A Semi-Supervised Method for Segmenting Multi-Modal Data

Liyue Zhao
School of Electrical Engineering and Computer Science
University of Central Florida
Email: lyzhao@cs.ucf.edu

Gita Sukthankar
School of Electrical Engineering and Computer Science
University of Central Florida
Email: gitars@eecs.ucf.edu

***Abstract*—Human activity datasets collected under natural conditions are an important source of data. Since these contain multiple activities in unscripted sequence, temporal segmentation of multimodal datasets is an important precursor to recognition and analysis. Manual segmentation is prohibitively time consuming and unsupervised approaches for segmentation are unreliable since they fail to exploit the semantic context of the data. Gathering labels for supervised learning places a large workload on the human user since it is relatively easy to gather a mass of unlabeled data but expensive to annotate. This paper proposes an active learning approach for segmenting large motion capture datasets with both small training sets and working sets. Support Vector Machines (SVMs) are learned using an active learning paradigm; after the classifiers are initialized with a small set of labeled data, the users are iteratively queried for labels as needed. We propose a novel method for initializing the classifiers, based on unsupervised segmentation and clustering of the dataset. By identifying and training the SVM with points from pure clusters, we can improve upon a random sampling strategy for creating the query set. Our active learning approach improves upon the initial unsupervised segmentation used to initialize the classifier, while requiring substantially less data than a fully supervised method; the resulting segmentation is comparable to the latter while requiring significantly less effort from the user.**

I. Introduction

Multimodal datasets of human activity have become increasingly important in a range of applications including user interfaces, surveillance and eldercare. In particular, datasets of humans performing daily activities in natural settings are of particular value since they consist of data acquired under unconstrained environments. For analysis, researchers typically need to segment this data into segments that contain individual activities; for instance, data acquired during cooking could consist of activities such as "beating an egg" or "kneading dough". Manually segmenting such datasets is prohibitively time consuming for researchers, since even a short household task generates a large volume of data. The goal of our work is to present an interactive method for segmenting such datasets that makes the best use of a researcher's limited time. While our method is applicable to a variety of multimodal datasets, in this paper we focus primarily on motion capture data (see Figure 1).

Prior work in the graphics community typically assumes that motion capture data is acquired in short takes in which the subject performs only one or two motions at the direction of the animator. By contrast, we examine the problem of analyzing long sequences of motion capture data from a study of the activities of human daily living. The human subjects performed long tasks, containing many types of motions, without direction from an animator, while multimodal data (video, audio, accelerometer, and motion capture) was collected [1]. Even relatively short household tasks generate a large motion database; the standard capture rate of 120 frames per second results in 72,000 frames in just ten minutes of capture.

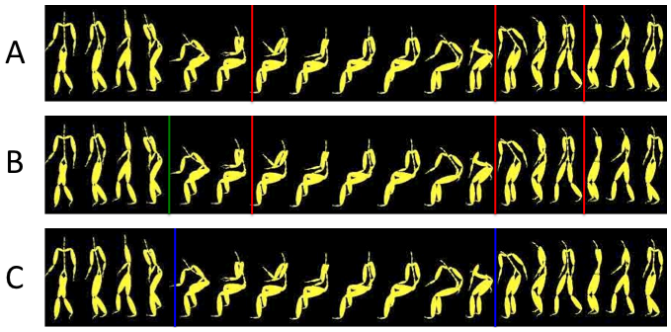

Fig. 1. (A) unsupervised segmentation; (B) new split (green) hypothesized using cluster refinement; (C) final segmentation generated by active learning.

The main problem with traditional active learning paradigms for training SVMs is sampling bias. Most approaches query samples close to the current estimated decision boundary since they assume that all data to be labeled are linearly separable. Unfortunately, this assumption is invalid in our datasets. Since the motion capture data are highly complex, even similar poses of a human may be labeled as different actions. Hence querying samples near by the decision boundary can ignore useful informative instances for the classification. Hoi et al. [2] report a similar issue in the domain of content-based image retrieval.

