2007

Multimodal Diaries

Fernando De la Torre
Carnegie Mellon University

Carlos Agell
Carnegie Mellon University

MULTIMODAL DIARIES

Fernando De la Torre *Carlos Agell*

Robotics Institute. Carnegie Mellon University. 5000 Forbes Av., Pittsburgh, PA 15213 USA

ABSTRACT

Time management is an important aspect of a successful professional life. In order to have a better understanding of where our time goes, we propose a system that summarizes the user's daily activity (e.g. sleeping, walking, working on the computer, talking, ...) using all-day multimodal data recordings. Two main novelties are proposed:

- A system that combines both physical and contextual awareness hardware and software. It records synchronized audio, video, body sensors, GPS and computer monitoring data.
- A semi-supervised temporal clustering (SSTC) algorithm that accurately and efficiently groups large amounts of multimodal data into different activities.

The effectiveness and accuracy of our SSTC is demonstrated in synthetic and real examples of activity segmentation from multimodal data gathered over long periods of time.

1. INTRODUCTION

Effective time management is an important aspect of a successful professional life. Many techniques exist to effectively manage your time, and several of them include the popular "to-do list", that is, making an inventory of your daily schedule. In this paper, we propose a system that summarizes the user's daily activity from multimodal sensors. Building a system that can compute statistics on the activities done over the day is a first step towards building intelligent personal agents able to manage time more efficiently.

Two main novelties are discussed. Firstly, a system that combines both physical and contextual awareness sensors to record synchronized audio-visual, body sensing, global position and computer monitoring data is described; secondly, we propose a semi-supervised clustering algorithm able to temporally segment multimodal data efficiently and accurately. Fig. 1 shows an example of an office scenario recording, where several synchronized modalities are used to summarize the user's activities.

2. SENSORS

In this section, we describe the sensors and the features used for temporal segmentation of activities.

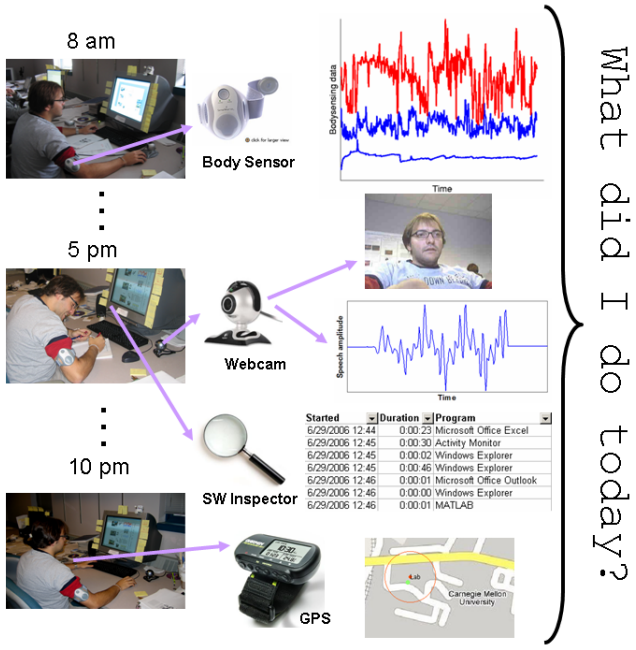

Fig. 1. Syncronized multimodal recording (body sensor, video, audio, computer monitoring and GPS.)

2.1. Physical Awareness Sensors

We use the SenseWear armband from BodyMedia (fig. 2.a) as a physical awareness device. The SenseWear armband [1, 2] combines five different sensors: 2-axis accelerometers, galvanic skin response, skin temperature, heat flux, and near-body ambient temperature. The SenseWear is synchronized with the computer, and it is worn during the entire day on the upper right arm. We have used sleeping, physical activity, lying down, and standing up as features for activity segmentation.

2.2. Context Awareness Sensors

For context awareness, we have used the Logiteck Quickcam camera to record audio and video, a computer monitoring software to record the programs that the user is running, and a wearable GPS device (see fig. 1)

The video features consist of a binary stream, where 1 corresponds to detecting the user's face and 0 otherwise. We use the OpenCV face detector from Viola and Jones [3], and color filters to reduce the number of false positives. To extract

Fig. 2. a) Body sensor fixed on the upper arm. b) Wearable GPS Personal Navigator.

audio features, we compute the Mel Frequency Cepstral Coeffcients (MFCC) over 20 ms. These features are used to train a Support Vector Machine (SVM) [4] for classifying the audio signal into four states: user talking, other people talking, typing, and silence.

