Video-based Animal Behavior Analysis From Multiple Cameras

Xinwei Xue and Thomas C. Henderson

***Abstract*— It has become increasingly popular to study animal behaviors with the assistance of video recordings. The traditional manual human video annotation is a time and labor consuming process and, the observation results vary between different observers. Hence an automated video processing and behavior analysis system is desirable. We propose a framework for automatic video based behavior analysis systems, which consists of four major modules: behavior modeling, feature extraction from video sequences, basic behavior unit (BBU) discovery and complex behavior recognition. In this paper, we focus on BBU discovery using the affinity graph method on the feature data extracted from video images. We present a simple yet effective way of fusing information from multiple cameras in BBU discovery, and we present and analyze results on artificial mouse video using single, stereo and three cameras. Overall the results are encouraging.**

I. INTRODUCTION

A professor in the medical school and his research group are studying the genetics of certain diseases. In one instance, this requires the determination of the time the lab mouse spends grooming itself, as shown in Figure 1. The traditional way to do this is to first videotape the mouse for a period of time, and then an observer watches the video and records the behaviors of the mouse manually. This is a time and labor consuming process. Moreover, the observation results vary between different observers. Thus it would be a great help if the behaviors could be accurately derived from an automated video processing and behavior analysis system.

Fig. 1. Mouse in a cage

In fact, live subject behavior study has become a very important research area, in which the behavior of various animals or humans is studied for many different purposes. In the context of an animal, the behaviors may include movements (motion), posture, gestures, facial expressions, etc. Animal behavior study originates from areas including biology, physiology, psychology, neuroscience and pharmacology, toxicology, entomology, animal welfare, and so on. The animals mostly studied are mice, rats or rodents, and other animals including ants, poultry, pigs and the like. There are many reasons for studying human behavior, such as smart surveillance, virtual reality, advanced user interfaces, and human motion analysis.

It has become increasingly popular to study behavior with the help of video recordings, since video recordings can easily gather information about many aspects of the situation in which humans or animals interact with each other or with the environment. Also, the video recordings make offline research possible.

A. Automatic Animal Behavior Analysis Framework

We propose a four-module framework for video animal behavior analysis: behavior modeling, feature extraction, basic behavior unit (BBU) discovery, and complex behavior analysis, as shown in Fig 2 (see [1] for a detailed description on relationships between the four blocks enclosed in the dashed box).

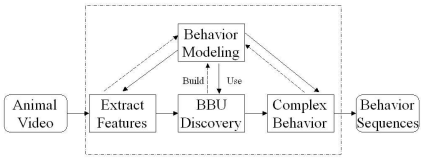


Fig. 2. Work-flow for Video Based Behavior Analysis

Behavior modeling. This step interacts with the other three modules. We need to define, characterize (represent), and model the interesting behaviors in terms of three factors: physical (spatiotemporal) features; the relationship between these behaviors; and the relationship between the animal and its environment.

Feature extraction. To be able to distinguish behaviors, we need to be able to extract sufficient spatiotemporal physical features of the object from video sequences that represent different behaviors. The features may include: the object's position, posture, speed, contour or region pixels, kinematics and dynamics, motion patterns, etc. We may also need to extract features of the environment. This process usually requires the ability to detect and track objects from video sequences.

Discovery of basic behavior units (BBUs), or behavioral segmentation. BBUs are the behavior primitives and higher level analysis will be carried out in terms of these. A BBU

Xinwei Xue is a Ph.D. candidate at School of Computing, University of Utah, Salt Lake City, Utah 84112, USA xwxue@cs.utah.edu

Thomas C. Henderson is with the Faculty of School of Computing, University of Utah, Salt Lake City, Utah 84112, USA tch@cs.utah.edu

can be defined as an activity that remains consistent within a period of time, and that can be represented by a set of spatiotemporal features. This step is based upon successful feature extraction. For the mouse-in-cage example, the BBUs of a mouse in a cage can be resting, exploring, eating, etc. The process of BBU extraction involves mapping the extracted physical features to distinctive behavior units, hence classifying subsequences of the video frames into a sequence of BBUs.

