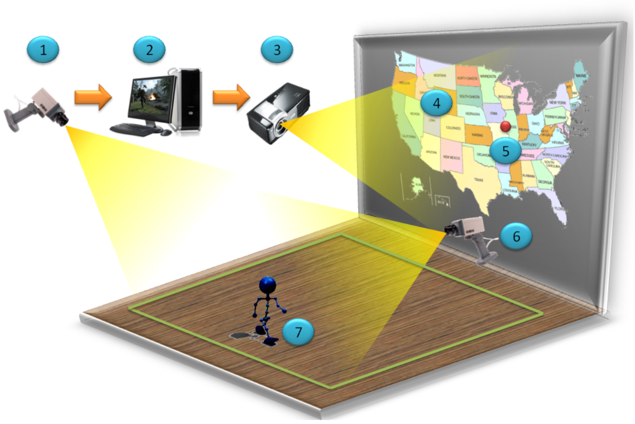


Figure 1: The physical set up of the proposed system: 1) Camera on the ceiling for localization, 2) Processor, 3) Projector displaying output, 4) Projected map, 5) Mapped user location, 6) Camera from the front for gesture recognition, 7) User.

In this paper we provide an alternative – an easy to use, and intuitive tool for users to physically explore data collections. These collections can range from a user's own geo-located data collection, to existing locative data collections from the web, to non-location based collections for which the user can provide some organizational structure. Our system uses two modalities of interaction: movement and gesture. *Movement* (walking around the physical space) enables users to quickly explore the digital space of a collection while maintaining a physical sense of motion through space. *Gesture* allows the user to directly manipulate individual data items. This interaction makes use of computer vision algorithms computed on standard commercial camera inputs and does not require the user to wear any distinctive markers, or gloves. The physical setup (shown in fig 1) consists of: a server for processing, a camera viewing the user from above for localization, a camera viewing the user from the front for gesture recognition, and a projector to display the

MM'10, October 25–29, 2010, Firenze, Italy.
Copyright 2010 ACM 978-1-60558-933-6/10/10 ...$10.00.

map, current user location, and data. Optional monitoring speakers can be added for data collections involving sound.

Our tool is adaptable to user needs and enables exploration for a variety of data collection types, including images, sound, and video. We have also designed a graphical user interface (fig 4) that allows users to quickly create an interaction with any geographically organized data collection, or to define their own interaction for non-location based collections. Currently we have evaluated our system on three data sets: a collection of geo-tagged photographs from Flickr, a geo-located audio collection from soundcities, and a non-location based database from the Museum of Modern Art website. Initial results and feedback from users is promising and we plan to extend these ideas to include additional functionality and a broad range of collections.

Our tool has three main components:

- A recognition component that uses computer vision techniques to localize the user in the physical space, and recognize the user's gestures (described in sec 2).
- A GUI that enables the user to import and define a mapping between the digital data space and the physical user space, and also allows the user to define a mapping between gestures and their desired data manipulation tasks (described in sec 3).
- A presentation component that projects the data map, user location, and data onto the wall, or plays back sound (described in sec 4).

1.1 Related Work

Due to space constraints we just mention some of the most relevant projects out of the extensive previous research related to location information, digital data and gesture based interfaces. Pejic et al [4] use a knowledge base formed by tracking user actions to suggest tourist location information of special interest to a user. Yang et al [6] discuss the design and implementation of a campus spatial information service based on Google maps that provides users with rich and interactive information in the form of pictures, descriptions, and links. Most relevant to our work, Michael et al [2], present an interactive map for digital library collections. Their system automatically derives descriptors for video that can then be displayed on a map synchronously with viewing. Similarly, we also present an interactive map, but focus on developing a more general tool for users to explore their own geo-located databases containing various types of digital media.

2. LOCALIZATION & GESTURE RECOGNITION

Our system provides two computer vision based recognition components: *localization* to determine where in the physical space the user is located at any given time, and *gesture recognition* to determine which gestures a user is executing. Depending on the predicted user location, the corresponding data items from a collection will be displayed. Depending on the predicted user gestures, various data manipulation actions will be executed.

2.1 Localization

Localization is performed using the top camera looking down on users in the scene. Because we have control over our physical space, a simple background subtraction based localization method can be utilized. First an initial background sequence is collected without any users, and the mean background frame is calculated. Then once the system begins to function, this mean background frame is subtracted from each subsequent frame. Thresholding this difference image and removing small components provides the set of pixels containing users. The location of each user is computed as the centroid of each remaining component. The video from the top camera also helps to determine the start of the gesture cycle based on knowledge of whether the user is walking or standing at a particular location performing gestures.