Another research question is how to initialize the classifiers without burdening the users by requesting a large initial set of labels. Our approach applies a clustering strategy to aggregate similar motion data after performing an unsupervised segmentation of the data. Sub-clusters are merged or divided based on whether they are mixed (contain multiple classes of labels) or pure (single class) clusters. Determination of cluster type is based on querying several samples in each cluster to identify whether it's a pure or mixed cluster. Although we can't

guarantee that the cluster is actually pure from the tiny subset of labels, we can definitely identify mixed clusters and remove these from our training set. Based on the initial segmentation and clustering, we automatically propagate the small set of user-provided labels across a larger training set to train the SVM classifiers.

However, the resulting hyperplane is not optimal since the labels based on clustering are themselves unreliable. To improve our segmentation, we employ an active learning SVMs approach [3] which asks the user to label those unlabeled samples that lie closest to the initial classification hyperplane. The resulting classifier finds the optimal decision boundary after querying a small number of samples.

II. Related Work

In this section, we describe two related approaches to the problem of motion capture segmentation: 1) unsupervised segmentation based on intrinsic data dimensionality, and 2) a supervised interactive support vector machine training paradigm.

Barbic et al. [4] introduced several approaches to motion capture segmentation based on the general concept that there is an underlying generative model of motion and that cuts should be introduced at points where the new data diverges from the previous model. In one of their proposed methods, principal component analysis (PCA) is used to create a lower-dimensional representation of the motion capture data at the beginning of a motion sequence. The main insight is that if the observed motion diverges from the data used to create the PCA basis, such as when the actor starts to perform a new action, then projecting the data of the new action using the old model will lead to large reconstruction errors. The moment that reconstruction errors increase quickly will occur at or near action boundaries.

However, in practice this approach leads to several problems. The method relies on building the PCA basis with frames from the current action, which requires about 300 frames or 2.5 secs of data. Unfortunately in our dataset, action changes can occur within that time frame, yielding a mixed basis capable of representing both actions without large reconstruction errors. Hence this technique cannot be used to accurately segment datasets with many short duration actions. Additionally, since PCA is a completely unsupervised approach, it is unable to distinguish between an activity that consists of multiple actions and boundaries between two semantically unrelated activities.

If user labels could be easily obtained, segmentation can be done in a completely supervised fashion using interactive SVMs to label the data [5]. Initially, users label a small training set of data. Then with kernel function Φ, the SVM classifier maps the training data into a high dimensional space which makes the data linearly separable. Since the partition hyperplane may not fit the unlabeled data, the user can add new labels to the training set and retrain the classifier. The method strives to balance classification accuracy and the user's labeling workload. However, their selection of new samples are based on the empirical judgment of the user and therefore susceptible to human error.

Our approach draws from both these methods, using an unsupervised PCA segmentation to initialize the clustering and a semi-supervised method to train the SVM classifiers. Unlike the interactive SVM segmentation proposed by [5], our approach utilizes the unlabeled data sets in the initial training. In the second phase, we automatically determine which instances from the unlabeled data are most useful to solicit labels from the user in the next iteration. Thus, the user is freed from selecting unlabeled samples and merely needs to label a small number of informative instances; this eliminates human bias and aims to reduce the amount of data that requires manual attention. In the next section, we provide details of our initialization and training method for semi-supervised support vector machines.

III. Methods

In the first stage, our SVM classifiers are initialized with a small set of training data. In the active learning stage, the classifiers are iteratively trained by having the users provide labels for a small set of automatically selected samples. Although the classifiers can be initialized by having the user provide labels for randomly sampled frames, we demonstrate that we can improve on that by selectively querying and propagating labels using a clustering approach.

A. Data Clustering

Several methods have been proposed to cluster data in geometric space [6], [7]. Since the motion segmentation problem is based on continuous time data sequences, it is possible to base the clustering on temporal discontinuities in the data stream. We use the PCA segmentation approach [4] outlined in the previous section to provide a coarse initial segmentation of the data.

