Unsupervised Learning of Human Expressions, Gestures, and Actions

Stephen O'Hara, Yui Man Lui, Bruce A. Draper
Department of Computer Science
Colorado State University
Fort Collins, CO, USA
{svohara, lui, draper}@cs.colostate.edu

Abstract— **This paper analyzes completely unsupervised clustering of human expressions, gestures, and actions in video. Lacking any supervision, there is nothing except the inherent biases of a given technique to guide grouping of video clips along semantically meaningful partitions. This paper evaluates two contemporary behavior recognition methods, Bag of Features (BOF) and Product Manifolds (PM), for clustering video clips of human facial expressions, hand gestures, and full-body actions. Our goal is to better understand how well these very different approaches to behavior recognition produce semantically useful clustering of relevant data.**

We show that PM yields superior results when measuring the alignment between the generated clusters over a range of K-values (number of clusters) and the nominal class labelling of the data set. A key result is that unsupervised clustering with PM yields accuracy comparable to state-of-the-art supervised classification methods on KTH Actions. At the same time, BOF experiences a substantial drop in performance between unsupervised and supervised implementations on the same data sets, indicating a greater reliance on supervision for achieving high performance. We also found that while gross motions were easily clustered by both methods, the lack of preservation of structural information inherent to the BOF representation leads to limitations that are not easily overcome without supervised training. This was evidenced by the poor separation of shape labels in the hand gestures data by BOF, and the overall poor performance on full-body actions.

I. Introduction

The ability to recognize human behaviors is important for human-machine interaction, video surveillance, and intelligent robotics. Recent research has focused on forced-choice classification tasks over short video clips. Benchmark data sets typically include pre-segmented clips that show only a single behavior from less than a dozen possibilities. The performance task is to classify the clips. While the forced-choice paradigm has led to notable performance gains over the past five years, it leaves many questions unanswered regarding the larger challenge of detecting and recognizing human behaviors in less structured contexts and in continuous streams of input.

Recent results from the Contest on Semantic Description of Human Activities (SDHA Challenge) [1] indicate that existing space-time feature-based approaches perform well on classification, yet detection in continuous videos remains difficult. Additionally, it is desirable to develop learning methods that require minimal supervision because of the difficulty in curating and labelling large data sets and because of the difficulty in generalizing many forced-choice algorithms to uncontrolled environments. Human behavior recognition in streaming video, under real world conditions, is the challenge facing those trying to detect suspicious pedestrian behavior in subway stations, trying to automatically annotate a movie, trying to build household robotics to assist the elderly, and so on. We need to move beyond the paradigm of forced-choice classification of short video clips.

Partially supported by DARPA Contract W911NF-10-2-0066

This paper addresses one aspect of the larger challenge, the unsupervised grouping of behaviors outside of the forced-choice closed-world assumption. Our larger goal is to develop behavior detection and recognition techniques that will be applicable to the "persistent stare" nature of video surveillance. Towards the larger goal, we first seek to understand how contemporary action recognition techniques lend themselves to open-ended clustering (where the value of k is unknown.)

Lacking any supervision, there is nothing except the inherent biases of a given technique to guide grouping of video clips. One might expect extremely poor alignment between the unsupervised clustering and the desired labels (classes) of a given data set. Perhaps surprisingly, this is not always so. In fact, a recent Product Manifold technique for measuring the similarity (distance) of videos generates clusters on the KTH Actions benchmark that are within a few percent of the best supervised classifiers on the same. This is an important result, and forms a key contribution of this paper.

We also show that over three different data sets, the Product Manifold distance measure consistently clusters the data more accurately with respect to the nominal class labelling than a competing Bag of Features method. To understand what algorithm biases may explain this phenomenon, we explore alternative labellings of the data to measure how well they align with a given aspect of similarity among the video clips.

More specifically, we compare a mainstream Bag of Features approach to a Product Manifold based method recently proposed by Lui et al [2]. Each method generates a pair-wise distance matrix to which the same clustering mechanism is applied. We apply both methods to three data sets representing facial expressions, hand gestures, and full-body actions, and evaluate the results of both unsupervised techniques against multiple possible labellings of the data.

This paper is organized as follows. Section II provides background on each method and related literature. Section III

describes our experimental methods, implementation details, and an overview of the data sets. Section IV presents our results with related analysis. We conclude with a summary and description of future work in Section V.

II. Background

This section provides relevant background information on the Bag of Features and Product Manifold methods, and a short discussion of related evaluations. We refer the interested reader to a recent survey on human action recognition [3] for additional background information.

A. Bag of Features

The Bag of Features approach has become one of the most popular methods for human action recognition in short video clips [4], [5], [6], [7], [8], [9], [10], [11], [12]. As adapted from similar methods of image classification and retrieval, Bag of Features approaches represent video clips as unordered sets of local space-time features. Features are quantized into discrete vocabularies, or codebooks. The space-time features in a video are assigned to their nearest neighbors in the codebook. The Bag of Features representation is typically a normalized histogram, where each bin in the histogram is the number of features assigned to a particular code divided by the total number of features in the video clip. Activity classification is often done by applying Support Vector Machines with appropriate kernels (χ^2 is common) to the Bag of Features representations.

