OFFICE PRESENCE DETECTION USING MULTIMODAL CONTEXT INFORMATION

Xiao Huang and Juyang Weng[*]

Michigan State University
Computer Science Departement
East Lansing, MI, 48824, USA
Email: {huangxi4, weng}@cse.msu.edu

Zhengyou Zhang

Microsoft Research
One Microsoft Way
Redmond WA 98052, USA
Email: zhang@microsoft.com

ABSTRACT

An office presence detection system is presented in this paper. Context information from multi-sensory inputs is integrated to infer a user's activities in an office. We design a layered architecture to model human activities with different granularities. An IHDR (Incremental Hierarchical Discriminant Regression) tree is used to automatically generate models for acoustic signals from unsegmented auditory streams, with a high adaptive capability to new settings. Hidden Markov Models (HMM) are implemented to detect human motion patterns. The outputs of the above two components are fed into high-level HMMs to analyze human activities. Experimental results of the real-time prototype system are reported.

1. INTRODUCTION

Context-aware system [1] has drawn increasing attention from researchers and engineers. Context is defined as "the situational information relevant to the interaction between a user and an application." Context consists of not only immediate multi-sensory inputs from streams of video, audio and computer interactions (mouse and keyboard information) but also other aspects of a user's information such as past states and intentions. Context-awareness is the key component of the next generation human-computer interaction technique, which tends to measure the information about "where," "what," "when," and "who."

A significant portion of previous work focused on recognizing human activities based on a single modality input in a specific environment. There are two popular probabilistic approaches in visual activity recognition: Hidden Markov Model (HMM) and Bayesian Belief Network (BBN). One of the earlier attempts to apply HMMs to activity recognition is found in [2]. Since then, a lot of extensions of HMMs have been tried to model different human activities. Variable-length HMM [3] is applied to exercise behavior recognition. Brand & Oliver used Coupled-HMM [4] to detect interactions between multiple people. Buxton & Gong [5] adopted BBN for visual surveillance. However, there have been few studies on human activity recognition based on multiple sensory inputs. In [6] a Layered HMM architecture integrates information from multimodal inputs to recognize six activities in an office. One limitation of HMM is that it is only a computational model in the sense that HMM is not designed to be generated automatically from observations. Engineers have to manually design the system for given settings in advance, which affects its adaptive capability in unknown environments.

In this paper, we propose a real-time office presence detection system to monitor the behaviors of a single human in an office. A layered architecture is implemented. Currently, the overall system is hand-designed (computational model only) but a major part (IHDR) is automatically generated (computational model and model generator), which is the major difference between this work and [6]. This novel component is crucial for unknown environments where neither the vocabulary of conversation nor speaker population or phone type is known in advance (e.g. as a consumer product). The IHDR (Incremental Hierarchical Discriminant Regression) [7] tree automatically generates the representation for acoustic signals and incrementally learns new patterns without training from users, which is one of the essential requirements of adaptive applications. Based on visual-sensory input, low-level HMMs classify the human's motion pattern. High-level HMMs are implemented to integrate the outputs from the above two low-level components and to infer human's activities. In what follows, we first review the setting of the system. The system architecture is presented in Section 3. Then we discuss the experimental results and conclude with a summary and discussion about future works.

2. SETTING OF A CONTEXT-AWARE SYSTEM

A typical setting of a context-aware system is shown in Fig. 1. A table and a chair are in an office. On the table, there are a telephone, a personal computer, a video camera

[*] Thanks to Microsoft Research for funding.

and a microphone. Here are the sensors we use to collect context information: 1) Microphone: one mini-microphone with an audio-pickup range of up to $30'$-$40'$ is used for sound classification. 2) USB camera: a USB 2.0 camera sampled at 15 f.p.s. is used to detect human motions. For a computer, to recognize visual patterns and sound patterns is not an easy task. Inferring human activities from multiple sensors is more difficult since high-level reasoning is required. The following section shows how to build such a system through a layered architecture.

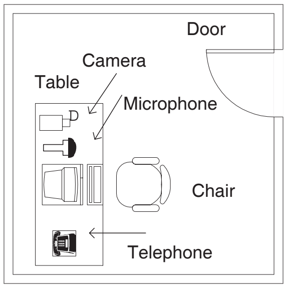

Fig. 1. Typical setting of a context-aware system.