Each raw motion capture frame can be expressed as a pose vector, $\mathbf{x} \in \Re^d$, where $d = 56$. This high-dimensional vector can be approximated by the low-dimensional feature vector, $\theta \in \Re^m$, using the linear projection:

$$\theta = \mathbf{W}^T(\mathbf{x} - \mu), \tag{1}$$

where $\mathbf{W}$ is the principal components basis and μ is the average pose vector, $\mu = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i$. The projection matrix, $\mathbf{W}$, is learned from a training set of $N = 300$ frames of motion capture data. $\mathbf{W}$ consists of the eigenvectors corresponding to the m largest eigenvalues of the training data covariance matrix, which are extracted using singular value decomposition (SVD). Transitions are detected using the discrete derivative of reconstruction error; if this error is more than 3 standard deviations from the average of all previous data points, a motion cut is introduced.

We found that this method provides a better starting point than traditional unsupervised clustering methods, such as k-means, which do not consider temporal information. Many of the clustering errors generated by the coarse segmentation are detected by pruning clusters based on a small set of labels solicited from the user. We ask the user to label the endpoints of the coarse segmentation and perform a consistency check on

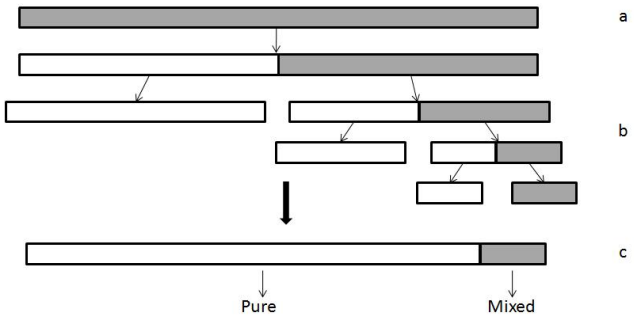

Fig. 2. (a) initial mixed data; (b) candidate clusters created by comparing the labels of the endpoints; (c) final clusters created by merging pure subsequences and discarding mixed subsequences (used to initialize SVMs).

the labels; if both endpoints have the same label, the segment is potentially pure; however if the labels of the endpoints disagree, we add a new cut in the middle of the segment and query the user for the label of that point. Clusters shorter than a certain duration (1% of total sequence length) are eliminated from consideration. The remaining clusters are used to initialize the support vector machine classifiers; labels from the end points are propagated across the cluster and the data is used to initialize the SVMs. The details of the segmentation method is illustrated in Figure 2. This process requires the user to label only 20–30 frames.

B. Active Learning

The clusters created by the coarse PCA segmentation, and refined with the user queries, are used to train a SVM classifier with both labeled and unlabeled samples. Semi-supervised support vector machines are regarded as a powerful approach to solve the classification problem with large data sets. Learning a semi-supervised SVM is a non-convex quadratic optimization problem; there is no optimization technique known to perform well on this topic [8]. However, our solution is a little different with the traditional methods based on linear or non-linear programming. Instead of searching for the global maximum solution directly, we use a simple optimization approach which may not identify the optimal margin hyperplane but will help the classifier decide which unlabeled samples should be added into the training set to improve the classification performance. We then query the user for the class labels of each of the selected samples and add them back to the training set. Suppose the labeled samples are denoted by $\mathbf{L} = \{x_1, x_2, ..., x_l\}$ and the unlabeled samples are $\mathbf{U} = \{x_{l+1}, x_{l+2}, ..., x_n\}$, the SVM classification problem can be represented as finding the optimal hyperplane with labeled samples that satisfies the equation:

$$\begin{array}{ll} \min_{\mathbf{w},b,\epsilon} & C\sum_{i=1}^{l}\epsilon_i + \|\mathbf{w}\|_2 \\ s.t. & y_i(\mathbf{w}\cdot\mathbf{x}_i + b) \geq 1 \quad i = 1, ..., l \end{array} \tag{2}$$

where ϵ_i is a slack term such that if $\mathbf{x_i}$ is misclassified and C is the constant of the penalty of the misclassified samples. All possible hyperplanes that could separate the training data as $f(\mathbf{x_i}) > 0$ for $y_i = 1$ and $f(\mathbf{x_i}) < 0$ for $y_i = -1$ are consistent with the version space $\mathcal{V}$. Tong and Koller [3] have shown that the best way to split the current version space into two equal parts is to find the unlabeled sample whose distance in the mapping space is close to the current hyperplane $\mathbf{w}_i$. The description of our method is detailed in Algorithm 1.