There are many choices involved when implementing a Bag of Features approach. One must decide how to sample the video to extract localized features. Possible sampling strategies include space-time interest point operators, grids/pyramids, or random sampling. Each strategy comes with parameters including space and temporal scales, overlap, and other settings. From the sampled regions, an appropriate descriptor must be chosen to provide a balance between discrimination, robustness to small photometric and geometric perturbations, and compactness of representation. Wang et al provide an evaluation of popular space-time interest point detectors and features [13], yet there is no conclusive result indicating which combination of detector and descriptor is best. The results are data-set dependent. Beyond feature detection and extraction, other design choices include codebook size, quantization method (e.g. k-means), and distance function to be used in nearest-neighbor assignments.

Advantages of the Bag of Features approach include the relative simplicity of the representation compared to graphical or constellation models, and the lack of any requirement to pre-process the videos to localize salient parts, perform segmentation, track moving objects, or any other image processing task beyond feature detection. As such, they are attractive for use in unsupervised systems that are designed to sample their environment and learn patterns without prior knowledge. The disadvantages include the difficulty in knowing precisely why two videos are considered similar, as there is little semantic meaning in the representation. For example, it is possible to correctly classify videos due to co-varying, but semantically irrelevant, background artifacts in the data set.

B. Product Manifold

Geometric methods present an alternative approach to those based upon localized sampling. Geometric approaches attempt to map the high-dimensional video data into a lower dimensional space with some regular structure, such as a differentiable manifold. If a video can be represented as a point on a manifold, then the distance between two videos is the geodesic distance between the points. Assuming the geodesic distance can be efficiently computed or approximated, it can be used to classify or cluster the corresponding videos.

A state-of-the-art example of this approach is from a recent paper by Lui et al [2]. Representing a video clip as a 3rd order tensor (an (x,y,t) data cube), Lui applies a modified High Order Singular Value Decomposition (HOSVD) to generate a core tensor and three factor matrices – one for each of the three unfoldings of the 3D tensor into a 2D matrix. Each factor is represented by a Grassmannian manifold. A video clip maps to three points represented by the canonical angles on the three factor manifolds. The product manifold formed by combining the three factor manifolds maps the video to a single point – the Cartesian product of the three canonical angles. The distance between two video clips is the chordal distance on the product manifold, for which a simple closed-form solution exists. Lui shows that the Product Manifold distance coupled with a simple nearest neighbor classifier outperforms competing methods on Cambridge Gestures and KTH Actions data sets.

The advantages of the Product Manifold approach include the relatively small number of design choices, the lack of any training or lengthy codebook generation process, and its computational speed. Using Lui's MATLAB code, the time required to encode two 30-frame video clips and generate the product manifold distance was on the order of ten milliseconds. The disadvantage of this method is the requirement to use fixed-size cubes in the representation. The video clips from the data sets must be cropped or scaled to a uniform-sized cube. The method works best when the activity in the videos is roughly aligned, although it is important to note that Lui's reported results on the KTH dataset includes classes where the actor is moving in different directions and undergoing scale changes, etc.

C. Related Evaluations

To our knowledge, there has been little reporting of unsupervised clustering of human behaviors. Niebels et al employ Probabilistic Latent Semantic Analysis (pLSA) and Expectation Maximization for classifying human actions. While they claim their method is unsupervised, they take advantage of the forced-choice nature of the task to train their probabilistic model over a known number of latent topics (classes), while also using the labels during a validation stage for selecting an optimal vocabulary [10]. Ryoo et al report on the SDHA challenge, which includes tests designed to

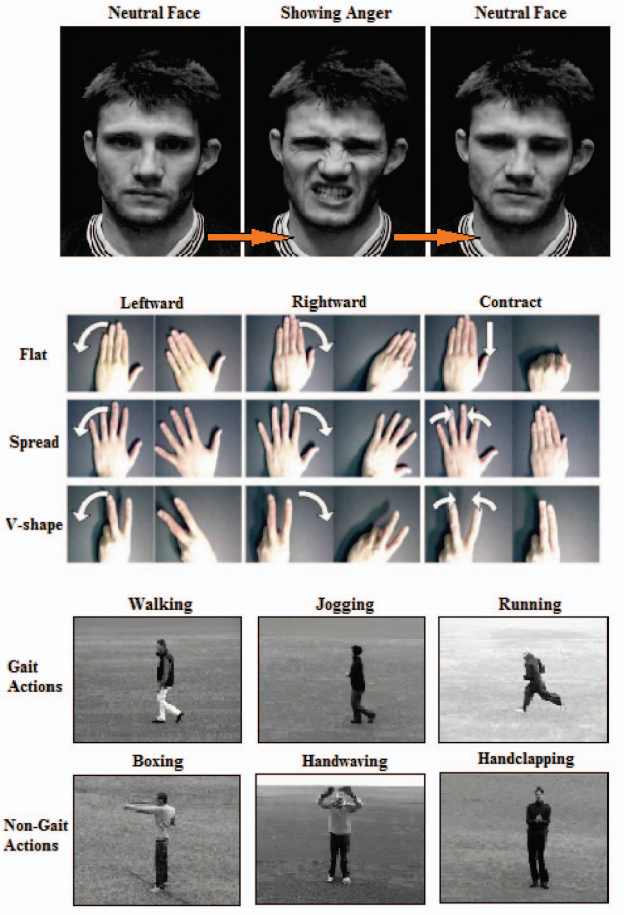

Fig. 1. From top to bottom: Expressions [5], Cambridge Gestures [16], and KTH Actions [4] data sets.