3. SYSTEM ARCHITECTURE

The goal of this project is to detect human activities in an office. The system architecture is shown in Fig. 2. Two kinds of sensory inputs are used: auditory and visual. Cepstral analysis is applied to raw auditory signals to extract cepstral features, which are fed into an IHDR tree to discriminate four kinds of auditory patterns: "Phone ring," "Conversation," "Uncertain noise," and "Silence." A motion detector captures motion information by computing the difference between two consecutive images. A sequence of motion activities is classified by low-level HMMs into four motion patterns: "Rest," "Moving near door," "Moving in the office," and "Out." The outputs from the above two components are combined together and sent to high-level HMMs, which infer the human activities in the office. There are totally four kinds of activities: "Conversation," "Other activity," "Rest," and "Nobody around." The detailed design of each component is shown in Fig. 3 and explained in the following sections.

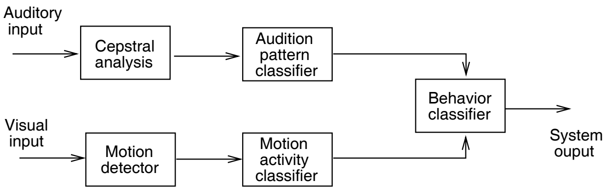

Fig. 2. Architecture of the context-aware system.

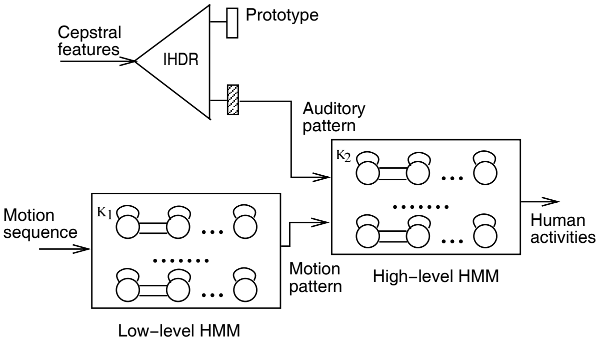

Fig. 3. A detailed architecture of the context aware system.

3.1. Auditory pattern classification using IHDR

The auditory data are digitized at 11025Hz by a normal sound blaster card. Cepstral analysis [8] is performed on the speech stream. 20 consecutive cepstral feature vectors together form a single auditory sensation vector, which covers about 0.4s. Continuous feature vectors are fed into an IHDR tree. No manually segmentation is needed. The function of the IHDR tree is to approximate a mapping $F : \mathcal{X} \mapsto \mathcal{Y}$ from a set of training samples $\{(x_i, y_i) \mid x_i \in \mathcal{X},\ y_i \in \mathcal{Y},\ i = 1, 2, \ldots, n\}$. The mapping is done through a coarse-to-fine tree structure. Each node of the tree is modeled by q Guassians. In this sense, the original d-dimensional (d=360) input space is mapped to a q-1 dimensional discriminant subspace. We only conduct Linear Discriminant Analysis (LDA) in the very-low dimensional subspace, which saves tremendous computational cost. Each Gaussian is represented by its first two-order statistics: mean and covariance matrix. Mean is updated incrementally as follows:

$$\bar{x}^{(n+1)} = \frac{n-\mu}{n+1}\bar{x}^{(n)} + \frac{1+\mu}{n+1}x_{n+1} \qquad (1)$$

where x_{n+1} is the $(n+1)$th sample, $\bar{x}^{(n+1)}$ is the mean after this sample is trained, μ is a parameter. If $\mu > 0$, the new input gets more weight than old inputs. We called this implementation the amnesic average. The covariance matrix can be updated incrementally by using the amnesic average too. Each leaf node generates quite a few primitive prototypes (block in Fig. 3), which represent different patterns (model generator). In testing phase, if a prototype is reached (shadowed block), the label associated with it would be the output (computational model). Furthermore, IHDR intrinsically has incremental online learning capability to adapt to new signals. This is why we use IHDR instead of traditional HMM to classify auditory patterns. For practical applications, the system has to work in different offices. Each office has different people and different telephones. In order to make HMMs adapt to new settings, a bank of new HMMs has to be created manually. In contrast, one IHDR is enough to handle all types of auditory signals.

