Temporal Segmentation and Activity Classification from First-person Sensing

Ekaterina H. Spriggs, Fernando De La Torre, Martial Hebert
Carnegie Mellon University.
{espriggs, ftorre, hebert}@cs.cmu.edu

Abstract

Temporal segmentation of human motion into actions is central to the understanding and building of computational models of human motion and activity recognition. Several issues contribute to the challenge of temporal segmentation and classification of human motion. These include the large variability in the temporal scale and periodicity of human actions, the complexity of representing articulated motion, and the exponential nature of all possible movement combinations. We provide initial results from investigating two distinct problems - classification of the overall task being performed, and the more difficult problem of classifying individual frames over time into specific actions. We explore first-person sensing through a wearable camera and Inertial Measurement Units (IMUs) for temporally segmenting human motion into actions and performing activity classification in the context of cooking and recipe preparation in a natural environment. We present baseline results for supervised and unsupervised temporal segmentation, and recipe recognition in the CMU-Multimodal activity database (CMU-MMAC).

1. Introduction

Temporal segmentation of human motion into actions is central to the understanding and building computational models of human motion and activity recognition. Research that addresses the problem of detection, recognition and synthesis of human human motion have gained substantial interests from both academia and industry over the last few years due to the large number of applications[1], [20], [13], [15], [22]. Unsupervised techniques for learning motion primitives from data have recently drawn the interest of many scientists in computer vision [9], [28], [27], [17] and computer graphics [4], [16], [3], [8]. Although previous research has shown promising results, recognizing human activities and factorizing human motion into primitives and actions (*i.e.* temporal segmentation) is still an unsolved problem in human motion analysis. The inherent difficulty of human motion segmentation stems from the large intra-person physical variability, wide range of temporal scales, irregularity in the periodicity of human actions, and the exponential nature of possible movement combinations. In this work we explore the use of Inertial Measurement Units (IMUs) and a first-person camera for overall task classification, action segmentation and action classification in the context of cooking and preparing recipes in an unstructured environment. As a first step to exploring this space, we investigate the feasibility of standard supervised and unsupervised Gaussian Mixture Models (GMMs), Hidden Markov Models (HMMs), and K-Nearest Neighbor (K-NN) techniques for action segmentation and classification on these two modalities. Furthermore, to alleviate the need of manual annotation, we also investigate the use of unsupervised techniques and compare performance with the supervised methods.

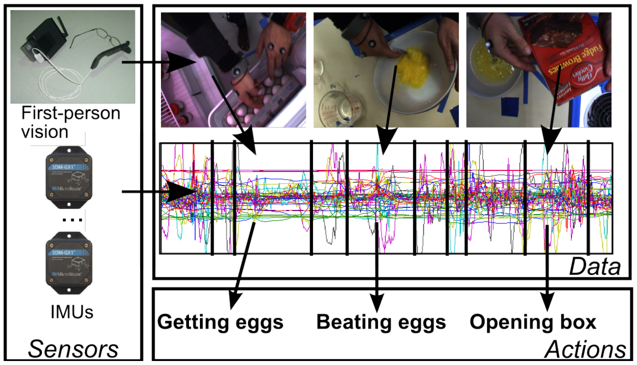

Figure 1. Action segmentation and classification from first-person sensors from the CMU-MMAC dataset.

This paper provides baseline results for recipe classification, action segmentation and action classification on the Carnegie Mellon University Multimodal Activity (CMU-MMAC) database [6]. The database contains data from human behavior in a natural kitchen environment, including sensor modalities that capture the user's perspective. Figure 1 illustrates the problem of temporal segmentation: given a stream of IMU data and first-person vision, we want to find a temporal decomposition and classification of the recipe the user is cooking.

The remainder of the paper is organized as follows. Sec-

978-1-4244-3993-5/09/$25.00 ©2009 IEEE

tion 2 discusses prior work, section 3 introduces the action database we used, and some of the challenges it presents are examined in section 4. Sections 5 and 6 show baseline experiments using the IMU sensors and first-person video from the CMU-MMAC database. Finally, Section 7 concludes the paper and outlines future work.

2. Previous work

In the area of wearable and ubiquitous computing Schiele *et al*. [24] proposed an interactive computer vision and augmented reality system that autonomously provides media memories based on objects in the view. Object recognition is performed using multidimensional histograms of Gaussian derivatives from images collected by a wearable camera. Mayol and Murray [19] recognize hand activity by detecting objects subject to manipulation using a wearable camera on the shoulder. Data was collected from one subject and five events are recognized via their associated objects. IMUs for action recognition have been explored by several groups. For example, Lester *et al*. [14] use discriminative classifiers and HMMs to recognize a small set of ten actions (*e.g*. running, walking, *etc*.) from a multimodal data set (*e.g*. accelerometer, audio, light sensor on a sensor board) in an unconstrained environment.

