Motif Discovery and Feature Selection for CRF-Based Activity Recognition

Liyue Zhao[1] Xi Wang[1] Gita Sukthankar[1] Rahul Sukthankar[2,3]
zhaoliyue712@gmail.com wxjennifer@gmail.com gitars@eecs.ucf.edu rahuls@cs.cmu.edu
[1]*University of Central Florida* [2]*Intel Labs Pittsburgh* [3]*Carnegie Mellon University*

***Abstract*—Due to their ability to model sequential data without making unnecessary independence assumptions, conditional random fields (CRFs) have become an increasingly popular discriminative model for human activity recognition. However, how to represent signal sensor data to achieve the best classification performance within a CRF model is not obvious. This paper presents a framework for extracting motif features for CRF-based classification of IMU (inertial measurement unit) data. To do this, we convert the signal data into a set of motifs, approximately repeated symbolic subsequences, for each dimension of IMU data. These motifs leverage structure in the data and serve as the basis to generate a large candidate set of features from the multi-dimensional raw data. By measuring reductions in the conditional log-likelihood error of the training samples, we can select features and train a CRF classifier to recognize human activities. An evaluation of our classifier on the CMU Multi-Modal Activity Database reveals that it outperforms the CRF-classifier trained on the raw features as well as other standard classifiers used in prior work.**

***Keywords*-activity recognition, feature selection, CRF**

I. Introduction

Human activity recognition has become an increasingly important component of many domains such as user interfaces and video surveillance. In particular, enabling ubiquitous user assistance systems for elder care requires rapid and robust recognition of human action from portable sensor data. The CMU Multimodal Activity database [2] was collected to facilitate the comparison of different activity recognition techniques for recognizing household activities. The dataset contains video, audio, inertial measurement unit (IMU), and motion capture data. In this paper, we focus on the sole use of the IMU data, as it is a more cost-effective and less invasive sensor (see Figure 1). The primary disadvantage is that the IMU data is noisier and less informative than the richer motion capture and video data, making the classification problem substantially more difficult.

Prior work [7] has examined the relative benefits of using discriminative conditional random fields (CRFs) vs. commonly-used generative models (e.g., the Hidden Markov Models) on diverse activity recognition problems such as video recognition and analyzing Robocup team behavior. The consensus has been that many of features commonly used in activity recognition problems are based on overlapping time windows that nullify the observational independence assumptions of many generative graphical models. However, feature representation and selection impact both the performance and computational accuracy of CRFs. In this paper, we examine the problem of automatically discovering and selecting features to improve the performance of a CRF-based classification. The key to our method is the automatic discovery of repetitive and informative subsequences, *motifs*, in the noisy IMU data. We suggest that re-representing the signal in the form of motifs facilitates the creation of the candidate set of features.

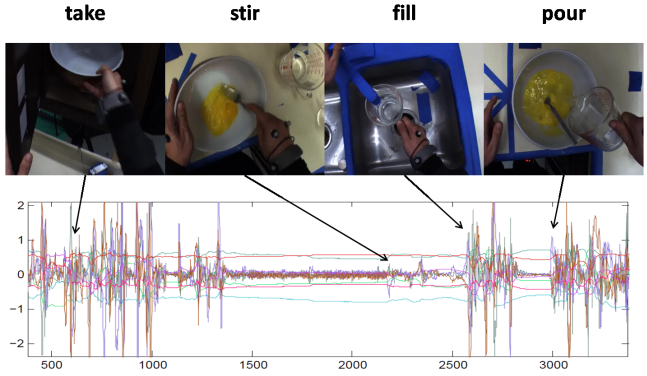

Fig. 1. CMU-MMAC IMU dataset: example actions and corresponding data.

II. Related Work

Unsupervised motif discovery has been demonstrated as a promising technique for identifying and matching imperfect repetitions in time series data. Researchers have primarily addressed two problems: 1) initial motif selection criteria [8] and 2) rapid and robust matching [1]. Another issue is how to generalize single-dimensional motif techniques to multi-dimensional data [9]. Our work takes a novel approach to this problem by identifying and matching motifs in each dimension separately and using the conditional random field to learn the linkage between motif occurrences across different dimensions and action class labels. In prior work [6], motifs have been used in conjunction with HMMs to distinguish six arm techniques from a single IMU, but the assumption was made that each action could be characterized by a single six-dimensional motif. In contrast, our data is substantially more complicated (full body data generated from from five IMU sensors) and our action set is composed of 14 unscripted actions performed during a cooking task (e.g., “stir brownie mix”, “pour mix in pan”, “get fork”, “break eggs”, and “pour oil”).