3.2. HMMs for motion pattern classification

Discrete HMMs are implemented for behavior recognition based on motion. An HMM is denoted by $\lambda = (A, B, \pi)$, where A is state transition probability matrix, B is the observation symbol probability matrix, π is the initial state distribution. Specification of an HMM involves the choice of the number of states N, the number of observations M. With training data, we can calculate λ by using Baum-Welch algorithm [9]. Given a model λ and a sequence of observation O=$\{O_1, O_2, ..., O_T\}$, the likelihood of the sequence is

$$P(O|\lambda) = \Sigma_i \alpha_t(i), 1 \leq i \leq N, \tag{2}$$

where $\alpha_t(i)$ is defined as

$$\alpha_t(i) = [\sum_i \alpha_{t-1}(i) a_{ij}] b_j(O_t), \tag{3}$$

where a_{ij} is the element of A and $b_j(O_t)$ is the probability for state j in the model λ of observing O_t. Usually a bank of K HMMs (λ_k $(1 \leq k \leq K)$) would be generated. The likelihood of the observation sequence in each model is $L_k = P(O|\lambda_k)$. Suppose the maximal likelihood is L_{max} and the minimal likelihood is L_{min}, the normalized likelihood is

$$L_k^{'} = \frac{L_k - L_{\text{min}}}{L_{\text{max}} - L_{\text{min}}}. \tag{4}$$

HMMs choose the pattern $\hat{k}$ with the largest normalized likelihood as output.

$$\hat{k} = \arg\max_k \{L_k^{'}\} \tag{5}$$

3.3. Integration component

Integration of the above two low-level information to infer human activities is difficult since different modalities have different updating frequency and they can be either related or unrelated. High-level HMMs are necessary for reasoning human activities based on information of low-level components. Usually the reasoning is conducted with a larger time granularity (for example, 3 seconds), while in motion pattern classification and auditory pattern classification, the granularity is 1 second. The outputs from low-level component are symbols, which can be fed into the integration HMMs. Since the vision component outputs 4 types of different patterns (so does the audition component), the combination of these two components gives 16 types of observations.

4. EXPERIMENTAL RESULTS

We conducted experiments for each of these three components.

4.1. Experimental results of the auditory component

We trained the IHDR with 3 kinds of male conversation signals, 2 kinds of female conversation signals, 12 kinds of phone ring signals and uncertain noise. The number of cepstral feature is about 12000. A half is used for training, another half is used for testing. Results are shown in Tab. 1. Conversation and uncertain noise signals are confusing sometimes (206 conversation vectors are recognized as uncertain noise). The overall recognition rate is 94.75%. We need to notice that the test is source-dependent. In other words, the testing set and the training set come from the same source. If we test the system with signals of a telephone we never trained, the performance would definitely drop. That's why IHDR is important for this application since it can incrementally learn new auditory patterns, while it may not be so easy for HMMs.

Table 1. Recognition rate of each auditory set. (C=Conversation; UN=Uncertain Noise; P=Phone; S=Silence)

Data	C	UN	P	S	Total	Rate
C	2253	206	7	4	2470	91.21%
UN	230	2090	0	0	2320	90.07%
P	9	19	1762	5	1795	98.16%
S	0	0	0	1141	1141	100%

4.2. Experimental results of the motion component

In this experiment, the parameters of the low-level HMMs are: N=6, M=4 and K=4. Different observations are: "Motion in door area," "Motion in room but not door," "Static & last motion in door area," and "Static & last motion in room but not door." The normalized likelihood of each motion pattern is shown in the first four plots of Fig. 4. The x-axis is time line about 400 seconds. The granularity of HMMs is 1 second. If $L_k^{'} = 1$, then pattern k is reported. The ground truth of activity sequences is shown in the fifth plot, which goes as follows: the user firstly moved in the room (pattern 1), rested for a while (2), moved around the door (3) and then went out (4). The motion behaviors are clearly recognized. A mistake occurs around step 320, the system classified pattern (3) as pattern (1) when the user moved near the boundary of the door and the remain parts of the room.

4.3. Experimental results of the high-level reasoning

The specifications of the high level HMMs are: N=4, M=16 and K=4. Normalized likelihood of each human activity is shown in the first four plots of Fig. 5. The ground truth of activity sequences is in the fifth plot, which goes like this: the user moved around in the rest (activity 1), rested (2), moved around again (1), talked for a while (3), moved to the door (1) and went out (4). The x-axis is time line about 800 time frames. "Nobody" and "Rest" are perfectly classified,