There exists an extensive graphics and computer vision literature that addresses the problem of grouping human actions. In the computer graphics literature, Barbic *et al*. [3] proposed an algorithm to decompose human motion into distinct actions by detecting sudden changes in the intrinsic dimensionality of the Principal Component Analysis (PCA) model. Jenkins *et al*. [11], [8] used the zero-velocity crossing points of the angular velocity to segment the stream of motion capture data. Jenkins and Mataric [12] further extended the work by finding a non-linear embedding, using Isomap [26], that reveals the temporal structure of segmented motion. Recently, Beaudoin *et al*. [4] developed a string-based motif-finding algorithm to decompose actions into action primitives and interpret actions as a composition on the alphabet of these action primitives. The algorithm allows for a user-controlled compromise between motif length and the number of motions in a motif. Rui and Anandan [23] used principal components of frame-to-frame optical-flow to discover temporal trajectories of human motion in video. Recently, Guerra-Filho and Aloimonos [9], [10] presented a linguistic framework for modeling and learning of human activity representations. The low level representation of their framework, motion primitives, referred to as *kinetemes*, are studied as the foundation for a kinetic language.

In work using cameras observing the subjects, Schuldt *et al*. [25] presented a method using local space-time features to capture six types of human actions (walking, jogging, running, boxing, hand waving and hand clapping) from video. Blank *et al*. [5] perform action recognition, detection and clustering on several outdoor actions on known background using local space-time saliency, action dynamics, shape structure and orientation. Efros *et al*. [7] recognize actions at a distance on a ballet, tennis, and a soccer dataset. They introduce a novel motion descriptor based on optical flow measurements in a spatio-temporal volume. In contrast to using cameras observing the subject from a distance, we investigate the first-person vision modality from the CMU-MMAC database.

From the action recognition literature using other sensors, Bao and Intille [2] investigated performance of recognition algorithms with multiple, wire-free accelerometers on 20 activities (*e.g*. running, walking, reading) using data sets annotated by the subjects themselves. Wu *et al*. [29] presented a DBN model which incorporates common-sense activity descriptions, RFID sensor events, and video data from a static camera to perform recognition of 16 daily kitchen activities. In contrast to the type of activities explored in this work, the CMU-MMAC database contains data from preparing complete recipes as experienced from the user's perspective.

This project differs from previous work in that it explores action classification in a non-instrumented environment, using modalities collected from the user's perspective and targeted at the class of actions observed in performing everyday cooking.

3. Dataset

The Carnegie Mellon University Multimodal Activity database (CMU-MMAC)[6] database contains multimodal measures of the human activity of subjects performing the tasks involved in cooking and food preparation. A kitchen was built and to date forty subjects have been recorded cooking five different recipes: brownies, pizza, sandwich, salad and scrambled eggs. The following modalities were recorded: • Video: (1) Three high spatial resolution (1024 x 768) color video cameras at 30 Hertz. (2) One low spatial resolution (640 x 480) color video cameras at 60 Hertz. (3) One low spatial resolution (640 x 480) color video cameras at 30 Hertz. (4) One wearable medium spatial resolution (800 x 600) camera at 30 Hertz.

• Audio: Five balanced directive microphones at 44100 Hertz and 16 bit/sample.

• Motion capture: A Vicon motion capture system with 12 infrared MX-40 cameras. Each camera records images of 4 megapixel resolution at 120 Hertz.

• IMU: (1) 5 3DM-GX1 IMUs, each with a triaxial accelerometer, gyro and magnetometer sensor sampling at 125 Hz. (2) 4 6DOFv4 Sparkfun Bluetooth IMUs, each with a triaxial accelerometer, gyro and magnetometer sensor sampling at 62 Hz.

• Wearable: (1) Wearable e-watch - triaxial accelerometer and light intensity sensors [18]. (2) Bodymedia Sensewear Pro 2 (Bodymedia, Pittsburgh, PA), measuring Heat Flux, Galvanic Skin Response, Skin Temperature and Near-Body Temperature. (3) RFID reader i-Bracelet at 1 Hz [29].