evaluate performance in less controlled environments and on continuous video [1]. However, there were no algorithms which performed well enough on the continuous video challenge to report results. Instead, the reported results focus on forced-choice classification, with top algorithms performing only marginally better than baseline Bag of Features. Wang et al perform an evaluation of local spatio-temporal features in the context of action classification, but again, the focus is on forced-choice tasks [13].

Beyond the forced-choice paradigm, another problem with previous evaluations of Bag of Features action recognition algorithms is credit assignment. As pointed out by others in the context of object/face recognition (see [14], [15]), it is difficult to know what a Bag of Features approach is responding to. Similarly, Lui's Product Manifold approach processes pixels, and thus one may wonder whether his classification accuracy is due to motion, appearance, lighting, or something else.

III. Method

At a high-level, our experimental method is to generate a pair-wise distance matrix using both methods over three data sets relating to human expressions, gestures, and full-body actions. We apply a well-known hierarchical agglomerative clustering routine to the distance matrices to produce dendrograms of the similarity structure between the samples. The dendrogram can be cut at varying levels in the hierarchy to produce different numbers of clusters, from coarser to finer-grained grouping. We vary the number of clusters, k, over a range of values and observe how well the unsupervised grouping of the video clips compares to the desired labels. While we use labels to *evaluate* the clustering, the formation of the distance matrices and subsequent hierarchical clustering is entirely unsupervised. More details of each of these aspects can be found below.

Our intent with this study is not to invent a new algorithm, but rather to provide a comparison of the relative strengths and applicability of two popular approaches to unsupervised grouping of human behaviors. We selected Piotr Dollár's Bag of Features implementation [5], popularly known as the "Cuboids" algorithm, because the well-documented code is readily available upon request from the author, can be used to generate a number of feature descriptors, and generates competitive results. We used Lui's MATLAB implementation of the product manifold algorithm.

A. Data Sets

We selected the following data sets for this study: Facial Expressions [5], Cambridge Gestures [16], and KTH Actions [4]. The samples in each data set are short video clips that exemplify a given expression, gesture, or action, respectively. Figure 1 provides an illustration of each data set.

The Expressions data consists of 6 classes {anger, disgust, fear, joy, sadness, surprise}, repeated in 4 sets. The four sets are comprised of two subjects under two different lighting conditions performing 8 repetitions of all expressions, for a total of 192 videos. Each video clip starts with the subject in a neutral expression, then transitions into one of the expressions, and then back to neutral.

The Cambridge Gestures data consists of 9 classes, repeated in 5 sets of varying lighting, with 20 samples per class per set, for a total of 900 video clips. Each sample is a close-up of a single hand on a uniform background performing one gesture. The 9 classes are divided into three shapes combined with three motions, as illustrated in Figure 1.

The KTH Actions data consists of 6 classes {walking, jogging, running, boxing, handwaving, handclapping}, demonstrated by 25 subjects, each in 4 different scenes, for a total of 600 video clips. The first three scenes are taken outdoors, with a fairly uniform background. The fourth scene is taken indoors, also with a uniform background. Scene 2 varies the scale or angle from Scene 1. Scene 3 varies the clothing of the subject. Three of the classes involve a human gait, while the other three involve stationary actions. The subject varies direction of travel (for the gait classes), and is not always well-centered in the stationary actions.

All three data sets were designed to evaluate forced-choice classification algorithms. For the sake of familiarity within the action and gesture recognition community, we elected to use these same data sets, but in an evaluation scheme that measures unsupervised clustering and how the clusters align

with different potential labellings of the data. Video samples may be similar along different aspects than the externally applied class label, and our evaluation helps illustrate which of those aspects the algorithm is sensitive to.

B. Bag of Features

For the Expressions data, we used the code provided by Dollár, essentially unmodified, because it was developed in conjunction with this data set. Our minor changes were those required to use the Bag of Features representations to generate a distance matrix instead of as input to supervised classification. The code employs the Cuboids detector (separable linear filters, as described in [5]) coupled with the Cuboids descriptor, which is a flattened vector of gradients reduced via PCA to 100 dimensions.

For the Gestures and Actions data sets, we employ the Cuboids detector coupled with Histogram of Oriented Flow (HoF) features. We found this combination to generate the best performance in our tests, and it has been shown to generate good classification accuracy on KTH Actions, as demonstrated by Wang et al's evaluation of space-time features [13]. The HoF descriptor has 440 dimensions, which we employ with no dimensionality reduction. For the Cuboids detector, we set the spatial scale $\sigma = 2$ and the temporal scale $\tau = 3$ for Gestures and $\tau = 4$ for KTH Actions, which agree with the settings in Wang's evaluation.