1051-4651/10 $26.00 © 2010 IEEE
DOI 10.1109/ICPR.2010.932

III. METHOD

The goal of action recognition in this context is to assign tags to each frame of IMU data (from among the set of 14 actions or "none of the above"). Since these actions were performed as part of a higher-level task (cooking a brownie), one should be able to exploit the temporal dependencies between actions (e.g., "get eggs" is likely to precede "break an egg").

Our training approach can be described as follows. First, we discover motifs in the data collection, essentially learning a mapping to convert a given local window of IMU data from a multi-dimensional time series signal to a sequence of discrete symbols. Second, we define a series of low-level binary-valued features over motifs and pairs of motifs. From a large pool of candidate features, we select those that are most informative using an iterative approach, described below. Next, we learn a Conditional Random Field whose observations are defined over the set of selected features and whose output is over the set of action labels. The incremental feature selection and CRF training are iterated until the training set error converges. The final CRF is then evaluated on the test set. Each of these stages is detailed below.

A. Motif discovery

The first step in motif discovery is to discretize the continuous IMU signal into symbolic subsequences. Figure 2(b) illustrates this process. The raw data $T = \{t_1, t_2, \ldots, t_n\}$ (black line) is transformed into a piecewise continuous representation $S = \{s_1, s_2, \ldots, s_m\}$ (green line) using the Piecewise Aggregate Approximation (PAA) algorithm, which computes an average of the signal over a short window. This is then mapped to a symbolic representation using "break points" (red lines) that correspond to bins; these are generated so as to separate the normalized subsequences (under a Gaussian assumption) into equalized regions. Thus, a continuous 1-D signal is represented as a sequence of discrete symbols.

To compare symbolic sequences in a manner that is both computationally efficient and robust to signal noise (i.e., corresponding to symbol mismatch), we propose a matching metric that relies on random projections. In our problem, we designate two motifs as matching if they agree on k symbol positions. Figure 2(c) gives an example with $k = 2$, where the randomly-selected columns 1 and 3 are used to compare motifs. In this example, motifs 1 and k, 2 and 3, and 4 and j all match. These matches can be summarized by incrementing entries in a symmetric match table (where rows and columns correspond to motifs), as shown in Figure 2(d). Accumulating counts in this manner using several different random projections can enable us to efficiently match long motifs in a manner that is robust to occasional symbol errors.

B. Feature selection for CRFs

A conditional random field (CRF) is an undirected graphical model $G = (V, E)$ that represents the conditional probabilities of a label sequence $\mathbf{y} = \{y_1, \ldots, y_T\}$, given the observation sequence $\mathbf{x} = \{x_1, \ldots, x_T\}$. For our problem, the labels correspond to activities, such as "open fridge", while the

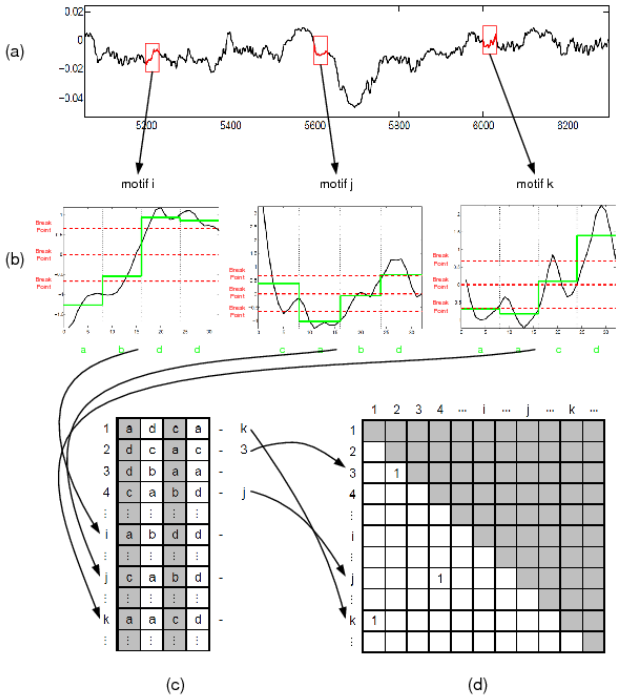

Fig. 2. Motif discovery (see text for details).

observations consist of a set of features computed over the raw data. In this paper, we focus on CRFs with a linear chain structure, where the current state is connected only to its temporally-neighboring states and the current observation.