To record the interaction of the user with the computer we use Activity Monitor from Softactivity. We classify at 1 Hz the programs that the user is running into three categories: work, non-work and internet surfing. Every time an unclassified program shows up the user is prompted to classify it into one of those groups.

To track people outdoors, we use a wearable wrist-strap GPS receiver. The GARMIN Foretrex 201 has an accuracy of 15 m and an updating frequency of 1Hz, see fig. 2(b). Using the coordinates logged by this device (longitude and latitude) and the time stamp, we can estimate the mean speed.

3. SEMI-SUPERVISED SPATIO-TEMPORAL CLUSTERING (SSTC)

Given the set of multimodal features described in the previous section, our goal is to segment the data into temporally coherent chunks. In this section, we extend standard clustering algorithms (e.g. K-means or spectral graph methods) to incorporate temporal coherence and semi-supervised information.

3.1. Discriminative Cluster Analysis (DCA)

DCA [5] is a clustering method that combines both clustering and discriminative dimensionality reduction in an unsupervised manner. DCA minimizes:

$$E_{DCA}(\mathbf{B},\mathbf{V},\mathbf{G}) = ||(\mathbf{G}^T\mathbf{G})^{-\frac{1}{2}}(\mathbf{G}^T - \mathbf{V}\mathbf{B}^T\mathbf{D})||_F \quad (1)$$

where $\mathbf{D} \in \Re^{d\times n}$ (see notation[1]) is a data matrix such that each column $\mathbf{d}_i$ corresponds to a sample of multimodal features at one time instant. $\mathbf{G} \in \Re^{n\times c}$ is a dummy indicator matrix, such that $\sum_j g_{ij} = 1$, $g_{ij} \in \{0,1\}$ and g_{ij} is 1 if $\mathbf{d}_i$ belongs to class C_j, c denotes the number of classes and n the number of samples. $\mathbf{B} \in \Re^{d\times k}$ and $\mathbf{V} \in \Re^{c\times k}$ are reduced rank approximation matrices. Considering the simpler case where $\mathbf{B} = \mathbf{I}_d$, after eliminating $\mathbf{V}$, eq. 1 is proportional to:

$$E_{DCA}(\mathbf{G}) \propto tr(\mathbf{D}^T(\mathbf{D}\mathbf{D}^T)^{-1}\mathbf{D}\mathbf{G}(\mathbf{G}^T\mathbf{G})^{-1}\mathbf{G}^T) \quad (2)$$

Relaxing the constraints on $\mathbf{G}$, so that $g_{ij} \geq 0$ and $\mathbf{G}\mathbf{1}_c = \mathbf{1}_n$, a gradient descent strategy can efficiently find a local optimum of eq. 2, see [5] for details.

3.2. Temporal term

DCA does not take into account any temporal coherence of the cluster labels or incorporate semi-supervised information. One of the benefits of relating the clustering problem to the optimization of an objective function (e.g. [5]) is that we can easily add additional constraints as a penalty term.

In order to penalize non-smooth changes (over time) on the labels, we encourage that $\mathbf{g}_i$ and $\mathbf{g}_{i+1}$ have similar values by minimizing: $E_t = \sum_{i=1}^{n-1} ||\mathbf{g}_i - \mathbf{g}_{i+1}||_2^2 = ||\mathbf{G}^T - \mathbf{G}^T\mathbf{P}||_F$, where $\mathbf{P}$ is a known permutation matrix (left shift of the identity matrix). Moreover, adding dynamic information and a normalization factor, the temporal term transforms to:

$$E_t = ||(\mathbf{G}^T\mathbf{G})^{-\frac{1}{2}}(\mathbf{G}^T - \mathbf{A}\mathbf{G}^T\mathbf{P})||_F \quad (3)$$

$\mathbf{A}$ encodes the state dynamics and the matrix $(\mathbf{G}^T\mathbf{G})^{-\frac{1}{2}}$ is a normalization factor for DCA. If $\mathbf{G}$ is known, the optimal $\mathbf{A}$ can be computed as: $\mathbf{A} = \mathbf{G}^T\mathbf{P}^T\mathbf{G}(\mathbf{G}^T\mathbf{G})^{-1}$.