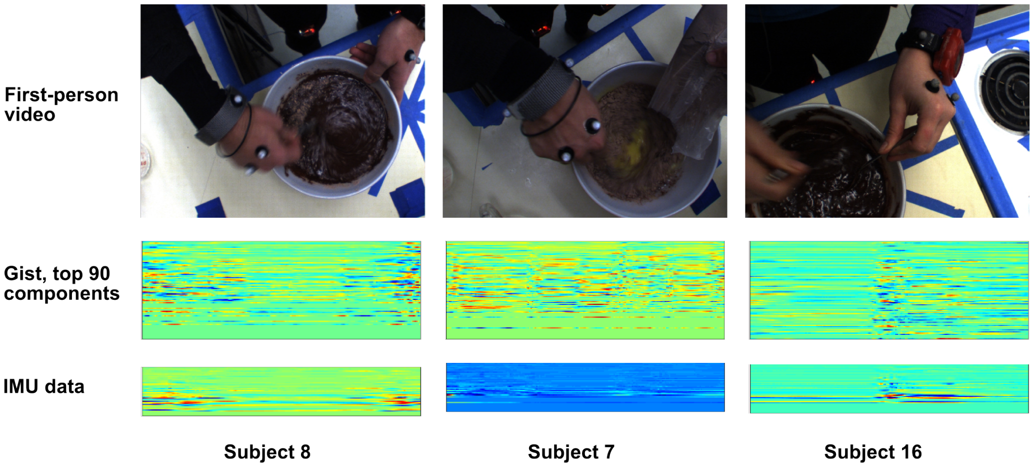

Figure 2. Examples of various ways subjects stirred the brownie mix - first pouring in the brownie mix, then stirring; stirring while pouring in the brownie mix; stirring while holding two utensils. Top row is first-person vision, middle row shows the top 90 components of the gist for 100 frames before the snapshot, and 100 frames after the snapshot, bottom row is IMU data for same time interval (approximately 7 seconds total length).

The various modalities were recorded using several computers, which were synchronized using the Network Time Protocol (NTP). The dataset can be downloaded from *http://kitchen.cs.cmu.edu/*.

For these initial results we consider two recipes performed by seven subjects - making brownies from a dry mix box, and making an omelet. We explore two sensor modalities - five IMUs located on each of the subject's wrists, ankles, and one on the waist, and the first-person vision camera. The average number of frames for each person for the brownies recipe is 11784, for the omelet recipe is 6875, and the data from the seven subjects consists of a total of 82489 frames for the brownies recipe, and total of 48131 for the omelet recipe, all at 30Hz sampling frequency.

This dataset differs from other activity recognition databases as it contains a multitude of cooking activities from a larger number of people. The subjects were asked to perform the recipes in a natural way, and no instructions were given as to how to perform each task. The actions vary greatly in time span, repetitiveness, and manner of execution. In addition to the variety of actions, this dataset contains key modalities that directly relate to the person's perspective - wearable IMUs and first-person vision.

4. Challenges

4.1. Data annotation

As an initial step to exploring the dataset, we first consider possible levels of annotating actions. After initial evaluation of the data, we have found that data labeling of everyday activities is ambiguous due to the various ways a task can be performed and described.

For instance, we can label at the recipe level (*e.g.* ”beat two eggs in a bowl”), at a more detailed action level (*e.g.* ”break an egg”), or at a very fine-grained level of simple movements (*e.g.*, ”reach forward with left hand”). As a first step to evaluating performance of action recognition on this dataset, we label 29 actions for seven subjects making brownies, as shown in Table 4.2.

Not all actions were performed by all subjects, and some frames belong to unlabeled actions (*e.g.*, frames in between two distinct actions are difficult to classify at the chosen level of annotation).

4.2. Variability in action execution

One of the big challenges in this dataset is the great variety of performing each of the daily kitchen actions observed, as no instructions of how to perform the recipe were given to the subjects. For example, one of the subjects pours the brownie mix in the bowl of beaten eggs and then stirs the ingredients, while another stirs while pouring in the brownie mix, and yet a third person stirs while holding a second utensil in the mix (see Figure 2). This diversity presents ambiguity in describing the action as either ”pouring in mix” or ”stirring mix,” or as a separate action ”pouring in mix while stirring.”

4.3. Object recognition and scene detection

Many of the objects in the dataset lack texture as the distinctive parts may not be visible from the typical viewpoint of the user, making it difficult to use object recognition and object tracking algorithms based on texture features (all objects used in this dataset were taken from the usual everyday kitchen inventory).

In addition, cooking involves transforming ingredients from one shape and color to another, *e.g.* breaking eggs and beating them, pouring in brownie mix, etc, rendering object tracking very difficult. The cooking ingredients constitute a significant number of the objects in view in first-person vision, and are thus an important part of activity un-

Open cupboard (bowls)	Get fork
Open cupboard (brownie)	Walk to fridge
Open fridge	Get eggs
Close fridge	Walk to counter
Break one egg	Beating egg(s)
Pour in water in bowl	Get oil from cupboard
Pour oil in cup	Put oil away
Open brownie box	Pour in brownie mix
Pour oil in bowl	Stir brownie mix
Get baking pan	Spray with Pam
Put Pam away	Set stove settings
Pour mix in baking pan	Put pan in oven
Pour tap water in cup	Put cap on
Get Pam from cupboard	Remove cap
Read recipe	

Table 1. List of 29 manually selected action classes for annotation of the brownies recipe.

derstanding. An alternative method to object recognition that can provide information about the objects in use is an RFID bracelet that reads tags on objects, as was successfully used in [29]. While we have tagged several objects in the kitchen, some key cooking ingredients used in this dataset (*e.g.*, eggs, forks) do not lend themselves to an easy and effective tagging.