Since CRFs are log-linear models, the potential function can be expressed as the linear sum of feature functions as $\exp\{\sum_j w_j f_j(y_{i-1}, y_i, \mathbf{x}, i)\}$, where w_j represents the weight corresponding to feature $f_j(y_{i-1}, y_i, \mathbf{x}, i)$. In this work, we learn both the features $f_j(.)$ (described in the next section) and their associated weight parameters w_j, detailed below.

Consider an annotated dataset with pairs of training samples $(\mathbf{X}, \mathbf{Y}) = \{(\mathbf{x}^{(1)}, \mathbf{y}^{(1)}), \ldots, (\mathbf{x}^{(N)}, \mathbf{y}^{(N)})\}$. The CRF weight vector $\mathbf{w} = \{w_1, ..., w_M\}$ can be obtained by optimizing the sum of conditional log-likelihood $L(\mathbf{Y}|\mathbf{X}; \mathbf{w})$. Then CRF training is, given a set of training samples $\mathbf{X}$ and $\mathbf{Y}$, finding the values for $\mathbf{w}$ that maximize the first derivative of the log-likelihood:

$$\frac{\partial L(\mathbf{Y}|\mathbf{X}; \mathbf{w})}{\partial w_j} = \sum_k \{ \sum_{f_j \in \mathcal{S}} F_j(y_{i-1}^{(k)}, y_i^{(k)}, \mathbf{x}^{(k)}, i) - \sum_y \sum_{f_j \in \mathcal{S}} p(\mathbf{y}|\mathbf{x}^{(k)}; \mathbf{w}) F_j(y_{i-1}, y_i, \mathbf{x}^{(k)}, i) \},$$

where $\mathcal{S}$ denotes the selected feature set (described below) and $F_j(y_{i-1}^{(k)}, y_i^{(k)}, \mathbf{x}^{(k)}, i)$ is defined as

$\sum_{i=1}^{T} f_j(y_{i-1}^{(k)}, y_i^{(k)}, \mathbf{x}^{(k)}, i)$. Since the CRF log likelihood is convex with respect to the weight vector $\mathbf{w}$, this training can be accomplished using standard optimization techniques such as conjugate gradient and limited memory BFGS [5].

A key aspect of the proposed method is that we automatically select informative features from a large pool of candidates defined over motifs. As validated in our experiments, this leads to a significant improvement over CRFs trained directly on the raw data. We define three types of binary features over our motifs to form a pool of over 4000 features, from which our goal is to select a small subset that can maximize the conditional log likelihood, without overfitting to the training data.

We adopt the following greedy forward selection procedure, where each feature in the pool is considered in turn and the best feature at each iteration is added. Specifically, we initialize the candidate set $\mathcal{C}$ with the pool of available features and the subset of selected features $\mathcal{S}$ to be empty. At each iteration, we evaluate every potential feature $f_\lambda \in \mathcal{C}$ individually by considering a CRF with $\mathcal{S} \cup f_\lambda$ and selecting the feature that maximizes the log-likelihood gain, $G(\lambda, f_\lambda) = L(\mathbf{Y}|\mathbf{X}; \mathbf{w}, \lambda) - L(\mathbf{Y}|\mathbf{X}; \mathbf{w})$, where λ denotes the weight corresponding to the potential new feature f_λ. This best feature is added to $\mathcal{S}$ and removed from $\mathcal{C}$. We continue selecting features until the CRF error (computed on a hold out set) ceases to improve.