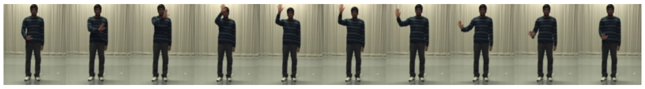


Figure 2: Example gesture for interaction – rotate clockwise. The user can provide a mapping between each possible gesture and each possible interaction for their particular collection.

2.2 Gesture Recognition

We evaluate two methods for gesture recognition. The first method is based on Motion History Images (MHI) and Principal Component Analysis (PCA). The second method is based on Optical Flow accumulated Local Histograms. Because we require a real-time system, we implemented, but could not make use of recent state of the art action recognition systems that require lengthier processing (*e.g.* [5]).

For either recognition approach, first the silhouette of the user is obtained from the front camera using background subtraction and thresholding to compute a bounding box around the user. This region is segmented and rescaled to a fixed size. Then the gesture descriptor corresponding to a video window ending in the current frame is generated. Gesture recognition is computed approximately every 3 video frames (the current frame is processed as soon as processing is finished for the previous frame). Classification is computed using nearest neighbors with an L_2 distance metric. If the current gesture is similar enough to a training gesture, then the gesture is recognized and sent to the server to trigger the corresponding manipulation action. Currently we recognize 9 different gestures (one example is shown in fig 2) though more could be incorporated based on user needs.

Motion History Images and Eigen Actions: MHI is a scalar-valued descriptor where intensity is a function of recentness of motion [1]. It not only describes where the action takes place for each pixel in a video sequence, but also how action evolves over time. This effectively embeds the motion information of an action into a single descriptor. In an MHI, H_t, the value of a pixel at (x, y), $H_t(x, y, i)$, is a function of the temporal history of motion at that point and can be defined as follows:

$$H_t(x,y,i) = \begin{cases} \tau & \text{if } D(x,y,i) = 1 \\ max(0, (H_t(x,y,i-1) - 1) & \text{otherwise} \end{cases}$$

where $D(x, y, i)$ is the ith binary image in an action sequence, and τ is the maximum duration a motion is stored. Examples of start frame, end frame and corresponding MHI for action "waving two hands" are shown in Figure 3 from left to right. Principal Components Analysis is employed for dimensionality reduction.

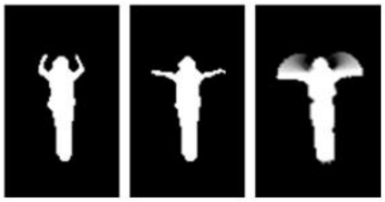


Figure 3: Example gesture: left – start frame, middle – end frame, right – computed MHI.

Optical Flow accumulated Local Histograms The second approach we evaluate is an optical flow accumulated local histogram. For this method the optical flow, F, of a sequence is computed to determine correspondences and motion of pixels between consecutive images. We utilize the motion estimation technique of Farneback et al [3] that computes the optical flow based on polynomial expansion. In order to get stable and robust information describing a whole motion, we accumulate the instantaneous motion between every two frames in a set of binned spatial histograms (where we used 8 bins per histogram). This increases robustness to small variations in gestures by different users.

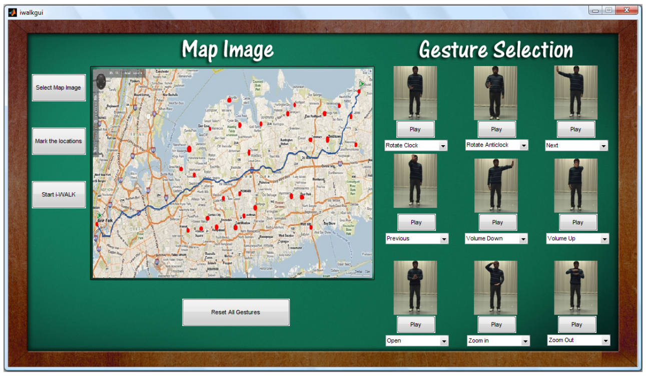


Figure 4: Graphical user interface for map selection and gesture assignment. The user can upload a map image and either denote map coordinates or provide a direct mapping between their specific data and map locations. This provides functionality for interacting with both geo-located data (where the data comes with mapped locations) and non-location based data (where the user provides the mapping between data and mapped space). The user can also preview and assign each predefined gesture (*e.g.* move right hand to right) to a particular media action (*e.g.* move to next image).