As a start, we explore action classification without object use information, concentrating on the type of scene as observed through the first-person camera. Specifically, we investigate if the global scene information from the first-person video has discriminative power for recipe identification and action classification. The first-person video exhibits substantial amount of rapidly changing pixel values as the subject performs the recipe steps. However, we note that most actions are performed while the background remains somewhat constant. For example, breaking eggs, beating eggs, pouring ingredients in a bowl, etc, are always performed while looking at the bowl on the counter, and not while looking in the fridge, which is associated with the action of fetching the eggs.

5. Unsupervised segmentation

As a first step to exploring features from first-person vision and IMU sensors in the context of daily activities, we investigate data segmentation through unsupervised techniques. We perform two tasks - recipe classification and unsupervised temporal segmentation on three data modalities: vision only, IMU sensors only, and combined vision and IMU sensor data. To evaluate the unsupervised results, each estimated cluster is displayed against the manually segmented data. A decision is made whether each cluster contains coherent chunks of frames as defined by the manual labels. We report the total number of frames in the learned clusters that correspond to the chosen action cluster.

5.1. Task classification from first-person vision

One of the key benefits of first-person vision is that it relates to the user's intentions. We expect that what the user sees should be correlated with the action they are performing. As an initial analysis of the first-person vision modality we explore scene type as one possible cue to determine what stage in the recipe each frame belongs to.

We investigate if global features capture the recipe type by modeling the sequence of scene transitions in time. Noting that many actions are performed while looking at a somewhat constant background, we consider the gist [21] of each frame as a possible way of describing the scene the person is looking at. In this context, the gist is used to discriminate between indoor locations (*e.g.* the counter, the fridge, the stove top, etc) as observed through the first-person camera, which is on average 2-3ft away from objects and surfaces.

We compute the gist of each frame at 4 scales and 8 orientations, discretized into 4x4 blocks, producing a 512 dimensional feature vector per video frame. We perform standard dimensionality reduction by concatenating the feature vectors of the seven subjects, performing PCA analysis and retaining smaller size feature vectors (32 or less). The data is then normalized to zero mean and variance of one. The data for the brownie and omelet recipe is reduced separately.

We investigate whether the extracted video features cluster into similar scenes by estimating a Gaussian Mixture Model in an unsupervised manner. Considering brownie and omelet recipes separately, we use the gist features from the seven subjects after dimensionality reduction.

For unsupervised scene segmentation, we estimate a GMM for several combinations of parameters. We explore various size feature vectors from the computed PCA components (3, 8, 16, 32) and number of clusters (20, 30, 40), with a threshold of 30 iterations, using 2 replicates and diagonal covariance. The GMM model is learned from the data from all seven people one recipe at a time, and each frame is assigned to a cluster.

In Figure 3 we visualize seven of the estimated clusters, noting that the majority of frames in this set roughly correspond to the manually labeled ”stirring” action. From a total of 20401 frames manually labeled as ”stirring,” 14432 were assigned to the set of these seven estimated clusters (70%), 5969 frames were not assigned to this set (29%). From the 5228 frames in the estimated clusters that belong to different actions (26%), most of them belong to actions involving ”pouring.” When changing the model parameters we observe that some manually labeled actions (*e.g.*”walk to fridge,” ”walk to counter,” ”take eggs,”) are more coherently clustered by the model. This suggests the need for individual classifiers for each action.

For the recipe classification, we describe the sequence of scenes for the two recipes by estimating an HMM with mixture of Gaussians outputs from the reduced gist features