We use a vocabulary of size 150 for all experiments, selected empirically among sizes ranging from 50 to 1000. The vocabulary was generated by k-means over a random sample of 10% of all the features extracted from the data set. The Bag of Features representation was formed for each video and a pair-wise distance matrix generated using the χ^2 histogram distance function. Due to the randomness inherent in the vocabulary creation, we repeated the process 20 times and chose the vocabulary that generated the best results. For the remainder of this paper, this approach will be labeled "BOF."

C. Product Manifold

We used the code provided by Lui with no modifications beyond those required to generate pair-wise distance matrices on different data sets. Each video clip is rescaled to a 20x20x32 tensor. Through the HOSVD, the tensors are projected onto the product manifold, and the pair-wise distances computed. For the remainder of this paper, this approach will be labeled "PM."

D. Cluster Accuracy

We define *cluster accuracy* as the percentage of samples that were of the majority in their respective clusters. The minimum score always occurs when $k = 1$, in which case the cluster accuracy is the ratio of the number of samples in the largest class to the total number of samples in the data set, N. At the other extreme, when $k = N$, the cluster accuracy will be 1.0, as all samples will be assigned unique clusters and thus there will be no cluster "impurity." We desire to minimize k while maximizing cluster accuracy. The computation is shown formally in Eq. 1, where C is the set of k clusters, x_i are the data points being clustered, and $|\cdot|$ indicates set cardinality.

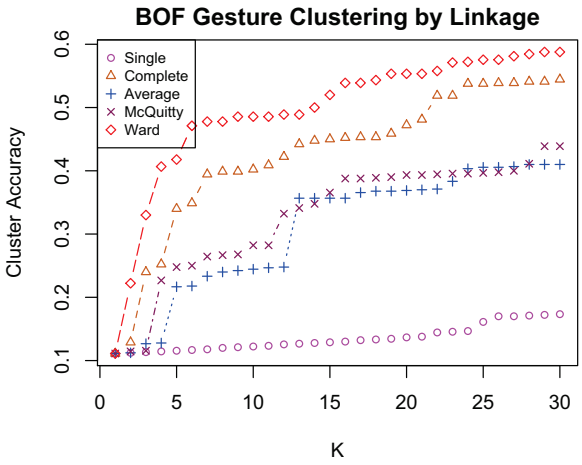

Fig. 2. Comparison of hierarchical clustering linkage methods. This graph is formed using the Bag of Features method on the Gestures data set, but performance was similar for both algorithms on all three data sets. All the other experiments in this paper employ Ward's linkage for hierarchical clustering.

$$\begin{gathered} C = \{C^1, C^2, \ldots, C^k\} \\ X_L^k = \{x_i | x_i \in C^k \wedge Label(x_i) = L\} \\ \sum_{i=1}^{k} \max_{L \in Labels} |X_L^k| / |C^k| \end{gathered} \tag{1}$$

E. Hierarchical Clustering

We use an agglomerative hierarchical clustering method to group similar video clips. We tested several linkage methods and found that Ward's algorithm, which seeks to minimize the incremental increase in cluster variance, had superior clustering results. In addition to Ward's method, we evaluated Single linkage (nearest neighbor between cluster members), Complete linkage (furthest neighbor), Average linkage (Average distance), and McQuitty's linkage (weighted average based on recursive agglomerations).

Figure 2 shows a comparison of different linkage methods when employing the Bag of Features algorithm for measuring the similarity of Gestures. In this figure, we plot the cluster accuracy of the Gesture class label against the number of clusters, K, which was varied from 1 to 30. We perform a single hierarchical clustering per curve, and we vary K by selecting the cut in the hierarchy that yields the appropriate number of clusters. When K is unknown, the full curve may be more indicative than any single point in measuring performance. Regarding the selection of linkage method, Figure 2 is representative of the results over all data sets and with both BOF and PM implementations – Ward's linkage is the best choice in all cases.

Hierarchical clustering is used because it produces deterministic results and it is easy to vary the number of clusters (k), by cutting the tree at an appropriate point in the hierarchy. In a completely unsupervised learning environment,

the number of class labels, k, is not known, so the different levels of similarity/generalization provided by hierarchical clustering are appropriate.

IV. Results

We compared Bag of Features and Product Manifold methods for clustering facial expressions, hand gestures, and full-body actions. Each set of experiments is described below. A summary comparison of the relative performance of BOF and PM is illustrated in Figure 3. This figure presents the performance curve when the generated clusters are compared against the nominal class labels provided by the data set. There are 6 classes in KTH Actions and Expressions, and 9 classes for Gestures, indicated by the vertical dotted black line. The solid red curve shows the cluster accuracy of PM over all K, the amber dotted curve shows BOF.

From this figure, PM tends to outperform BOF over all data sets. Also noteworthy is the significant performance difference in two of the three data sets, while one (Expression) yields comparable results. This may be in part because Dollár developed both the Expressions data set and the BOF implementation we adapted for this study, and thus the implementation may have a level of tuning for this data set not present in the others. However, we believe other factors are involved, which we present later.