3.3. Adding semi-supervised information

In this section, we add two types of semi-supervised information to the clustering: the *must-link* term and the *cannot-link* term.

Let $\mathbb{N}^s$ be the set of samples that belong to the same class. $\mathbf{e}_r \in \Re^n$ denotes an indicator vector for data point $\mathbf{d}_r$ so that $\mathbf{D}\mathbf{e}_r = \mathbf{d}_r$. We formulate the must-link supervised additive term as follows:

$$E_{s_{ML}} = \sum_{i,j\in\mathbb{N}^S} ||\mathbf{g}_i - \mathbf{g}_j||_2^2 = ||\mathbf{G}\mathbf{E}_{ML}||_F \quad (4)$$

where $\mathbf{g}_i - \mathbf{g}_j = \mathbf{G}(\mathbf{e}_i - \mathbf{e}_j)$ and $E_{s_{ML}} \in \Re^{c\times l}$ is a matrix with l columns corresponding to the number of pairs of data points that belong together, and each column contains the vector $\mathbf{e}_i - \mathbf{e}_j \in \mathbb{N}^s$ that defines a pair of must-link points.

An analogous cannot-link term ($E_{s_{CL}}$) can be defined out of the set $\mathbb{N}^D$ of cannot-link pairs, defining $\mathbf{E}_{CL} \in \Re^{c\times l_2}$ as a matrix, where each column contains the vector $\mathbf{e}_i - \mathbf{e}_j$ that defines a pair of cannot-link points.

[1] Bold capital letters denote a matrix $\mathbf{D}$, bold lower-case indicates a column vector $\mathbf{d}$. $\mathbf{d}_j$ represents the j column of the matrix $\mathbf{D}$. d_{ij} denotes the scalar in the row i and column j of the matrix $\mathbf{D}$ and the scalar i-th element of a column vector $\mathbf{d}_j$. All non-bold letters represent variables of scalar nature. $\mathbf{I}_k \in \Re^{k\times k}$ is the identity matrix. $tr(\mathbf{A}) = \sum_i a_{ii}$ is the trace of the matrix $\mathbf{A}$. $||\mathbf{A}||_F = tr(\mathbf{A}^T\mathbf{A}) = tr(\mathbf{A}\mathbf{A}^T)$ designates the Frobenious norm of a matrix.

3.4. Optimization

Combining all the terms, the semi-supervised spatio temporal clustering algorithm optimizes:

$$E_{sstc}(\mathbf{G}) = E_{cluster} + \lambda E_t + \beta_1 E_{s_{ML}} - \beta_2 E_{s_{CL}} \quad (5)$$

$E_{cluster}$ can be K-means, DCA or spectral clustering, see [5] for the details. The parameters $\lambda, \beta_1, \beta_2$ are normalization factors to make E_t, $E_{s_{ML}}$, $E_{s_{CL}}$ and $E_{cluster}$ comparable in terms of energy.

To cluster, we perform gradient descent in eq. 5 with a line search strategy. To impose non-negativity constraints on g_{ij}, we parameterize $\mathbf{G}$ as the Hadamard product of two matrices $\mathbf{G} = \mathbf{V} \circ \mathbf{V}$ [5] and the updates are given by the following expressions:

$$\mathbf{V}^{n+1} = \mathbf{V}^n - \eta \left(\frac{\partial E_{cluster}(\mathbf{V}^n)}{\partial \mathbf{V}} + \lambda \frac{\partial E_t(\mathbf{V}^n)}{\partial \mathbf{V}} + \beta_1 \frac{\partial E_{s_{ML}}(\mathbf{V}^n)}{\partial \mathbf{V}} - \beta_2 \frac{\partial E_{s_{CL}}(\mathbf{V}^n)}{\partial \mathbf{V}} \right) \quad (6)$$