Recognition of complex behaviors. A complex behavior is a behavior consists of multiple BBUs with spatial or temporal relationships between them. It is in a higher level of behavioral hierarchy. Once basic behaviors are discovered, complex behaviors can be constructed and analyzed based upon the relationship between animal's basic behaviors, the interactions of the animal with environment, and with other animals.

In this paper, we focus on BBU discovery with multiple cameras. For simplicity, we tested our proposed algorithm on synthetic mouse video where we know the ground truth. The rest of the paper is organized as follows: The related work on the BBU discovery method is reviewed in Section II. The proposed method on single and multiple cameras is presented in Section III. The experimental results are discussed in Section IV. Finally, the conclusions and future directions are concluded in Section V.

II. Related Work

Video surveillance is a popular research area where the focus has been mostly on humans and vehicles. Most of the existing techniques extract basic behaviors (or actions) directly based upon one or more features extracted (trajectory, motion, posture, etc.) from the detection and tracking results. Pattern recognition techniques (template matching, clustering analysis) are used to classify the video sequence into actions or behavior units, as discussed in the survey papers [2], [3], [4], [5]. These methods are effective in their specific applications. The idea is to utilize all the available distinguishing features to perform classification.

Recently, new approaches based on data (or feature) variance or similarity analysis have been developed for discovering BBUs: PCA-related techniques, and affinity graph-based techniques.

PCA is a classical data analysis tool. It is designed to capture the variance in a dataset in terms of principle components, which is a set of variables that define a projection that encapsulates the maximum amount of variation in a dataset and is orthogonal (and therefore uncorrelated) to the previous principle component of the same dataset. This technique first calculates a covariance matrix from the data, then performs the singular value decomposition (SVD) to extract the eigenvalues and eigenvectors. The eigenvector corresponding to the largest eigenvalue is the *principle component*.

The affinity graph method is also an eigenvalue decomposition technique, or spectral clustering technique. It captures the degree of similarity between the data sequences. Different from the PCA technique, it computes an affinity matrix based upon an affinity measure (e.g., distance, color, texture, motion, etc.) instead of a covariance matrix. The eigenvectors extracted by SVD go through a thresholding step to segment out the first cluster. Then it goes on to process the next eigenvector to find the second cluster, and so on.

PCA-related techniques. Jenkins [6] employs a spatiotemporal nonlinear dimension reduction technique (PCA-based) to derive action and behavior primitives from motion capture data, for modularizing humanoid robot control. They first build spatiotemporal neighborhoods, then compute a matrix D of all pairs' shortest distance paths, and finally perform PCA on the matrix D. Barbic et al. [7] propose three PCA-based approaches which cut on where the intrinsic dimensionality increases or the observed distribution of poses changes, to segment motion into distinct high-level behaviors (such as walking, running, punching, etc.).

Affinity graph method. The affinity graph method has mostly been applied in image segmentation, as summarized in [8]. Recently, this method has been applied to event detection in video [9], [10]. Though not exactly the same approach, the concept of similarity matrix for classification are applied in gait recognition [11] and action recognition [12].

Different affinity measures have been proposed to construct the affinity matrix. In image segmentation, distance, intensity, color, texture and motion have been used [13]. In video-based event detection, as in [10], a statistical distance measure between video sequences is proposed based on spatiotemporal intensity gradients at multiple temporal scales. [9] uses a mixture of object-based and frame-based features, which consist of histograms of aspect ratio, slant, orientation, speed, color, size, etc., as generated by the video tracker. Multiple affinity matrices are constructed based on different features, and a weighted sum approach is utilized for constructing the final affinity matrix.

The most closely related methods to our work are [9] and [10]. [10] constructs an affinity matrix from temporal subsequences using a single feature, while the former constructs the affinity matrices for each frame based upon weighted multiple features.

We are particularly interested in discovering animal behaviors from video sequences. We propose a framework for discovering basic behaviors from temporal sequences based on multiple spatiotemporal features. In our approach, we combine the advantages of the approaches from [9] and [10]: 1) We construct one affinity matrix based on a feature vector consisting of a set of weighted features. The combined features provide us with more information. 2) We construct the affinity matrix on a subsequence of the frame features (multiple-temporal scale), instead of on one frame. Thus we can encode the time trend feature into the problem, and capture the characters of the temporal gradual changes. 3) We apply the affinity graph technique to multiple cameras. Stereo or multiple cameras has been used in human posture classification [14], [15], where either multiple 2D information fusion or reconstructed 3D information is used. Approaches other than the affinity graph method are used. In

our work, we use the multiple camera image information in the simplest way to demonstrate the effectiveness of multiple cameras.