A key result shown in the summary results is the performance of PM on KTH Actions. At $K = 6$, the cluster accuracy is 90.7%, which is only slightly below state-of-the-art *classification* methods that rely on supervised training, and improves over Niebels et al [10] results of 81.5% with their unsupervised pLSA method (which takes advantage of a known K value.) Equally interesting, perhaps, is the poor performance of BOF on the same. In supervised settings employing Support Vector Machines, BOF methods using the same interest point detector and HoF features as selected for our study score near 90% on KTH Actions (as reported in [13]). We validated a similar result with our setup. The substantial drop in performance experienced by BOF when moving from supervised to unsupervised methods was somewhat surprising, especially in light of the lack of a corresponding drop with PM. This may illustrate the relative importance of supervision in achieving high classification rates with the two methods. We explore these and other aspects in more detail, presented according to data set, below.

A. Expressions

We compared the clusters generated on the Expressions data to four labellings: the nominal Expression label from the data set (6 classes), the Set label (4 sets), and labels for Subject (2) and Lighting (2). Figure 4 shows the results. Although the performance of the two methods is similar for Expression, Set, and Lighting labels, BOF clustering is much more closely aligned to subject identity than PM, as evidenced by the significantly higher curve.

With BOF, the Subject labelling generates less cluster impurity than Expressions. While the higher curve is indicative of the fact that there are only two subjects as opposed to six expressions, it is also true that with PM, the Subject labelling does not behave the same way. The subject identity is seemingly less useful to PM when grouping the expression video clips than it is with BOF. This leads to the speculation that if this small data set were expanded to include many more subjects, the sensitivity to subject identity evidenced by BOF may lead to decreased cluster accuracy when labelling by Expression, while PM performance might be less effected.

B. Gestures

We evaluated BOF and PM clustering against the following labels applied to the Cambridge Gestures data set: Gesture (the nominal class label, 9 classes), Set (5 sets with varying lighting conditions), Direction of motion (3 motions as per Figure 1), and Shape (Flat, Spread, and V-Shape). Results are shown in Figure 5.

One immediately obvious aspect of Figure 5 is that both methods generate clusters that are nearly completely separable along direction of motion (98% accuracy at all k ranges for BOF and 100% for PM). At the same time, Gesture class labelling is nearly identical in performance to Shape labelling for both methods. We hypothesize that the hierarchical clustering groups the data first by motion direction, and later by shape. Further, because the overall performance of BOF is much lower than PM, it may be that PM is doing a much better job differentiating shape, while BOF struggles in this regard. This would not be surprising because BOF discards locations of features in the representation. As such, the histogram of space-time features located near the fingertips of the spread hand and flat hand may look very similar, and thus difficult to differentiate. The Product Manifold method, however, treats all pixels equally, preserving location information, and thus having less confusion between the hand shapes. To test this hypothesis, we further investigate the details of how clusters align to labels in the Gesture data set.

Given the strong affinity for both methods with the three gesture directions, we investigated the cluster accuracy when comparing the nominal class labels (Gesture) to clusters when $K = 3$. The result in Table I shows that both methods nearly perfectly cluster along motion direction, as expected from Figure 5.

TABLE I
GESTURE LABELS COMPARED TO 3 CLUSTERS WITH BOF AND PM.

	BOF Cluster ID			PM Cluster ID		
Label	1	2	3	1	2	3
Flat Left	99	0	1	100	0	0
Spread Left	100	0	0	100	0	0
V Left	100	0	0	100	0	0
Flat Right	0	98	2	0	100	0
Spread Right	1	99	0	0	100	0
V Right	1	97	2	0	100	0
Flat Contract	0	4	96	0	0	100
Spread Contr.	0	6	94	0	0	100
V Contract	0	2	98	0	0	100

When we raise K from 3 to 9, the number of nominal classes in the Gesture data set, we see that the two algorithms

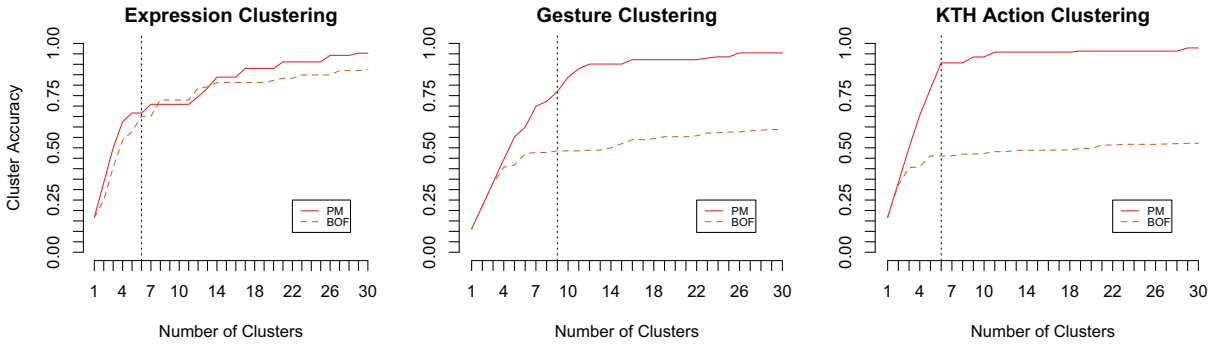

Fig. 3. PM vs BOF methods compared against the nominal class labels on all three data sets. Vertical line indicates number of nominal classes in the data set (6,9,6, respectively). PM cluster accuracy on KTH Actions (90.7% at K=6) is comparable to the scores of state-of-the-art supervised classification methods on the same.