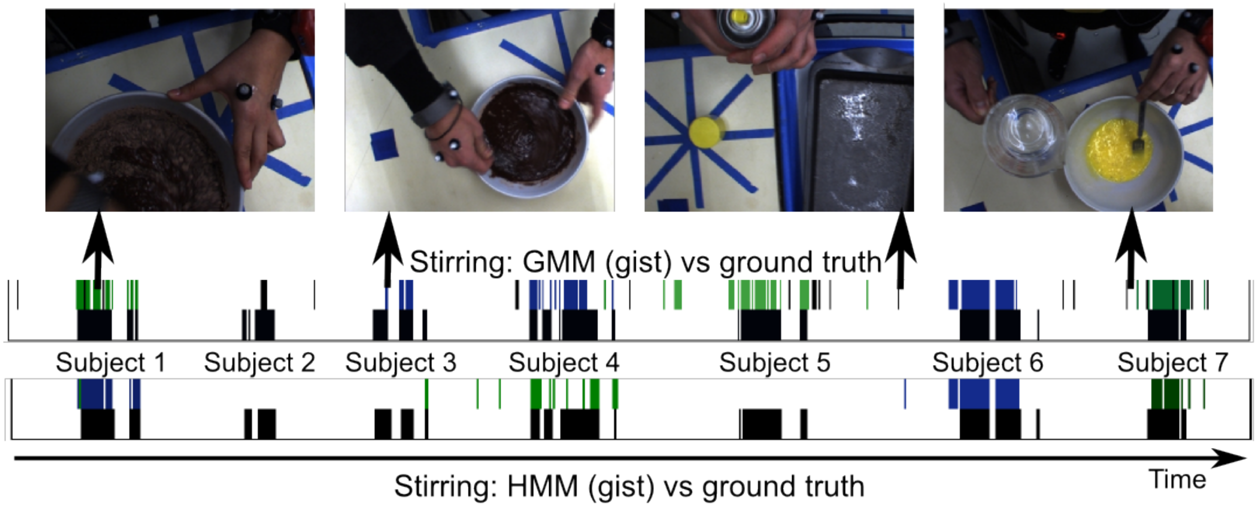

Figure 3. Unsupervised segmentation of the brownies recipe sequence using only the gist. The GMM segmentation uses 32 PCA components and 30 clusters, the HMM uses 16 PCA components and 29 states. Several of the unsupervised clusters capture the action of ”stirring,” as shown in the first two snapshots. The black clusters are the ground truth for this action. Two frames which belong to other actions are also shown. Note that the estimated clusters group actions per person and don’t generalize across people.

in an unsupervised manner. The combinations of parameters considered are: 1, 2, and 3 mixture of Gaussians output, 20, 30, and 40 hidden states, using feature vectors of length 3, 8, and 16, spherical covariance, and a maximum of 10 iterations for convergence.

Evaluation of recipe classification performance was performed by learning an HMM for the brownie recipe from all but one subjects, learning an HMM for the omelet recipe from all subjects, computing likelihood of the withheld sequence under both models, and classifying it as the more likely type. This is repeated in a cross-validation manner, withholding all people in turn from both recipes. Best average classification performance of 92.8% (13 out of 14 tests correctly predicted) was reached with an HMM with 32-dimensional feature vectors, 40 hidden states, and 3 mixture of Gaussians outputs. We note that the higher dimensional vector tests perform better, with the number of states and mixtures having less effect.

In addition to recipe classification, the data for all people can also be segmented using the classes produced by computing the Viterbi path from the estimated HMM from each recipe separately. Figure 3 shows the frames from four clusters that best match the frames labeled as ”stirring,” along with the GMM segmentation and the manually labeled data.

Compared to the GMM performance, the HMM clustering fails to cluster the manually labeled frames into coherent chunks. For the HMM model using 16 PCA components and 29 states, a total of 9865 (48%) of the 20401 frames labeled as ”stirring” were clustered together in the chosen set of clusters. The total number of frames in these HMM clusters is 12287, where 2422 frames (20%) belonging to actions with different manual labels. Experimenting with the model parameters, we note that a small number of PCA components produces clusters spread randomly on the timeline (compared to the manually labeled data), and a larger number of components produces coherent segmentations of a few actions. However, these segmentations do not generalize across people - one cluster models an action from one subject, while another cluster models the same action from another subject.

5.2. Action segmentation from IMU sensors

Inspired by prior work with accelerometer sensors for classification of various actions [2], we explore unsupervised techniques for recipe classification and data segmentation using the IMU sensors from this dataset. Previous work has successfully performed classification of repetitive actions like walking, running, washing windows, plates, etc, from accelerometer data [2] by computing features using a sliding window. However, some of the actions in this dataset span a very short amount of time, while others are performed over a longer period. It is not clear at this point how to extract IMU features using a sliding window framework.

For this initial analysis we smooth the sensor data by taking the mean of every four frames, while sub-sampling the 125Hz signal to 30Hz. The resulting data points are 45-dimensional feature vectors - each of the five accelerometer, gyro and magnetometer sensors report values for 3 axes. We perform PCA on the concatenated IMU data from the seven subjects separately per recipe and retain a smaller size feature vectors (32 or fewer). The data is then normalized to have zero mean and one variance. We estimate an HMM with a mixture of Gaussians output for the two recipes separately from the seven subjects using the IMU features after the dimensionality reduction. We tested 3, 8, 16, and 32-dimensional feature vectors, 10, 20, and 30 hidden states,

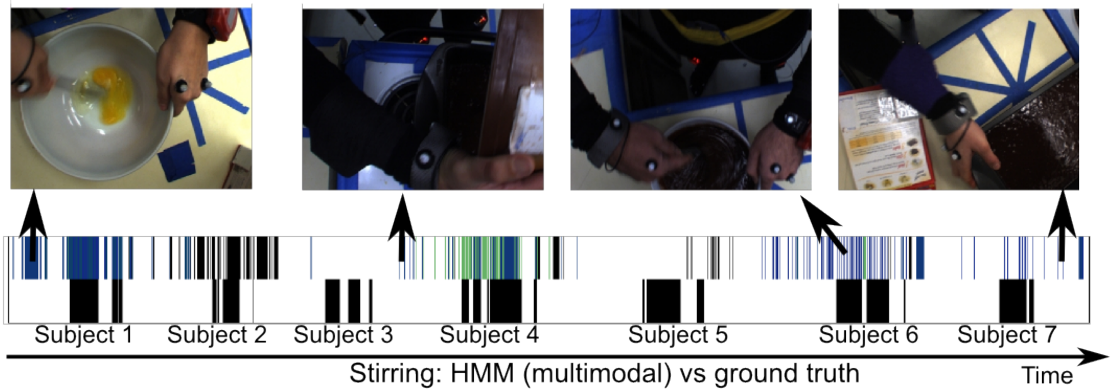

Figure 4. Unsupervised segmentation using HMM on the multimodal data. Displayed are a set of seven estimated clusters which roughly correspond to the action of "stirring." The HMM uses 16 dimensional features and 30 states. 50% of the manually classified frames fall in these estimated clusters.