III. BBU Discovery with Multiple Cameras

A. Affinity Graph Method

We propose to use the affinity graph method, an unsupervised learning method to discover basic behavior units. Firstly, the spatiotemporal features are extracted from video frames, as in the Feature Extraction block, shown in Figure 2. Then we take a subsequence (of length *T*) of the features extracted from video images as an element, and calculate the affinity measure between each pair of elements to construct the affinity matrix. Each element overlaps with the next element by a couple of frames, as shown in Figure 3, like a sliding window.

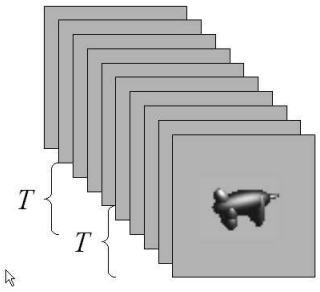


Fig. 3. Demonstration of Video Image Subsequence

This is done by choosing an *element* for consideration. Next a matrix is constructed in which each (i, j) entry gives an affinity (or similarity) measure of the i^{th} and j^{th} elements. The eigenvalues and eigenvectors of the matrix are found, and the eigenvalues give evidence of the strength of a cluster of similar elements. As described in [13], [16], if we maximize the objective function $w_n^T \mathcal{A} w_n$ with affinity matrix $\mathcal{A}$ and weight vector w_n linking elements to the n^{th} cluster, and requiring $w_n^T w_n = 1$, then the Lagrangian is:

$$w_n^T \mathcal{A} w_n + \lambda(w_n^T w_n - 1)$$

which leads to solving $\mathcal{A} w_n = \lambda w_n$. Therefore, w_n is an eigenvector of $\mathcal{A}$. The eigenvector corresponding to the largest eigenvalue is used to partition the data into two clusters. Then we can iteratively partition the eigenvector corresponding to the next significant eigenvalue until there are no more major clusters [13].

After the eigenvector is generated by Single Value Decomposition (SVD), a thresholding technique is applied to partition the eigenvector. In [16], manual threshold selection is used, while in [17], the median or a threshold (by search) that minimizes the CUT or NCUT value (see [17]) is used. Here we take a different approach. We first calculate the accumulative histogram of the eigenvector, and smooth it with a Gaussian kernel, and then find the first threshold value that has gradient value smaller than certain percentage of the number of bins, say 10 percent. This seems to be effective for our experiment.

The affinity measure we use is the exponential function as used in [13], [16], [17]:

$$aff(e1, e2) = exp\{-((f(e1) - f(e2))^t (f(e1) - f(e2))/2\sigma_I^2)\}$$

Our approach differs from the closest literature [9], [10] as described in the related work in four aspects: 1) We construct one affinity matrix based on a feature vector consisting of a set of *weighted* features, instead of calculating affinity matrices for each feature. The combined features provide us with more information. 2) We propose a sequential hierarchical BBU segmentation based upon the distinguishing power of the features. We first use this method to split the video sequences into static and dynamic groups, and then further split the each of the static and dynamic groups into BBUs. 3) We construct the affinity matrix on a *subsequence* of the frame features (multiple-temporal scale), instead of on one frame. Selecting the optimal affinity measure, and time scale (length of the subsequence) is our next step. 4) We also apply this approach to multiple cameras scenario.

B. Affinity Graph Method for Single and Multiple Cameras

For one camera case, each *element* consists of a stack of spatiotemporal features extracted from a subsequence (of length *T*) of video images. Here we denote each element as $E[T][D]$ (D is the feature dimension). For multiple cameras that capture the video synchronously, we simply construct the affinity matrix based on elements that concatenate features from the multiple cameras: e.g., the length of the new feature vector for each image is doubled or tripled and so on. So each element is now $E[T][n * D]$ (n is the number of cameras). This is simple, but we are going to show that it is effective.