Input: The complete data set with labeled set $\mathbf{T}$ and unlabeled set $\mathbf{U}$
Output: The optimal SVMs hyperplane to separate the available data into two groups
Initialization: Calculate the initial hyperplane by using SVMs on the clustering data set $\mathbf{T}$;
while *the variation classification hyperplane is not stable* **do**
  Calculate the distance $\mathbf{d}$ between unlabeled set $\mathbf{U}$ and the current SVMs hyperplane $\mathbf{w_i}$;
  Query the unlabeled sample x_{l+i} with the smallest distance d_i;
  Manually label the sample x_{l+i};
  Update the labeled set as $\mathbf{T} = \mathbf{T} \cup \{x_{l+i}\}$ and unlabeled set as $\mathbf{U} = \mathbf{U} \setminus \{x_{l+i}\}$;
  Re-train the SVM classifiers $\mathbf{w_{i+1}}$ with labeled set $\mathbf{T}$;
end

Algorithm 1: Proposed active learning algorithm.

The traditional initialization method arbitrarily selects samples to include in the training sets. However, randomly choosing samples may lead to sampling bias which make the SVM classifier unable to achieve the global maximum. In our approach, the labels of samples in each viable cluster are set as the majority labels of querying samples. This converts learning a semi-supervised SVM into a classical SVM optimization problem. However, the clustering strategy does not guarantee that the decision boundary is optimal since the clustering step is not reliable. It merely gives a good initial hyperplane; active learning is still required to perfect the solution.

In our experiments, the SVM classifier was implemented with the SVM-KM toolbox using a polynomial kernel [9]; multi-class classification is handled using a one vs. all voting scheme. Instead of using a *hard margin* for the SVM, our method relies on a *soft margin* restriction in classification. A hard margin forces both labeled and unlabeled data out of the margin area, whereas the soft margin allows unlabeled samples to lie on the margin with penalties. With limited training samples, we find that the hard margin restriction is so restrictive that it may force the separating hyperplane to a local maximum.

IV. Results

For these experiments we used the publicly available Carnegie Mellon Motion Capture dataset (http://mocap.cs.cmu.edu/) collected with a Vicon motion capture system. Subjects wore a marker set with 43 14mm markers that is an adaptation of a standard biomechanical marker set with additional markers to facilitate distinguishing the left side of the body from the

Method	Initialization	Active Learning Query
A (proposed)	cluster refinement	margin-based
B	random	margin-based
C	random	random

TABLE I
SUMMARY OF METHODS EVALUATED IN FIGURE 3.

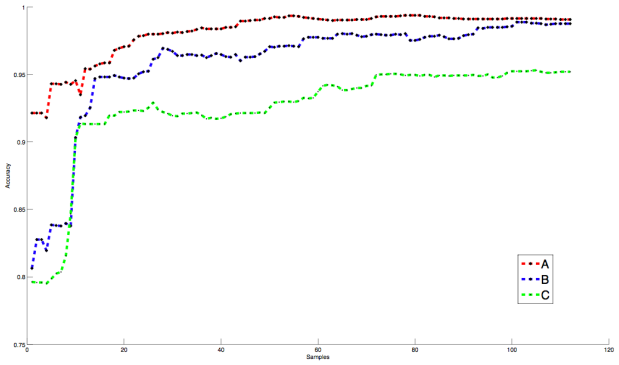

Fig. 3. Improvement in quality of segmentation as additional labels are acquired using active learning. The proposed method (A) benefits through intelligent initialization and margin-based selection of active learning queries.

right side in an automatic fashion. The dataset contains a bunch of sequences with different human actions; to evaluate our method we selected 15 sequences that include actions such as running, swinging, jumping, and sitting. Table I summarizes the characteristics of the three methods evaluated in our experiments. The first baseline (C) is trained using data that is sampled at random (with uniform distribution) from the activity sequence. The second (B) is initialized using a random segmentation but employs our proposed margin-based approach for generating instances for the user to label. The third (A) is our proposed approach and employs an unsupervised clustering to initialize the segmentation followed by margin-based sampling for identifying informative active learning query instances.