3. USER INTERFACE

In order to use our tool, users must provide definitions for the two axes of interaction – movement (localization) and gesture. The movement axis is defined by creating a mapping between the data collection space and the physical interaction space (fig 1). For geo-located data collections the mapping between the user's physical location and the data space is direct and can be associated by the user by either clicking on points on the uploaded map for each geographic location present in the collection or by specifying the bounding map locations. Figure 4 shows an example mapping between geo-tagged photo locations and a map. For non-location based data collections we also provide functionality for users to provide their own mapping between physical locations and data space. Figure 6, right, shows a user defined mapping between museum collection media and a map. Orange circles indicate locations that the user can use to access corresponding museum data.

The gesture axis is defined by asking the user to create a mapping between the gestures provided by our system and the action manipulations for their collection. Gesture videos can be played and then associated with a specific data manipulation action of the user's choice from a predefined set. Some manipulations are applicable to images (*e.g.* rotate, zoom, etc). Other manipulations are applicable to sound or video (*e.g.* play, volume up, etc). Currently we implement 9 different data manipulation actions though additional actions can easily be incorporated depending on user needs.

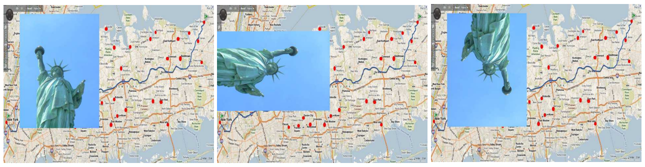


Figure 5: Example user interaction with a geo-tagged photo database collected from Flickr. The user's physical location in the space has been automatically mapped to an actual location – New York City – and gesture recognition has allowed the user to rotate the current photo.

4. DISPLAY

We use Cycling 74's Max/MSP/JITTER to display the user uploaded map, current user location in the digital data space, and the data items the user is interacting with, and playback or manipulate sound and video. MAX is an interactive graphical programming environment for music, audio, and media. JITTER extends MAX with video and matrix data processing. The Open Sound Control (OSC) library is used over a UDP Port to communicate between Max/MSP/JITTER and Matlab (utilized for camera image acquisition, localization, and gesture recognition).

Once we have localized the user and determined his/her current gesture (if a gesture is being executed), we send his/her current (x, y) location and gesture identification (ID) number over open sound control. The location auto populates an ordered list of data items associated with that location and also provides information to associate a video of footsteps moving over the background map when the user is walking around. This provides the user with helpful feedback about where (s)he is located in the digital data space. The transmitted gesture ID number triggers the appropriate data manipulation action where the gesture to action mapping has been provided beforehand through the user interface. For sound based collections speakers are used to audibly display the data items.

4.1 Results & Evaluation

We have successfully tested our system on three different types of digital media collections and performed real-time

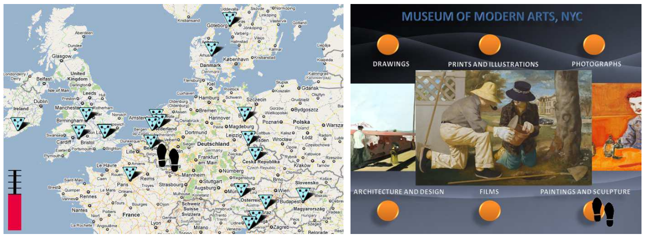


Figure 6: Example user interaction with a geo-located sound database (left) and a museum collection (right). For the sound database the user can explore the audio collection in a locative manner and perform data manipulation tasks via gestures such as "play", "lower volume" etc. For the museum collection since the data is not geo-located a priori, the user can use our tool to provide a mapping between the display space and their collection – in this case associating each type of artistic media (drawings, photographs etc) with a location on the display. Footprints show the mapping between the user's current location in the space and their location within the collection.