C. Feature Extraction and Selection

As in the framework shown in Figure 2, features need to be extracted and selected prior to performing BBU discovery. Here are the critical questions to answer for feature extraction:

1) What agent variables are necessary for BBU identification?
2) What features allow recovery of variable values?
3) What methods to use to extract those features?
4) How does feature error relate to BBU error?

Let's take the mouse-in-cage scenario as an example. We are interested in such BBUs as resting, exploring, eating, and grooming. Assume that we can extract the mouse using simple background subtraction technique. The important variables in distinguishing these behaviors include the kinematics and dynamics (of animal head, body and limbs), posture, and shape, etc. Here for kinematics and dynamics variables, we can extract the position, speed, and orientation change of the mouse; for posture and shape, we can calculate the orientation of the mouse, the bounding box filling ratio and aspect ratio. For the periodic motion like eating, and grooming, thought it is hard to extract the limb motion from the video images, we may instead compute position, speed, orientation, aspect ratio, features from motion history image (MHI) [18], which could reflect these limb and body motion

pattern. Overall, we need to extract spatiotemporal video features that best correspond to the necessary variables for BBU identification.

IV. Results & Analysis

For simplicity, we use a 2000-frame synthetic video (where we have the ground truth) to test our approach. No training session is performed. In design of the artificial mouse video, the mouse behaviors in the real mouse video is mimicked. The environment includes the artificial mouse, a food source (a ball in the middle) and walls. There is no occlusion in the synthetic video.

A. Synthetic Video Generation

We synthesized several clips of mouse-in-cage scenario with ellipsoids, which consists of four behaviors: resting (staying still), exploring (moving around), eating (reaching up to the 'food,' the little ball above the mouse), and grooming (standing on tail with two front legs brushing the head with slight body motion), as shown in Figure 4. The little sphere in the center of image represents 'food.'

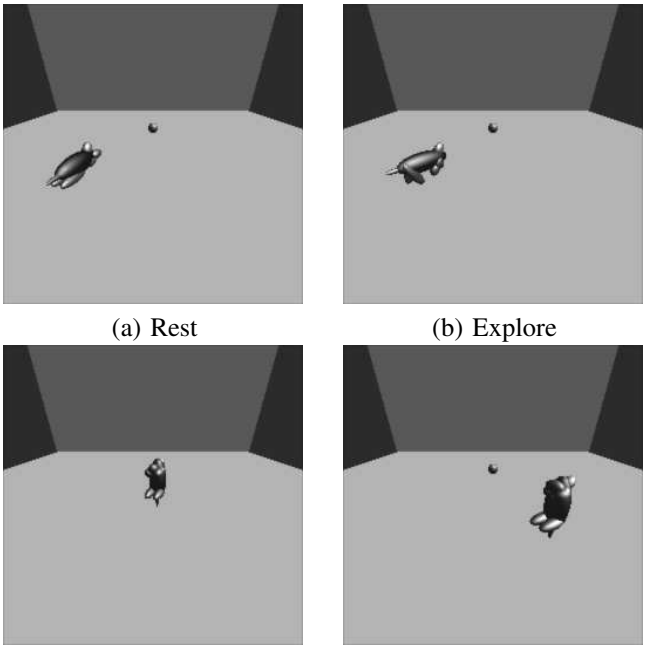
This 2000-frame synthetic video sequence consists of 8 rest segments, 4 segments of eating (reaching up), 2 grooming segments, and the rest are exploring segments. The synthetic behavior transition probability follows a sigmoid function as described the in two-state problem presented in [19]. The labeled behavior sequence is shown in Figure 5.

This synthetic video, though makes it easier for tracking (i.e., by background subtraction, or region competition), is very helpful in studying the effectiveness of the proposed technique. The mouse moves randomly around in a 3D space, sometimes closer to the viewer, and sometimes farther away thus smaller. It may face the camera or away. This requires that the features we use be scale invariant. Also, since the mouse moves in random directions, it increases the technical challenge for BBU discovery.

For multiple cameras, we simply record the video in multiple locations and record the sequences. The three images captured by three cameras are shown in Figure 6.