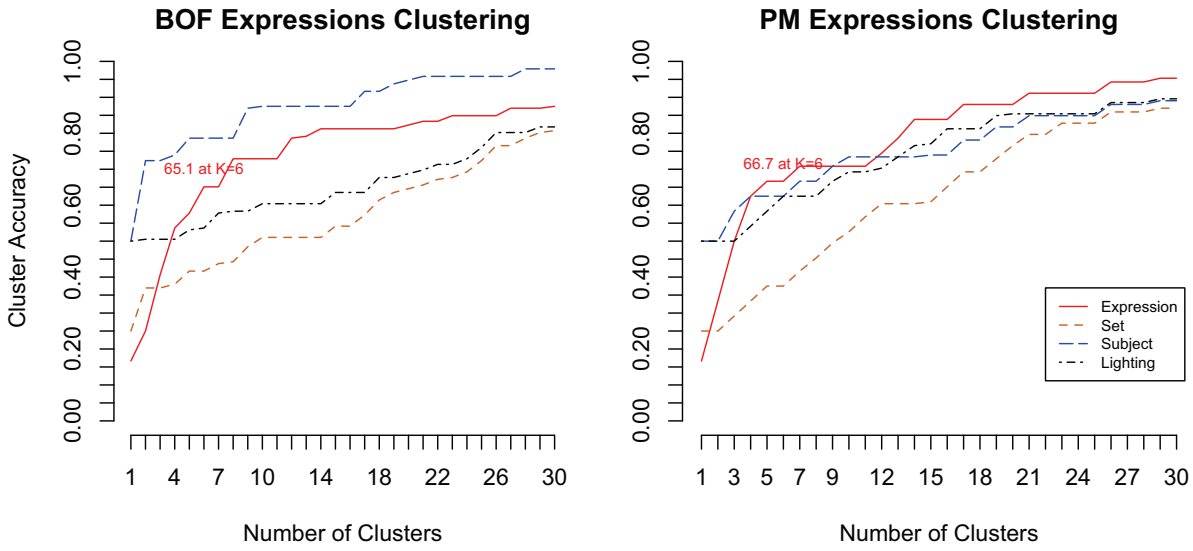

Fig. 4. Clustering of Expressions data. BOF clusters are more closely aligned to the Subject labelling, but both methods perform similarly on other labels.

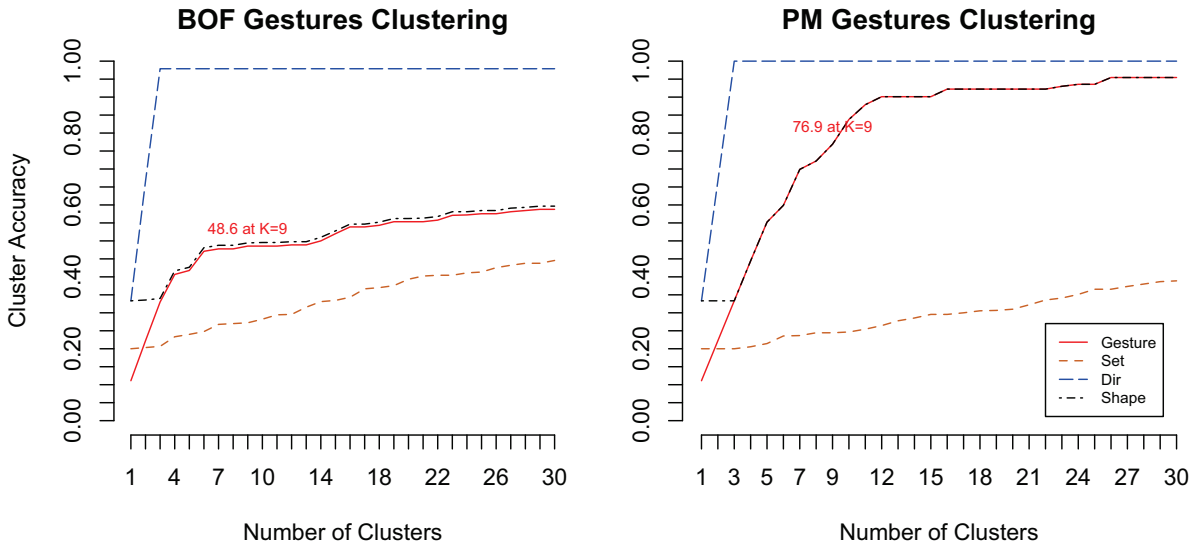

Fig. 5. Clustering of Gestures data. Direction of motion is nearly perfectly separated by both methods. Both methods show that Gesture class labelling is limited by the Shape.

behave differently, as shown in Tables II and III. While PM begins to differentiate based on shape, BOF struggles to do so. BOF maintains its confusion between shapes within the same direction, while PM manages to cleanly separate Leftward motion into the three Shapes, and partially separate the Contraction motion as well. This evidence supports our hypothesis that shape is a secondary aspect of the clustering behind motion, and it proves to be the limiting factor on the overall agreement between the class labels and the clusters.