and 1,2, and 3 mixtures of Gaussian outputs.

For recipe classification we perform the same cross-validation experiment as for the first-person vision: we train models for both recipes by withholding one person, and then classify the withheld sequence using the most likely model. Recipe classification performance is 100% from an HMM with 3-dimensional feature vectors, 20 hidden states, and 3 mixture of Gaussians outputs. We note that higher dimensional vectors performed worse (85% performance from an HMM with 8-dimensional vector, 20 hidden states and 3 mixtures).

We also explore unsupervised segmentation of the IMU data using an HMM. However, no coherent clusters are produced for the parameters used - the frames from the resulting clusters are widely spread along the timeline.

5.3. Action segmentation from multi-modal data

The goal is to explore the combination of first-person vision and IMU sensors for recipe classification and action segmentation using unsupervised algorithms. The first-person video and the IMU features are integrated by concatenation after normalizing the features by their norm, and then computing PCA for all seven people together, separately per recipe.

For the recipe classification task we estimate HMM models for both recipes by withholding one subject at a time and classifying the test sequence according to the most likely model. We explored models using 3, 8, 16, and 32-dimensional features, 20, 30, and 40 hidden states, and 1, 2, and 3 mixture of Gaussians output. From the HMM parameter options we tested, the best recipe classification performance was 92.8% using an 8-dimensional feature vector, 30 hidden states, and 3 Gaussian mixtures. We note that the number of hidden states and mixtures did not affect the outcome as much as the dimension of the feature vectors used.

Figure 4 shows a comparison between one of the estimated unsupervised segmentations and the available manual action annotations.

While unsupervised segmentation cannot clearly convey the action being performed, the segmentations resulting by changing the model parameters show promise in discovering some of the manually labeled actions. Different actions are clustered better for different model parameters, suggesting the need for multiple levels of segmentation.

6. Supervised action classification

To evaluate action classification on this dataset we consider standard supervised algorithms that use the 29 manually annotated actions for the brownies recipe (see Table 4.2), with chance at roughly 3%. From the combined 82496 data points from the seven subjects, 81.4% (67191) of the frames are annotated. Note that the action "stirring the brownie mix" comprises approximately 25% of these frames. To handle the lack of fully annotated data, we remove the unlabeled frames from the dataset and train only on the labeled frames. Two models were considered - a supervised HMM and a K-Nearest Neighbor model.

6.1. Action classification using supervised HMM

We train an HMM on the three data modalities after dimensionality reduction by providing the class labels in the training stage of the model. We explored 5, 8, 16 and 32-dimensional feature vectors, and 1, 2, and 3 mixtures of Gaussian outputs, with 29 states. We estimate an HMM from six people, and test performance on the withheld person, repeating this for all seven people. We classify each test frame as belonging to one of the 29 classes from the manual annotation.

Using gist features, the best average frame classification achieved over all seven people was 9.38% (with chance at 3%), using an HMM with 16-dimensional feature vector, 29 states, and two mixtures of Gaussians outputs. Using IMU data alone, average classification performance was 10.4% from an HMM with 32-dimensional feature vectors, 29 states, and 2 mixtures of Gaussians outputs. Combining both modalities, we reach 12.34% average frame classification performance, using an HMM with 16-dimensional feature vector, 29 states, and 2 mixtures of Gaussians outputs.