B. BBU Discovery Results & Analysis

Here we present BBU discovery results on single and multiple cameras. We experimented with the following features extracted from the silhouette of the artificial mouse, as the result of contour tracking or background subtraction: position (centroid of the blob), speed (of the blob centroid), orientation (principle axis of the blob), orientation change, aspect ratio (width/height), aspect ratio change, and similar features of the motion history image (MHI) [18]. We used a subsequence of length 10 ($T = 10$) and slides one frame at a time in the experiments.



(a) Rest (b) Explore

(c) Eat (d) Groom

Fig. 4. Synthetic Mouse-in-Cage Scenario Video Clips

We have tried two approaches: one using combined weighted features in the BBU detection step, the other using a sequential inference approach. The experiment results show that the global motion (i.e, the speed) of the blob is a good feature for segmenting out the frames with no or slight motion. The orientation and its change, and features of MHI are good to separate the grooming (slight global motion, with locomotion) from resting behavior, and separate the reaching up behavior from the exploration behavior. Based upon this observation, we come up with the idea of sequential hierarchical BBU segmentation with the affinity method:

1) Select the feature set with most distinguishing power, and perform affinity method with these features. This will segment the image into several segments.

2) Select the next feature set with most distinguishing power, and perform BBU segmentation with these features on the segments produced by previous step.

3) Repeat step 2) with all or the rest of the features.

Here we first segment the video into static and dynamic sequences using the affinity measure on speed feature in step 1. Then the rest of the features are used to segment the *groom* behavior from the *rest* behavior, and segment the *reached* behavior from the *explore* behavior.

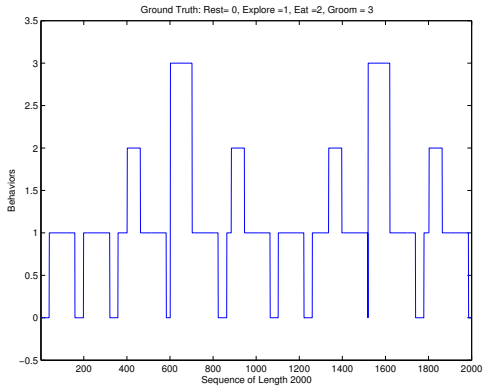
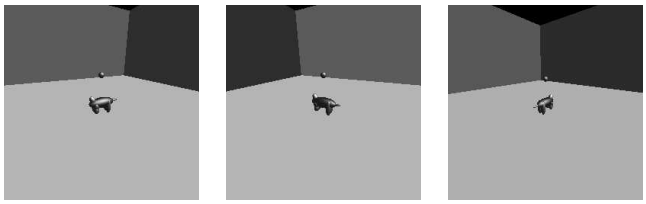


Fig. 5. Behaviors in the synthetic video sequence. Rest= 0, Explore = 1, Eat = 2, Groom = 3



(a) Left (b) Middle (c) Right

Fig. 6. Images Captured by Three Cameras

In our experiment, the BBU segmentation result using multiple cameras achieves better detection accuracy than using only single camera. We have run 5 experiments with one, two and three cameras, with each experiment having random variable controlling the moving speed and direction of the artificial mouse. The results shows unanimous better result with more cameras. The average error rates are about 10%, 8% and 6% for single, stereo and three cameras, respectively (this does not include the errors in the interval between each behavior transition, to account for the size of the subsequence window). Figure 7 compares the static frames discovery results between ground truth, and the best results of single camera, stereo, and three cameras. Figure 8 shows the best BBU result of the corresponding cases among the 5 experiments.

In computational aspect, constructing the affinity matrix and SVD process are two major computation components. The computation time for constructing the affinity matrix is proportional to the square of the number of elements n ($n = nFrames/T$). In our experiment for the 2000-frame synthetic sequence ($T = 2000/10$), it takes about 115 seconds and 3 seconds, respectively to compute these two components and overall about 2 minutes in Matlab on a 1.6GHz laptop with 768MB RAM.