Unfortunately, a straightforward implementation of this procedure is extremely time consuming since it requires an expensive computation for every potential feature at each iteration. In particular, the normalization term of the CRF, $Z(\mathbf{x}^{(k)})$ must be calculated every time the gain $G(\lambda, f_\lambda)$ is evaluated. Motivated by work on kernel CRFs [3] and image segmentation [4], we employ a first-order approximation method. The log likelihood function $L(\mathbf{y}|\mathbf{x}; \mathbf{w}, \lambda)$ can be approximated by its first-order Taylor expansion:

$$L(\mathbf{Y}|\mathbf{X}; \mathbf{w}, \lambda) = L(\mathbf{Y}|\mathbf{X}; \mathbf{w}) + \lambda \left[\frac{\partial L(\mathbf{Y}|\mathbf{X}; \mathbf{w}, \lambda)}{\partial \lambda} \mid_{\lambda=0} \right].$$

In this equation, the second term can be expressed as:

$$\frac{\partial L(\mathbf{Y}|\mathbf{X}; \mathbf{w}, \lambda)}{\partial \lambda} \mid_{\lambda=0} = E[f_\lambda, \lambda] - \tilde{E}[f_\lambda, \lambda],$$

where $\tilde{E}[f_\lambda, \lambda] = \sum_k \sum_{i=1}^{T} f_\lambda(y_{i-1}^{(k)}, y_i^{(k)}, \mathbf{x}^{(k)}, i)$ represents the empirical expectation and $E[f_\lambda, \lambda] = \sum_k \sum_{i=1}^{T} \sum_{y'} p(y'|\mathbf{x}^{(k)}, \mathbf{w}, \lambda) f_\lambda(y_{i-1}, y_i, \mathbf{x}^{(k)}, i)$ is the model expectation. Employing this approximation achieves significant computational benefits in practice.

Our proposed method is agnostic to the choice of features. Motivated by Vail et al.'s work on activity recognition for robots [10], we employ the following three types of features. In our case, these are computed over motif patterns rather than the raw data, and all are two-valued features. The indicator function $\delta(.)$ is 1 if its argument is true and 0 otherwise.

1) **Identification features**
   $f(y_{i-1}, y_i, \mathbf{X}, i) = \delta(y_i = \text{motif}_k)$. These features constitute the basic units of actions and are computed at a node level. They verify that the action label at time t corresponds to motif k.
2) **Transition features**
   $f(y_{i-1}, y_i, \mathbf{X}, i) = \delta(y_{i-1} = \text{motif}_j)\delta(y_i = \text{motif}_k)$. These capture the first-order Markov transition properties between adjacent motifs. The transitions may appear both between different actions or within the same action and are designed to overcome the lack of synchronization between motifs computed over different dimensions of a multi-dimensional signal.
3) **Observation features**
   $f(y_{i-1}, y_i, \mathbf{X}, i) = \delta(y_i = \text{motif}_k) g_i(\text{motif}_k)$. In this definition, $g_t(\text{motif}_k)$ represents the magnitude average of motif k. These features make the magnitude information for a motif available to the CRF; that information is lost in a typical symbolic motif representation. Observation features recover it by returning the mean magnitude of the motif.

IV. Experiments

Our experiments employ the publicly-available CMU Multi-Modal Activity Dataset (CMU-MMAC) [2]. In particular, we focus on the inertial measurement unit (IMU) portion of the dataset, which was collected by five MicroStrain 3DM-GX1 sensors attached to the subject's waist, wrists and ankles.

The dataset consists of unscripted recipes performed by several subjects in a kitchen. Thus, there is considerable variability in the manner in which the task is performed. The data corresponding to a given recipe consists of approximately 10,000 samples collected at 30 Hz over a period of about 6 minutes. Each frame includes 3-axis absolute orientation, angular velocity and instantaneous acceleration from each of the five sensors, leading to a 45-dimensional feature vector. Our experiments focus on the recipes that have been manually annotated into a series of actions (e.g., "open fridge" or "stir brownie mix"); these correspond to the "make brownies" task. We downsample the raw data by a factor of 10.

Figure 3 verifies that the proposed method (red line) achieves a steady decrease in training error as features are incrementally selected. We confirm that randomly selecting promising features (black line) is significantly worse than gain-based selection.