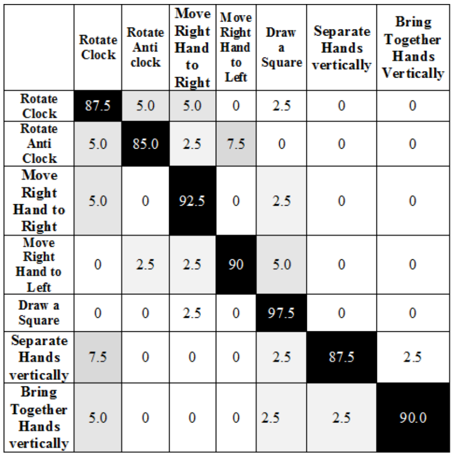
evaluations of the data exploration interaction. So far user feedback has been promising.

The first database consists of images and videos collected from the internet and organized according to their associated geo-tags (latitude-longitude). Images are collected from Flickr using the Flickr API and videos are collected from Youtube using the Google API (see fig 5). Here our experiments found that users were able to move around the digital data space to explore images associated with different geographic locations, quickly sift through images associated with a particular location, and manipulate images in useful ways (zoom in, go to the next image, etc).

The second database consists of sounds collected from the publicly available database www.soundcities.com which have associated geo-tags and moods. Here one experiment we performed was to allow users to walk around the space and play the corresponding tracks from each location (see fig 6). We also created an interaction for exploring the collection according to mood (ambient, birds, industrial etc) by having the user associate each possible mood with a location on the user uploaded map. This allowed users to explore the collection along another dimension.

The third database consists of images collected from the online Gallery of Museum of Modern Art in New York. For this collection, we created an interface that organized the data items according to media type (paintings, photographs, etc – see fig 6). This enabled type based browsing and viewing, though the set of possible data organizations is limitless.

Gesture Recognition: The effectiveness of our system greatly depends on its ability to recognize gestures accurately. Hence we have quantitatively evaluated our gesture recognition component under real-time usage. We created a training set of 9 videos performed by 3 different users for each gesture, and then evaluated recognition of each gesture performed 20 times by two users standing at varying locations across the physical space of the installation. To consider actual usage conditions, we perform these evaluations during actual usage rather than in the offline manner considered by many action recognition systems. The observed confusion matrix for MHI+PCA is shown in Figure 7, showing that gestures are recognized quite accurately in our real-time system. For the optical flow based method recognition, rates were similar, but slightly lower.



	Rotate Clock	Rotate Anti clock	Move Right Hand to Right	Move Right Hand to Left	Draw a Square	Separate Hands vertically	Bring Together Hands Vertically
Rotate Clock	87.5	5.0	5.0	0	2.5	0	0
Rotate Anti Clock	5.0	85.0	2.5	7.5	0	0	0
Move Right Hand to Right	5.0	0	92.5	0	2.5	0	0
Move Right Hand to Left	0	2.5	2.5	90	5.0	0	0
Draw a Square	0	0	2.5	0	97.5	0	0
Separate Hands vertically	7.5	0	0	0	2.5	87.5	2.5
Bring Together Hands vertically	5.0	0	0	0	2.5	2.5	90.0

Figure 7: Confusion matrix for 9 gestures recognized using a MHI + PCA approach, computed during real-time interaction with the system to reflect actual usage accuracy.

5. CONCLUSION & FUTURE WORK

We demonstrate an intuitive tool for users to create interactions with large digital media collections. This tool can be used for a variety of different data types (video, images, sound), and can be applied to any geo-located data collection or collection for which the user provides a mapping. Performance evaluation of the system indicates its ability to operate in real time and successfully track and recognize gestures. Planned future work includes further data manipulation capabilities for data selection and removal, on-the-fly data editing, and the ability to display large sets of data at once to sift through and explore the data more easily.

6. REFERENCES

[1] A. F. Bobick and J. W. Davis. The recognition of human movement using temporal templates. *IEEE PAMI*, 23:257–267, 2001.

[2] M. Christel, A. Olligschlaeger, and C. Huang. Interactive maps for a digital video library. *ACM Multimedia*, 2000.

[3] G. Farnebäck. Two-frame motion estimation based on polynomial expansion. In *Proc of the 13th Scandinavian conference on Image analysis*, 2003.

[4] A. Pejic, S. Pletl, and B. Pejic. An expert system for tourists using google maps api. In *Intelligent Systems and Informatics*, 2009.

[5] C. Schüldt, I. Laptev, and B. Caputo. Recognizing human actions: A local svm approach. In *Proc. ICPR*, pages 32–36, 2004.

[6] Y. Yang, J. Xu, J. Zheng, and S. Lin. Design and implementation of campus spatial information service based on google maps. In *MASS*, 2009.