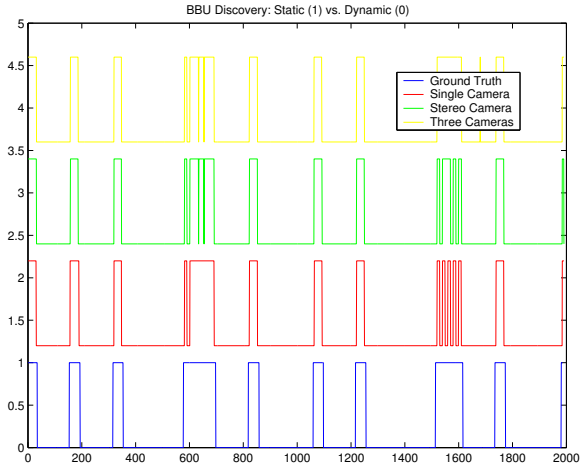


Fig. 7. Discovery of Static Frames: Top row: Three Cameras; Second Row: Stereo Cameras; Third Row: Single Camera; Bottom Row: Ground Truth

The errors come from two major sources, one is the selection of features. In the BBU detection, the distinguishing power of the features is essential. Better spatial-temporal features need to be further explored. The other is the choice of affinity measure and the optimal selection of parameters (such as subsequence length, skip length, weights of features, value of sigma in affinity measure, and the threshold selection for bipartition the eigenvector, etc.), which is the next step of this research.

V. CONCLUSIONS AND FUTURE WORK

We proposed a framework for video based animal behavior analysis. We proposed to apply the affinity graph method to perform BBU discovery using features extracted from single, stereo and multiple cameras, and presented the experimental results on synthetic mouse video. The results are encouraging and promising.

Meanwhile, we have noticed that in applying the affinity method in BBU discovery, optimal feature (spatio-temporal features) and parameter (size of subsequence, and number of frames to skip) selection is critical for the successful behavior clustering.

Also, we are going apply this method to the real mouse video. Our next step will be conducting complex video animal behavior analysis and uncovering underlying behavior models. Mapping the behavior model mechanism in the four blocks in Figure 2 is another area of our research effort. For multiple camera cases, where the cameras shall be deployed to get optimal information[20], and how the more

complicated information fusion techniques can be applied here will also need to be studied in the future.

REFERENCES

[1] X. Xue and T. C. Henderson, "Video-based animal behavior analysis," University of Utah, TechReport UUCS-06-006, June 2006.

[2] J. Aggarval and Q. Cai, "Human motion analysis: A review," *Computer Vision and Image Understanding*, vol. 73, no. 3, 1999.

[3] W. Hu, T. Tan, L. Wang, and S. Maybank, "A survey on visual surveillance of object motion and behaviors," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 34, no. 3, pp. 334–351, 2004.

[4] T. Moeslund and E. Granum, "A survey of computer vision-based human motion capture," *Computer Vision and Image Understanding*, vol. 81, no. 3, pp. 231–268, 2001.

[5] L. Wang, W. Hu, and T. Tan, "Recent developments in human motion analysis," *Chinese Journal of Computers*, vol. 25, no. 3, pp. 225–237, 2002.

[6] O. C. Jenkins and M. J. Mataric, "Deriving action and behavior primitives from human motion data," in *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems(IROS)*, Lausanne, Switzerland, 2002, pp. 2551–2556.

[7] J. Barbic, A. Safonova, J.-Y. Pan, C. Faloutsos, J. K. Hodgins, and N. S. Polland, "Segmenting motion capture data into distinct behaviors," in *Proceedings of Graphics Interface 2004 (GI'04)*, Canada, May 2004.

[8] Y. Weiss, "Segmentation using eigenvectors: a unifying view," in *Proc. IEEE International Conference on Computer Vision*, Kerkyra, Corfu, Greece, 1999, pp. 975–982.

[9] F. Porikli and T. Haga, "Event detection by eigenvector decomposition using object and frame features," in *Workshop on Event Mining, IEEE CVPR*, Washington DC, 2004.

[10] L. Zelnik-Manor and M. Irani, "Event-based analysis of video," in *Proceedings of IEEE CVPR*, Hawaii, 2001.

[11] C. BenAbdelkader, R. G. Cutler, and L. S. Davis, "Gait recognition using image self-similarity," *EURASIP Journal on Applied Signal Processing*, vol. 4, pp. 572–585, 2004.

[12] A. A. Efros, A. C. Berg, G. Mori, and J. Malik, "Recognizing action at a distance," in *IEEE International Conference on Computer Vision*, Nice, France, 2003, pp. 726–733.