$$\frac{\partial E_t(\mathbf{V}^n)}{\partial \mathbf{V}} = \mathbf{V} \circ (2\mathbf{G}\mathbf{A}^T(\mathbf{G}^T\mathbf{G})^{-1}\mathbf{A} \ldots$$
$$-2\mathbf{G}(\mathbf{G}^T\mathbf{G})^{-1}\mathbf{A}\mathbf{G}^T\mathbf{G}\mathbf{A}^T(\mathbf{G}^T\mathbf{G})^{-1} \ldots$$
$$+4\mathbf{G}(\mathbf{G}^T\mathbf{G})^{-1}\mathbf{G}^T\mathbf{P}\mathbf{G}\mathbf{A}(\mathbf{G}^T\mathbf{G})^{-1} \ldots$$
$$-2\mathbf{P}^T\mathbf{G}\mathbf{A}^T(\mathbf{G}^T\mathbf{G})^{-1} - 2\mathbf{P}\mathbf{G}(\mathbf{G}^T\mathbf{G})^{-1}\mathbf{A})$$
$$\frac{\partial E_{s_i}(\mathbf{V}^n)}{\partial \mathbf{V}} = \mathbf{V} \circ (2\mathbf{E}_i\mathbf{E}_i^T\mathbf{G})$$

Optimizing eq. 5 w.r.t $\mathbf{G}$ is a non-convex optimization problem that, without a good starting point, is likely to get stuck at a local minimum. To improve clustering results, we use a top-down approach where a multiresolution scheme is employed. That is, we first decimate the data and apply the clustering scheme at the lowest resolution level and propagate the result to higher levels. The multiresolution scheme has two main benefits: it is faster and more accurate; and, the first order temporal constraints (i.e. E_t) imposed in the lower resolution are expanded to higher order terms in the full resolution.

4. EXPERIMENTS

In this section, we report results in both synthetic and real examples of the proposed semi-supervised temporal clustering.

4.1. Synthetic experiments

Fig. 3 shows a two-level piece-wise constant signal at levels 5 and 10, with added Gaussian noise (N(0,0.3)) and some glitches. These glitches occur naturally in our system by the discontinuity in the audio-visual classifiers. Ideally, we would like to segment this 1D signal into a square wave. Using standard DCA or k-means does not lead to a correct clustering because of the glitches and noise (see fig. 3 top). However, the multiresolution version of DCA with temporal consistency finds the desired solution (see fig. 3 bottom).

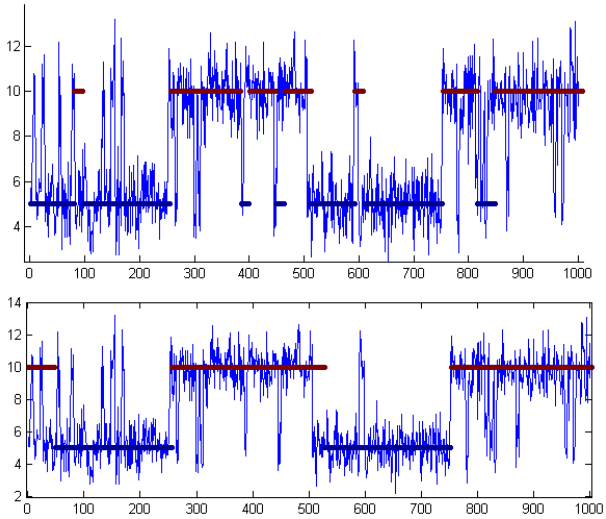

Fig. 3. Results of DCA and DCA with temporal coherence.

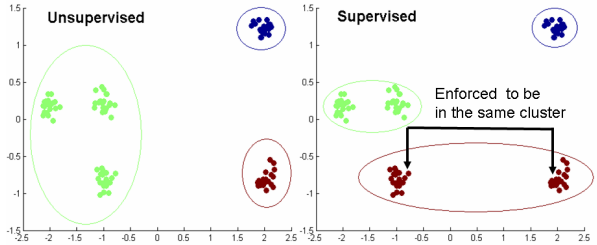

Fig. 4. Left: natural clustering. Right: clustering with semi-supervised information.

Fig. 4 illustrates the use of the semi-supervised term. Fig. 4 shows an example where five 2-dimensional Gaussians can be clustered differently on two clusters based on the supervised term (*must-link* pairwise constraint).

4.2. What did I do today?

For many people (including the authors), it often seems that at the end of the day not all the expected work has been done. Inevitably, the same question comes to mind: What did I do today?. In this section, we use the SSTC algorithm to segment our daily activity from multimodal data. Later, we provide statistics of the time spend in each task for user self-awareness (e.g. amount of time doing low-value jobs such as reading junk e-mail).

We have collected data every second, over a period of three days, in an office scenario for two different people. Fig. 1 shows a typical example of all the data gathered at a particular time instant. All this multimodal data (range from five to nine hours for each person per day) is manually classified into eight types of activity: sleeping, walking, away (inside the building), away (outside the building), working(no PC), internet surfing, working on computer and talking.