We evaluate the quality of segmentation using classification accuracy. Figure 3 shows how this accuracy improves with additional training data for each of the methods. Clearly, adding training data in a haphazard manner (C) leads to an inefficient form of active learning. The second method (B) demonstrates the benefits of our margin-based method for selecting queries for active learning. Finally, the accuracy curve for the proposed method (A) shows the boost that we obtain through intelligent initialization using unsupervised clustering. In comparison to a fully supervised SVM trained with 100 samples, our method achieves the same 95% accuracy with only half the data (40 samples).

These quantitative results are consistent with qualitative observations. Figure 4 shows a sample from our dataset where each segment is individually shaded. We compare the results from our proposed method (denoted "active learning") against those from a baseline unsupervised method (denoted "PCA"). Clearly, the segmentation generated by the proposed method

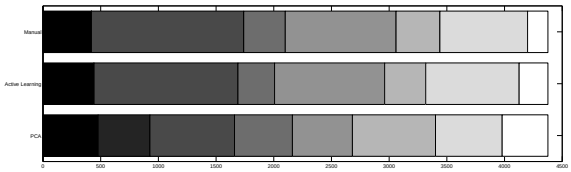

Fig. 4. Comparison of segmentation results. We observe that the segmentation generated by the proposed method (active learning, middle) is much closer to the ground truth (manual, top) than that generated by a standard unsupervised approach (PCA, bottom).

is much closer to the ground truth.

V. CONCLUSION

In this paper, we introduce a new approach for segmenting large motion capture databases by combining unsupervised clustering with active learning. We demonstrate that our segmentation technique is comparable to manual segmentation while requiring only a fraction of the labels needed by a fully-supervised method. In future work, we plan to analyze the effects of segmentation errors on higher-level analysis of human activity streams. Currently we can utilize this data by assuming that the all sensor data are time locked, enabling us to propagate the segmentation from the motion capture data to the video and accelerometer data streams. However due to data collection glitches, this is not always the case and propagating cuts across modalities results in segmentation errors. By directly applying our segmentation method to the other modalities, we can compensate for the lack of time locking.

ACKNOWLEDGMENT

This research was supported by the NSF Quality of Life Technology Center under subcontract to Carnegie Mellon.

REFERENCES

[1] F. Frade, J. Hodgins, A. Bargtell, X. Artal, J. Macey, A. Castellis, and J. Beltran, "Guide to the CMU Multimodal Activity Database," Carnegie Mellon, Tech. Rep. CMU-RI-TR-08-22, 2008.

[2] S. Hoi, R. Jin, Z. Jianke, and M. Lyu, "Semi-supervised SVM batch mode active learning for image retrieval," in *Proc. Computer Vision and Pattern Recognition*, 2008.

[3] S. Tong and D. Koller, "Support vector machine active learning with applications to text classification," *Journal of Machine Learning Research*, vol. 2, pp. 45–66, 2002.

[4] J. Barbic, A. Safonova, J.-Y. Pan, C. Faloutsos, J. Hodgins, and N. Pollard, "Segmenting motion capture data into distinct behaviors," in *Proceedings of Graphics Interface*, 2004.

[5] O. Arikan, D. Forsyth, and J. O'Brien, "Motion synthesis from annotations," *ACM Trans. Graphics*, vol. 22, no. 3, 2003.

[6] V. Sindhwani, P. Niyogi, and M. Belkin, "Beyond the point cloud: frame transductive to semi-supervised learning," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2005.

[7] F. Zhou, F. Frade, and J. Hodgins, "Aligned cluster analysis for temporal segmentation of human motion," in *Proceedings of the IEEE Conference on Automatic Face and Gesture Recognition*, 2008.

[8] O. Chapelle, V. Sindhwani, and S. Keerthi, "Optimization techniques for semi-supervised support vector machines," *Journal of Machine Learning Research*, vol. 9, pp. 203–233, 2008.

[9] S. Canu, Y. Grandvalet, V. Guigue, and A. Rakotomamonjy, "SVM and Kernel Methods Matlab Toolbox," Perception Systmes et Information, INSA de Rouen, Rouen, France, 2005.