Restating an earlier point, with no supervision, it is the inherent biases of the two methods that dictate which generates clusters that are better aligned with semantically-meaningful partitions. In this case, the bias of BOF to ignore relative spatio-temporal positions causes it to fail in many instances to match the nominal gesture label.

TABLE II
LABELLING GESTURE COMPARED TO 9 CLUSTERS WITH BOF.

	Cluster ID								
Label	1	2	3	4	5	6	7	8	9
Flat Left	95	4	0	0	0	0	0	1	0
Spread Left	46	54	0	0	0	0	0	0	0
V Left	26	74	0	0	0	0	0	0	0
Flat Right	0	0	24	35	9	13	17	2	0
Spread Right	0	1	30	25	20	6	18	0	0
V Right	0	1	25	26	18	4	24	2	0
Flat Contract	0	0	1	2	0	0	1	92	4
Spread Contr.	0	0	2	3	0	0	1	40	54
V Contract	0	0	1	0	1	0	0	52	46

TABLE III
LABELLING GESTURE COMPARED TO 9 CLUSTERS WITH PM.

	Cluster ID								
Label	1	2	3	4	5	6	7	8	9
Flat Left	93	0	0	0	0	0	0	7	0
Spread Left	3	0	0	0	94	0	0	3	0
V Left	0	0	0	0	0	0	0	100	0
Flat Right	0	43	57	0	0	0	0	0	0
Spread Right	0	85	15	0	0	0	0	0	0
V Right	0	62	38	0	0	0	0	0	0
Flat Contract	0	0	0	100	0	0	0	0	0
Spread Contr.	0	0	0	0	0	80	20	0	0
V Contract	0	0	0	17	0	0	41	0	42

C. Actions

We chose the following labels to apply to the KTH Actions data set: Action (the nominal class label, 6 classes), Scene (4 scene types), Gait (2 types: gait or non-gait actions, as per Figure 1), Location (2 types: indoors and outdoors, 75% are outdoors), and Subject (25 people). Results are shown in Figure 6. We did not expect either method to align clusters against the Subject label, as the individuals can be hard to discern, and Scene 3 uses changes of clothing to further make identifying the subject difficult. Separating the actions based on Gait labelling proved easy for both methods. Although the performance curve for Location appears high, the base rate is 75% outdoors, and the results did not rise much above that minimum score. Clustering based on PM distances was very closely aligned to the nominal class labels, as shown by the 90.7% cluster accuracy at $K = 6$, which as mentioned before, is competitive with supervised classification approaches.

Unlike with Gestures, we did not discover a semantic labelling that best explains the performance of the nominal class labels. Given the high performance of PM clustering on the class labels, one is led to believe that the classes are inherently separable in most cases when using the PM representation, but not when using BOF.

Echoing an earlier statement, given that Support Vector Machines trained with similar BOF representations achieve classification accuracies in the upper 80's to lower 90's%, it was surprising that the clustering performance was so comparatively poor on KTH Actions. Because of this, we believe that supervised training may be more important for achieving high accuracy with BOF representations of full-body actions than it is for PM representations.

V. CONCLUSION

A. Discussion

Lacking any supervision, and outside a forced-choice paradigm, it is important to design representations that are amenable to clustering human activities along semantically meaningful aspects. We presented performance differences between Product Manifold and Bag of Features representations over three data sets representing, respectively, human expressions, hand gestures, and full-body actions. The pairwise distance matrices generated by Product Manifold representations of the video clips led to consistently superior clustering accuracy when compared with the nominal class labels of each data set. Further, unsupervised clustering of KTH Actions using the Product Manifold approach yields accuracy only slightly worse than the best supervised methods.

Both PM and BOF excelled at separating direction of gesture motion and easily separated Gait from Non-gait actions. There is no corresponding gross motion involved in the Expression data, which involves only local deformations of the face without any significant global motion within the frame.

We also found that while gross motions were easily clustered by both methods, the lack of preservation of structural information inherent to the BOF representation leads to limitations that are not easily overcome without supervised training. This was evidenced by the poor separation of Shapes in the hand gestures data by BOF, and the overall poor performance on full-body actions.

B. Future Work

In future work, we plan to apply the Product Manifold technique in less controlled environments and on continuous data streams. This will require applying a pedestrian detection and tracking algorithm for isolating the space-time data cube and scaling it to a fixed-size tensor. Given the fast computational speed of this approach, multiple hypotheses could be tested over various sliding windows while maintaining real or near-real time performance.