[13] D. Forsyth and J. Ponce, *Computer Vision: A Modern Approach*. Upper Saddle River, NJ: Prentice Hall, 2003.

[14] R. Cucchiara, A. Prati, and R. Vezzani, "Posture classification in a multi-camera indoor environment," in *Proc. IEEE International Conference on Image Processing (ICIP)*, vol. 1, Genoa, Italy, 2005, pp. 725 – 728.

[15] S. Pellegrini and L. Iocchi, "Human posture tracking and classification through stereo vision," in *Proc. of Intern. Conf. on Computer Vision Theory and Applicartions (VISAPP)*, Setubal, Portugal, 2006.

[16] P. Perona and W. Freeman, "A factorization approach to grouping," in *Proc. 5th European Conference of Computer Vision (ECCV)*, Freiburg, Germany, 1998, pp. 655–670.

[17] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE*, vol. 22, no. 8, pp. 888–905, 2000.

[18] J. W. Davis and A. F. Bobick, "The representation and recognition of action using temporal templates," in *Proc. IEEE CVPR*, San Juan, Puerto Rico, 1997.

[19] T. C. Henderson and X. Xue, "Construct complex behaviors: A simulation study," in *ISCA 18th International Conference on Computer Applications in Industry and Engineering (CAINE05)*, Hawaii, November 2005.

[20] S. Abrams, P. K. Allen, and K. A. Tarabanis, "Dynamic sensor planning," in *In Proceedings of IEEE International Conference on Robotics and Automation*, 1993.

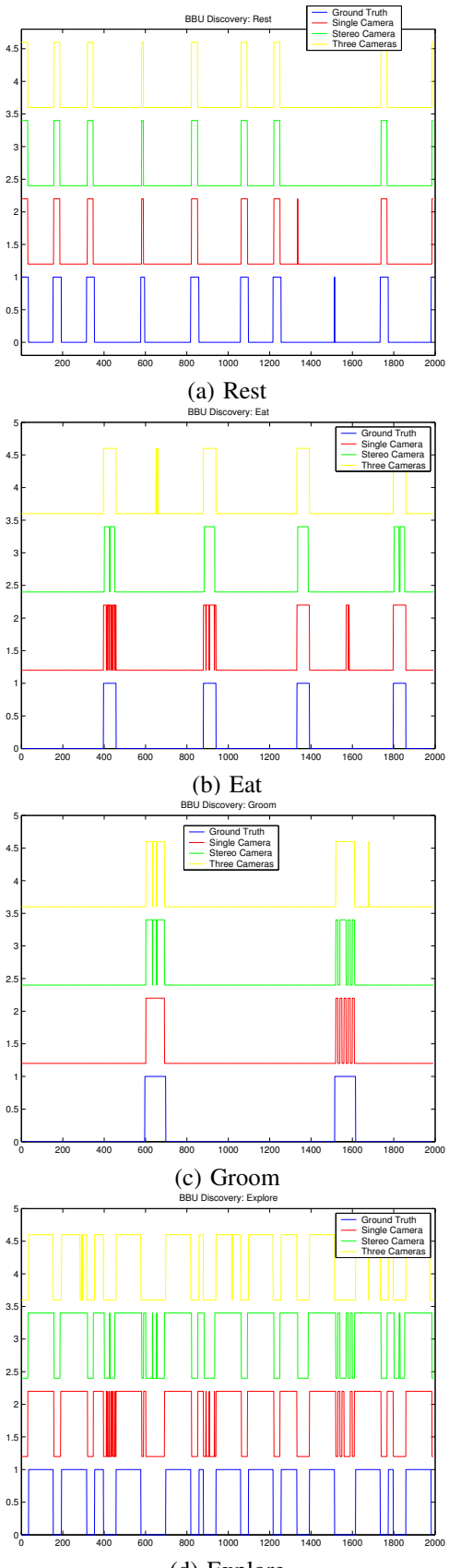


Fig. 8. BBU Discovery Result a) *rest* b) *Eat* c) *groom* d) *explore* Top row: Three Cameras; Second Row: Stereo Cameras; Third Row: Single Camera; Bottom Row: Ground Truth