Table I summarizes the classification results for 14 actions. We compare the proposed CRF method against several baselines: CRF on raw features, hidden Markov model (HMM) with various parameters and k-Nearest Neighbor classifier (kNN). For the HMM, we first reduced the dimensionality using PCA to 8, 16, and 32 features and then learned an HMM with mixture of Gaussian outputs. We fixed the number of hidden states to 14 (corresponding to the number of classes) and explored several (1, 2, 3) Gaussians in the mixture. We also compared these HMMs to an HMM employing the complete 45-dimensional feature space. The best results were achieved using a mixture of two Gaussians and a 32-dimensional feature vector. Interestingly, this baseline is outperformed using a straightforward kNN classifier. We note that even though a

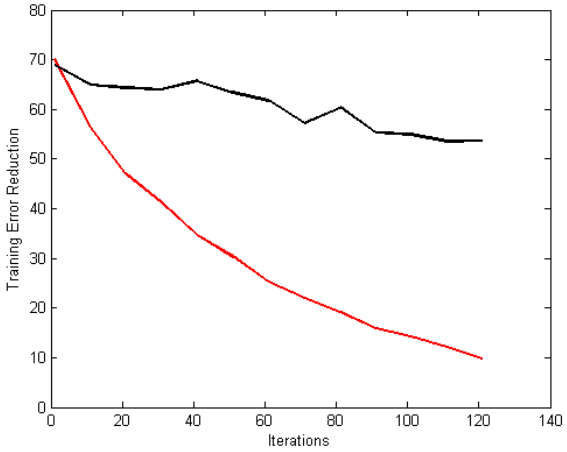

Fig. 3. Training error decreases with feature selection. Random feature selection (black line); Proposed method (red line).

TABLE I
COMPARISON OF CLASSIFICATION ACCURACY OF PROPOSED APPROACH AGAINST SEVERAL BASELINES.

Approach	Parameters	Accuracy
HMM	dim=8	8.22%
	dim=16	12.09%
	dim=32	25.60%
	dim=full (45)	16.74%
kNN	k=1	34.47%
	k=3	36.52%
CRF	raw features	30.02%
	proposed	**38.75%**

CRF trained on the raw features is not as good as the best baselines, feature selection using the proposed method enables us to clearly outperform all of baseline approaches.

V. CONCLUSION

This paper presents a method for generating and selecting features to improve the accuracy of CRF-based activity recognition. By linking our feature representation to the existence of 1-D motifs we can improve on classification performance over the raw IMU data. The CRF efficiently learns the cross-dimensional linkages between motifs, eliminating the need for multi-dimensional motif matching. We employ greedy feature selection in conjunction with a first-order approximation method based on reductions of the conditional log-likelihood error to achieve robust recognition while retaining computational feasibility.

ACKNOWLEDGMENTS

This research was supported in part by NSF award IIS-0845159 and by the NSF Quality of Life Technology Center under subcontract to Carnegie Mellon. We thank Tran Truyen for providing code and valuable advice.

REFERENCES

[1] B. Chiu, E. Keogh, and S. Lonardi. Probabilistic discovery of time series motifs. In *Proc. ACM KDD*, 2003.
[2] F. Frade, J. Hodgins, A. Bargtell, X. Artal, J. Macey, A. Castellis, and J. Beltran. Guide to the CMU Multimodal Activity Database. Technical Report RI-08-22, Carnegie Mellon, 2008.
[3] J. Lafferty, X. Zhu, and Y. Liu. Kernel conditional random fields: representation and clique selection. In *ICML*, 2004.
[4] A. Levin and T. Weiss. Learning to combine bottom-up and top-down segmentation. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2006.
[5] D. Liu and J. Nocedal. On the limited memory model for large scale optimization. *Mathematical Programming B*, 45(3), 1989.
[6] D. Minnen, T. Starner, I. Essa, and C. Isbell. Discovering characteristic actions from on-body sensor data. In *ISWC*, 2006.
[7] C. Sminchisescu, A. Kanaujia, Z. Li, and D. Metaxas. Conditional model for contextual human motion recognition. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2005.
[8] Y. Tanaka, K. Iwamoto, and K. Uehara. Discovery of time-series motif from multi-dimensional data based on mdl principle. *Machine Learning*, 58(2), 2005.
[9] A. Vahdatpour, N. Amini, and M. Sarrafzadeh. Toward unsupervised activity discovery using multi-dimensional motif detection in time series. In *Proceedings of International Joint Conference on Artificial Intelligence (IJCAI)*, 2009.
[10] D. Vail, J. Lafferty, and M. Veloso. Feature selection for activity recognition in multi-robot domains. In *Proceedings of AAAI*, 2008.