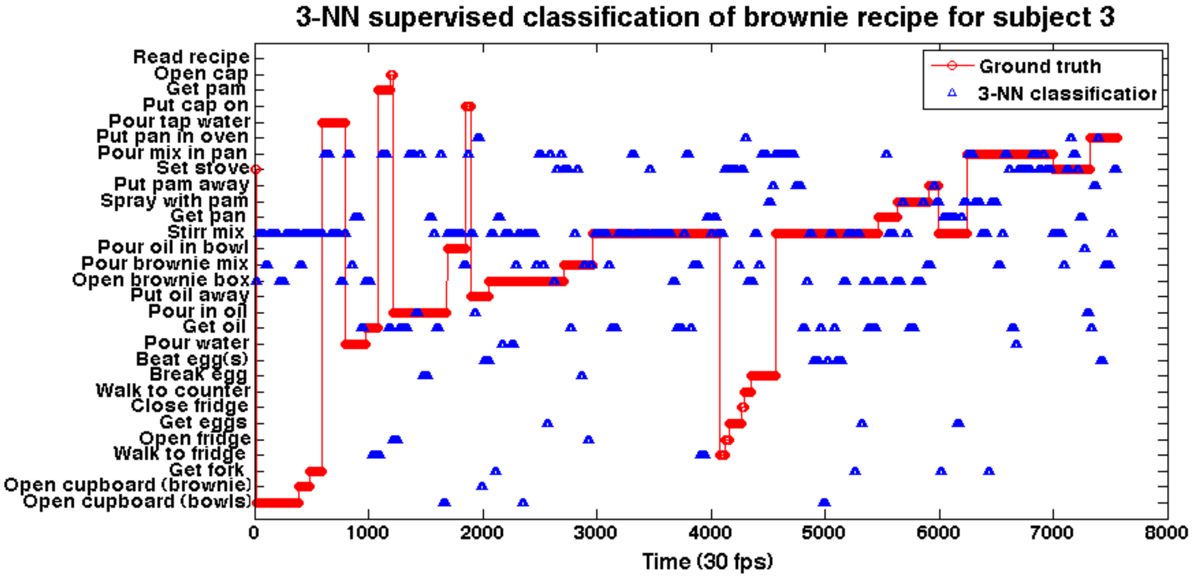

Figure 5. Classification performance from supervised 3-NN, merging 30 feature vectors without overlap, using the full 557 dimensional multimodal data. Plot shows the classification performance for subject 3: 61% of the frames were correctly classified. The ”stirring” action has more training frames than the other actions, and it is classified more often by NN.

6.2. Action classification using K-Nearest Neighbor

In the spirit of [7], we also explored matching test frames in the framework of nearest-neighbor. For the brownies recipe, we classify each frame from a withheld person’s sequence based on the grouping of the frames in the remaining six sequences. We explored 1-NN and 3-NN for classification, with a Euclidean distance and majority rule with nearest point tie-break options.

Using first-person data only, the average frame classification performance over the seven tests for 1-NN was 48.64% (chance is at 3%), when using the entire 512-dimensional feature vector. When using IMU data alone, performance was 56.8% using the full 45-dimensional feature vector. Best performance of 57.8% was achieved when using both modalities with the full 557-dimensional feature vector. Figure 5 shows the results from the multi-modal frame classification for one subject. We varied the number of neighbors used for classification and we also constructed new feature vectors by concatenating v consecutive vectors together without overlap, with $v = [2, 5, 10]$. We observe that NN models capture the ”stirring” action reasonably well, and by changing the parameters we get more coherent clusters for a few other actions. Similarly to the results from HMM and GMM, an unsupervised k-means algorithm produces one cluster for an action from one subject, and another cluster for the same action, performed by another person. By varying the parameters we obtain clusters for different people and different actions.

We argue that the high increase in performance of NN versus GMM and HMM is due to the high dimensionality of the data (NN with full dimensional features performs best) and also because more data is available for this action: 25% of the frames are manually classified as ”stirring.”

7. Discussion and future work

This work presents baseline results from unsupervised temporal segmentation and supervised activity classification from multimodal data. The performance of unsupervised methods is difficult to evaluate in general, however in this case we see promising results in multi-modal data segmentation compared with the chosen level of action annotation. Since multiple levels of action annotation are possible, comparing the unsupervised segmentation with manual labels is ambiguous. However, by varying the model parameters, we show that standard models (GMM, HMM and K-NN) capture some sets of distinct actions. In future work we will explore methods for more robust evaluation of the unsupervised results.

Overall task classification in the context of recipe classification between brownies and omelet from seven subjects shows promising results. We will perform this task on a larger sample - more subjects and more recipe types (this data is already available in the CMU-MMAC database).

From the supervised experiments, initial results show that using a simple K-NN model for frame classification outperforms the standard HMM and GMM models. The results suggest that the data has a high dimensionality which cannot be handled by GMM and HMM. We will explore more robust methods for feature selection and dimensionality reduction in future work. Furthermore, the explored models cluster actions per subject and do not generalize well across people. To address this issue, we will explore individual classifiers per action.

Overall, the presented baseline supervised results show that using gist and IMU data is a reasonable direction in the exploration of daily kitchen action classification. Initial results are promising and bring up many interesting questions regarding action classification in the CMU-MMAC database.

8. Acknowledgements

This work is based on work supported by the National Science Foundation under Grant No. EEEC-0540865.