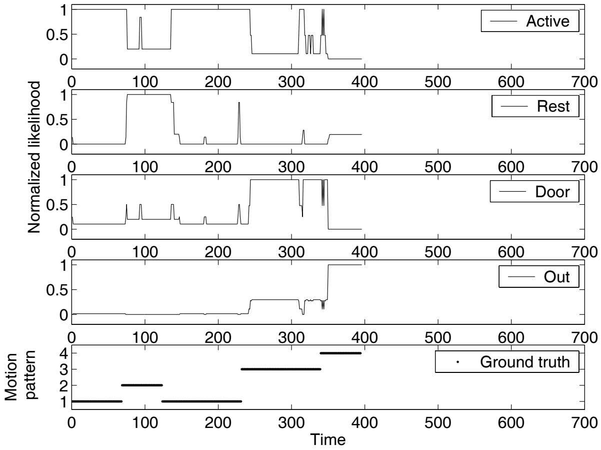

Fig. 4. Likelihood of motion patterns over time.

while "Conversation" and "Other activity" are messed a little because sometimes you can move and talk at the same time. 40 minutes of office activity are recorded (about 10 minutes for each activity). A half of the data is used for training; another half is used for testing. The recognition rate of human activities is shown in Tab. 2. About 5% of "Conversation" is incorrectly recognized as "Other activity", which is consistent with the results in Fig. 5.

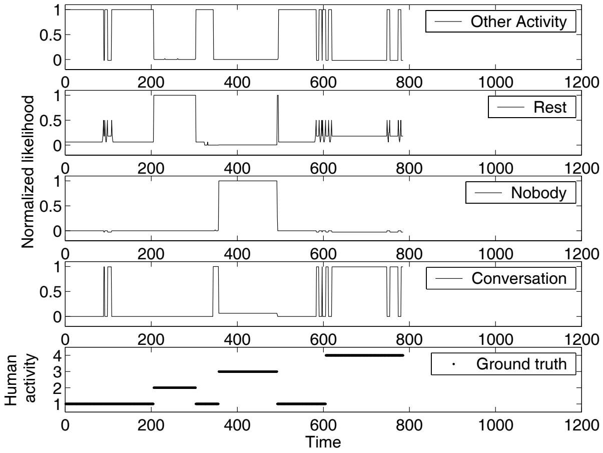

Fig. 5. Likelihood of human activities over time.

5. SUMMARY AND DISCUSSION

In this paper, a layered architecture is proposed for the detection of office presence. IHDR, as a model generator, generates representation for different auditory patterns and can easily adapt to new settings. Two levels of HMMs with different granularities handle motion pattern classification and integration of low-level outputs, respectively. The initial results of the prototype system are promising. However, to build a highly adaptive system for consumers needs further study. Here are some future works: 1) Maximize the adaptive capability. We plan to make all major components of the system generative using IHDR. 2) Minimize the user on-site training so that most adaptation is performed silently and automatically in the background. 3) Dynamic CPU usage adaptation so that the system does not tie up CPU cycles noticeably.

Table 2. Recognition rate of human activities (N=Nobody; OA=Other activity; R=Rest; C=Conversation).

Data	N	O	R	C
N	100%	0	0	0
O	0	94.31%	0	5.69%
R	0	0	100%	0
C	0	4.52%	0	95.48%

6. REFERENCES

[1] S. Shafer, B. Brumitt, and J. Cadiz, "Interaction issues in context-aware interactive environments," *Human Computer Interaction*, vol. 16, 2001.

[2] J. Yamato, J. Ohya, and K. Ishii, "Recognizing human action in time-seqential images using hidden markov model," in *Proc. of Int. Conf. on Computer Vision and Pattern Recognition*, 1992, pp. 379–385.

[3] A. Galata, N. Johnson, and D. Hogg, "Learning variable length markov models of behaviour," *Computer Vision and Image Understanding*, vol. 81, no. 3, pp. 398–413, 2001.

[4] M. Brand, N. Oliver, and A. Pentland, "Coupled hidden markov models for modeling interacting processes," in *Proc. of Int. Conf. on Computer Vision and Pattern Recognition*, 1996, pp. 994 – 999.

[5] H. Buxton and S. Gong, "Advanced Visual Surveillance using Bayesian Networks," in *Proc. of Int. Conf. on Computer Vision*, Cambridge, Massachusetts, June 1995, pp. 111–123.

[6] N. Oliver, E. Horvitz, and A. Garg, "Layered representation for human activity recognition," in *Proc. of Int. Conf. on Multimodal Interfaces*, 2002, pp. 3–8.

[7] W. Hwang and J. Weng, "Hierachical discriminant regression," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 22, no. 11, pp. 1277 –1293, 2000.

[8] J. Deller, J. Proakis, and J. Hansen, *Discrete-time processing of speech signals*, Institute of Electrical and Electronics Engineers Press, New York, 2000.

[9] L. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings IEEE*, vol. 77, no. 2, pp. 257–286, 1989.