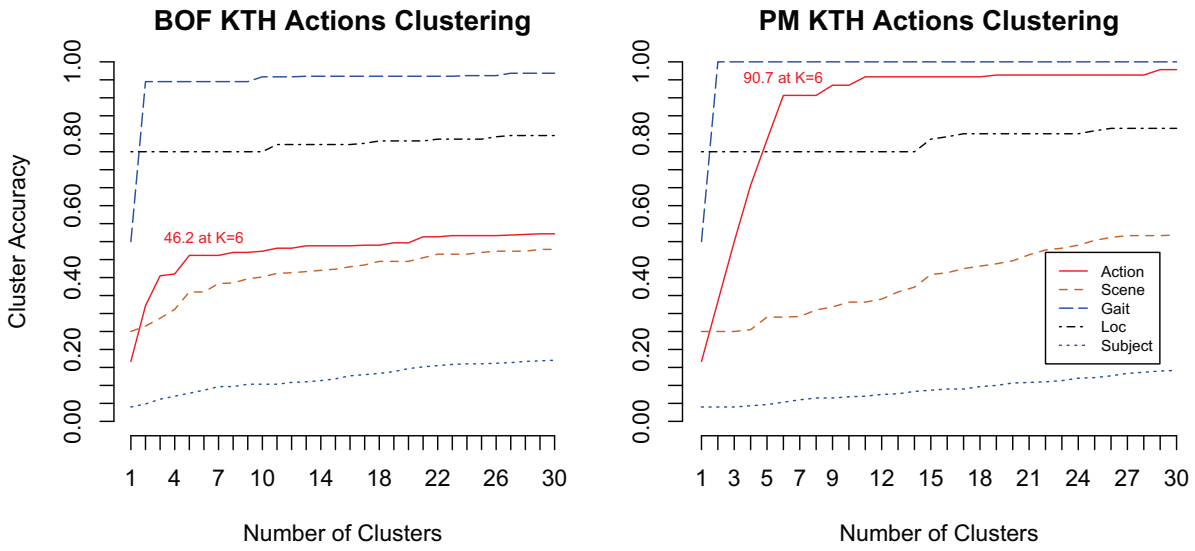

Fig. 6. Clustering of KTH Actions data. Both methods easily distinguish between Gait and Non-Gait actions. PM clustering performs excellently when judged against the nominal class labels.

The clustering method will likely need to be adapted to an online method with outlier rejection for learning salient behaviors in continuous data streams, while avoiding clustering noise. By combining online unsupervised learning methods with Product Manifold distance measures, we hope to make significant advances towards our larger goal of developing behavior recognition capabilities in less controlled, non forced-choice scenarios operating on continuous data streams. This capability is what is ultimately required for the "persistent stare" needs of the video surveillance community and for advancing human-robot interactions.

VI. Acknowledgements

We would like to thank Piotr Dollár for the code he has made available.

References

[1] M. S. Ryoo, C. C. Chen, J. K. Aggarwal, and A. Roy-Chowdhury, "An overview of contest on semantic description of human activities (SDHA) 2010," in *Proc. ICPR*, 2010.

[2] Y. M. Lui, J. R. Beveridge, and M. Kirby, "Action classification on product manifolds," in *Proc. CVPR*, 2010.

[3] R. Poppe, "A survey on vision-based human action recognition," *Image and Vision Computing*, vol. 28, no. 6, pp. 976–990, 2010.

[4] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: A local SVM approach," in *Proc. ICPR*, 2004.

[5] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *2nd Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005*, 2005.

[6] Y. Ke, R. Sukthankar, and M. Hebert, "Efficient visual event detection using volumetric features," in *Proc. ICCV*, 2005.

[7] I. Laptev, "On space-time interest points," *International Journal of Computer Vision*, vol. 64, no. 2, pp. 107–123, 2005.

[8] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 12, pp. 2247–2253, 2007.

[9] P. Scovanner, S. Ali, and M. Shah, "A 3-dimensional sift descriptor and its application to action recognition," in *Proceedings of the 15th international conference on Multimedia*. ACM, 2007.

[10] J. C. Niebles, H. Wang, and L. F. Fei, "Unsupervised learning of human action categories using spatial-temporal words," *International Journal of Computer Vision*, vol. 79, no. 3, pp. 299–318, 2008.

[11] K. Rapantzikos, Y. Avrithis, and S. Kollias, "Dense saliency-based spatiotemporal feature points for action recognition," in *Proc. CVPR*, 2009.

[12] A. Kovashka and K. Grauman, "Learning a hierarchy of discriminative Space-Time neighborhood features for human action recognition," in *Proc. CVPR*, 2010.

[13] H. Wang, M. M. Ullah, A. Klser, I. Laptev, and C. Schmid, "Evaluation of local spatio-temporal features for action recognition," in *Proc. BMVC*, 2009.

[14] J. Ponce, M. Hebert, C. Schmid, A. Zisserman, J. Ponce, T. Berg, M. Everingham, D. Forsyth, M. Hebert, S. Lazebnik, M. Marszalek, C. Schmid, B. Russell, A. Torralba, C. Williams, J. Zhang, and A. Zisserman, "Dataset issues in object recognition," in *Toward Category-Level Object Recognition*, ser. Lecture Notes in Computer Science. Springer Berlin / Heidelberg, 2006, vol. 4170, pp. 29–48. [Online]. Available: http://dx.doi.org/10.1007/11957959_2

[15] N. Pinto, J. DiCarlo, and D. Cox, "How far can you get with a modern face recognition test set using only simple features?" in *Proc. CVPR*, 2009.

[16] T. K. Kim, S. F. Wong, and R. Cipolla, "Tensor canonical correlation analysis for action classification," in *Proc. CVPR*, 2007.