References

[1] C. B. Abdelkader, L. S. Davis, and R. Cutler. Motion-based recognition of people in eigengait space. In *FGR*, pages 267–274, 2002. 1

[2] L. Bao and S. S. Intille. Activity recognition from user-annotated acceleration data. *Pervasive Computing*, pages 1–17, 2004. 2, 5

[3] J. Barbic, A. Safonova, J.-Y. Pan, C. Faloutsos, J. K. Hodgins, and N. S. Pollard. Segmenting motion capture data into distinct behaviors. In *Graphics Interface*, pages 185–194, 2004. 1, 2

[4] P. Beaudoin, S. Coros, M. van de Panne, and P. Poulin. Motion-motif graphs. In *ACM SIGGRAPH / Eurographics Symposium on Computer Animation*, 2008. 1, 2

[5] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. *Computer Vision, IEEE International Conference on*, 2:1395–1402, 2005. 2

[6] F. de la Torre, J. Hodgins, A. Bargeil, and X. Martin. Guide to the cmu multimodal activity (cmu-mmac) database. http://kitchen.cs.cmu.edu/. In *Technical Report CMU-RI-TR-08-22, Robotics Institute, Carnegie Mellon University, March 2008*, April 2008. 1, 2

[7] A. A. Efros, E. C. Berg, G. Mori, and J. Malik. Recognizing action at a distance. In *In ICCV*, pages 726–733, 2003. 2, 7

[8] A. Fod, M. J. Matarić, and O. C. Jenkins. Automated derivation of primitives for movement classification. *Autonomous Robots*, 12(1):39–54, 2002. 1, 2

[9] G. Guerra-Filho and Y. Aloimonos. Understanding visuo-motor primitives for motion synthesis and analysis. *Comp. Anim. Virtual Worlds*, 17:207–217, 2006. 1, 2

[10] G. Guerra-Filho and Y. Aloimonos. A language for human action. *Computer*, 40(5):42–51, 2007. 2

[11] O. C. Jenkins and M. J. Matarić. Deriving action and behavior primitives from human motion data. In *IROS*, volume 3, pages 2551–2556, 2002. 2

[12] O. C. Jenkins and M. J. Matarić. A spatio-temporal extension to Isomap nonlinear dimension reduction. In *ICML*, 2004. 2

[13] J. Lee, J. Chai, P. S. A. Reitsma, J. K. Hodgins, and N. S. Pollard. Interactive control of avatars animated with human motion data. *ACM Trans. Graph.*, 21(3):491–500, 2002. 1

[14] J. Lester, T. Choudhury, N. Kern, G. Borriello, and B. Hannaford. A hybrid discriminative/generative approach for modeling human activities. In *In Proc. of the International Joint Conference on Artificial Intelligence (IJCAI*, pages 766–772, 2005. 2

[15] Y. Li, T.-S. Wang, and H.-Y. Shum. Motion texture: a two-level statistical model for character motion synthesis. *ACM Trans. Graph.*, 21(3):465–472, 2002. 1

[16] G. Liu and L. McMillan. Segment-based human motion compression. In *ACM SIGGRAPH / Eurographics Symposium on Computer Animation*, pages 127–135, 2006. 1

[17] C. Lu and N. J. Ferrier. Repetitive motion analysis: Segmentation and event classification. *IEEE Trans. Pattern Anal. Mach. Intell.*, 26(2):258–263, 2004. 1

[18] U. Maurer, A. Rowe, A. Smailagic, and D. P. Siewiorek. ewatch: A wearable sensor and notification platform. *Wearable and Implantable Body Sensor Networks, International Workshop on*, 0:142–145, 2006. 2

[19] W. W. Mayol and D. W. Murray. Wearable hand activity recognition for event summarization. In *Wearable Computers, 2005. Proceedings. Ninth IEEE International Symposium on*, pages 122–129, 2005. 2

[20] J. C. Niebles, H. Wang, and L. Fei-Fei. Unsupervised learning of human action categories using spatial-temporal words. *Int. J. Comput. Vis.*, 2008. 1

[21] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175, May 2001. 4

[22] D. Ormoneit, H. Sidenbladh, M. J. Black, and T. Hastie. Learning and tracking cyclic human motion. In *NIPS*, pages 894–900, 2000. 1

[23] Y. Rui and P. Anandan. Segmenting visual actions based on spatio-temporal motion patterns. In *CVPR*, pages 1111–1118, 2000. 2

[24] B. Schiele, N. Oliver, T. Jebara, and A. Pentland. An interactive computer vision system dypers: Dynamic personal enhanced reality system. In *Computer Vision Systems*, pages 51–65. Springer Berlin / Heidelberg, 1999. 2

[25] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: a local svm approach. In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, volume 3, pages 32–36 Vol.3, 2004. 2

[26] J. B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000. 2

[27] P. K. Turaga, A. Veeraraghavan, and R. Chellappa. From videos to verbs: Mining videos for activities using a cascade of dynamical systems. In *CVPR*, 2007. 1

[28] D. D. Vecchio, R. M. Murray, and P. Perona. Primitives for human motion: a dynamical approach. *15th IFAC World Congress on Automatic Control*, 2002. 1

[29] J. Wu, A. Osuntogun, T. Choudhury, M. Philipose, and J. M. Rehg. A scalable approach to activity recognition based on object use. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8, 2007. 2, 3, 4