We use our SSTC algorithm to temporally cluster this

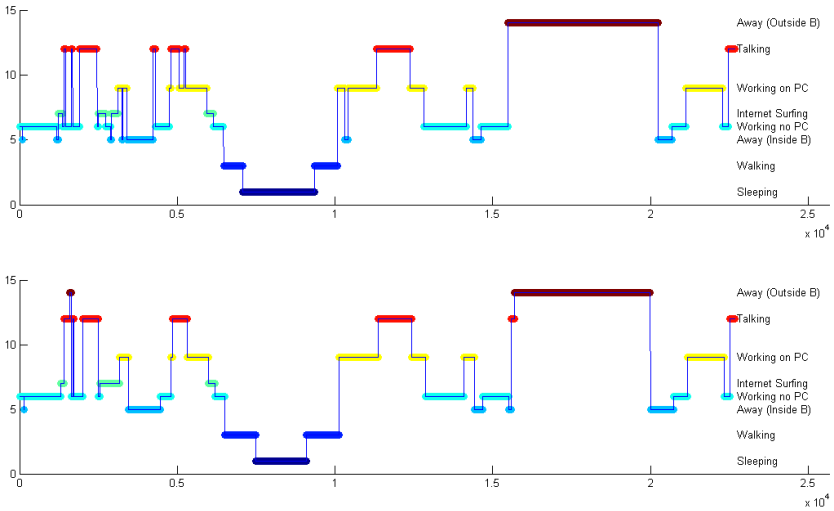

Fig. 5. Labeled Data and Output of the clustering algorithm

data. The program runs on Matlab, and we use a multiresolution strategy with 7 levels. Fig. 5.b shows the results of using semi-supervised spatio-temporal DCA clustering for six hours and 10 minutes of recording.

We compare the results obtained from several clustering methods: k-means, DCA, DCA + temporal term (DCA+TT) and DCA + temporal + semi-supervised term (DCA+TT+SST). The accuracy of the clustering is given by the number of correct samples over the total number of samples. This accuracy measure requires correct and precise labeling information for each day (user annotated data). Table 1 shows the clustering accuracy for all the algorithms described in section 4 with the data collected. A video with the results of the clustering for a particular user can be downloaded from $www.cs.cmu.edu/\sim ftorre/IcmeVideo.mpg$.

Algorithm	one level	Multiresolution
K-means	69.23 $\pm$ 1.9 %	–
DCA	74.01 $\pm$ 1.2 %	75.51 $\pm$ 4.4%
DCA+TT	76.13 $\pm$ 1.3 %	77.79 $\pm$ 5.0 %
DCA+TT+SST	83.42 $\pm$ 1.1 %	89.51 $\pm$ 4.8 %

Table 1. Accuracy for the different clustering methods.

It is also interesting to analyze which activities are easier to cluster. Table 2 reports the clustering accuracy for each of the activities.

5. CONCLUSION

In this paper, we have proposed a context and physical awareness system to monitor the daily activities of a user. To temporally segment the activities, we have extended traditional clustering algorithms by adding side information and temporal consistency to the clusters. We are currently working on extending the number of activities and analyzing which factors make the user more productive.

State	Accuracy
Sleeping	64.05$\pm$5.2%
Walking	91.02$\pm$2.1%
Working no PC	69.45$\pm$2.5%
Internet Surfing	89.78$\pm$1.4%
Working PC	92.10$\pm$2.7%
Talking	60.26$\pm$8l.1%
Away(inside)	73.21$\pm$6.3%
Away(outside)	94.12$\pm$1.3%

Table 2. Activity accuracy.

Acknowledgements This work has been partially supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. NBCHD030010.

6. REFERENCES

[1] A. Krause, D. Siewiorek, A. Smailagic, and J. Farringdon, "Unsupervised, dynamic identification of physiological and activity context in wearable computing," *International Symposium on Wearable Computers*, vol. 00, pp. 88, 2003.

[2] A. Teller and J. Stivoric, "The bodymedia platform: continuous body intelligence," in *Workshop on Continuous archival and retrieval of personal experiences*, 2004.

[3] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *CVPR*, 2001.

[4] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*, Cambridge university press, 2000.

[5] F. De la Torre and T. Kanade, "Discriminative cluster analysis," in *International Conference on Machine Learning*, 2006.