Joint Sparsity-based Representation and Analysis of Unconstrained Activities

Raghuraman Gopalan
Video and Multimedia Technologies Research Dept., AT&T Labs-Research, Middletown, NJ 07748 USA
raghuram@research.att.com

Abstract

While the notion of joint sparsity *in understanding common and innovative components of a multi-receiver signal* ensemble *has been well studied, we investigate the utility of such joint sparse models in representing information contained in a* single *video signal. By decomposing the content of a video sequence into that observed by multiple spatially and/or temporally distributed* receivers*, we first recover a collection of common and innovative components pertaining to individual videos. We then present modeling strategies based on subspace-driven manifold metrics to characterize patterns among these components, across* other *videos in the system, to perform subsequent video analysis. We demonstrate the efficacy of our approach for activity classification and clustering by reporting competitive results on standard datasets such as, HMDB, UCF-50, Olympic Sports and KTH.*

Figure 1. Frames sampled from a football game. While the video can be lengthy, a succinct description of its content can be expressed by a broad theme (a contest between two teams in a stadium with several onlookers) and certain interesting aspects of the game such as various strategies used by players for offense and defense, different crowd reactions, player celebrations and so on. We pursue such an intermediate representation of a video based on the principles of joint sparsity, and perform activity classification and clustering using modeling techniques on Grassmann manifolds.

1. Introduction

Understanding information contained in videos is a well-studied problem that has found utility in applications such as activity recognition, object tracking, search and retrieval, summarization among others [39]. The focus of this work is on activity analysis (that includes actions, events, and any temporal semantic in general), a problem that has seen a gradual transition from constrained data acquisition settings with static background and simple foreground object motions [37] to unconstrained videos collected from YouTube [21, 1] and surveillance scenarios [29].

As with other visual recognition tasks, there has been emphasis on both designing features to represent action patterns and learning strategies to derive pertinent information from features to perform robust activity classification. While there has been a gamut of feature representations ranging from spatio-temporal volumes [13, 3] and trajectories [32, 42] to local interest point descriptors [36, 20] and action attributes [11, 24], they have been complemented by modeling techniques such as feature pooling [5], pyramid matching [22], and local expert forests [26] to address robustness issues in dealing with unconstrained videos.

Since many videos can be qualitatively described in terms of a broad theme(s) and certain interesting aspects that stand out, for instance a football video sequence in Figure 1 where the theme could be a set of players in the field being watched by several onlookers and interesting aspects could be various offensive and defensive strategies exercised by the players, we seek to obtain an *intermediate* representation of videos portraying such information. We observe that such an analogy has been studied in multi-receiver communication settings in the form of distributed compressive sensing [2], where the goal is to harness *joint sparsity* in terms of *common* and *innovative* components present in a signal ensemble collected at multiple receivers, under some assumptions on the properties of the ensemble.

The goal of this work is to study the utility of such joint sparse models in the context of a *single* video, by expressing the content of a video sequence into that of an ensemble observed by multiple *receivers* that are spatially and/or temporally distributed in that video. Starting with an ini-

tial representation of such a signal ensemble, in the form of spatio-temporal bag-of-features [22], we first recover the intermediate joint sparse representation of the video in terms of common and innovative *atoms* pertaining to the ensemble. We then present modeling strategies to perform activity classification and clustering by analyzing the subspace spanned by atoms corresponding to each video and then pursuing a Grassmann manifold interpretation of the space spanned by subspaces corresponding to all videos in the system (or dataset). Before getting into the details of our approach, we first overview recent efforts on unconstrained activity analysis. We refer the reader to [39] for a comprehensive review of literature spanning atleast two decades.

1.1. Related Work

Unconstrained activity analysis under multiple sources of variations such as camera motion, inter-object interaction, the associated background clutter and changes in scene appearance due to illumination, viewpoint etc. is receiving recent attention in part due to the proliferation of video content in consumer, broadcast and surveillance domains. One of the earliest attempts in analyzing such activities was by Laptev et al. [22] that proposed using space-time bag of features and pyramids with a non-linear support vector machine classifier to learn realistic human actions from movies. Liu et al. [25] addressed a more challenging set of videos collected from YouTube by obtaining visual vocabularies from pruned static and motion features. Learning a discriminative hierarchy of space-time features was pursued by [20], while [42] investigated dense trajectories corresponding to local features and [33] clustered such trajectories into action classes using graphical models. Signal coding-based approaches have been studied by [15] that used sparse coding principles to obtain a generic mid-level video representation termed 'video primal sketch', [41] that proposed efficient sparse random projection algorithms for video classification, [31] that presented an information maximization approach for learning sparse action attribute dictionaries, and [19] that encoded motion interchange to decouple image edges from motion edges to facilitate better understanding of events. Classifier/feature ensemble-type approaches have also been proposed for instance, the action-bank [35] that produces a semantically-rich action description by pooling inputs from several action detectors, scene-aligned pooling [5] that decomposes video features into concurrent scene components to capture diverse content and dynamic scene semantics of a video, and local expert forests [26] for score fusion to account for imbalanced event class distributions. More recently, there has been an increasing focus on designing multi-modal event attributes [11, 34] to capture visual, text and audio information prevalent in the social-media space, and classifier adaptation [9, 23, 8] to deal with novel event instances that were unseen during training. Besides this, there are works on interesting video applications such as understanding collective crowd behavior [50], detecting daily activities from first-person camera views [30], and early event detection [16], and evaluation studies of different features/techniques on standardized consumer [38] and surveillance [29] datasets.

The focus of this work is to analyze activities by obtaining a mid-level, intermediate video representation based on joint sparsity principles [2], which aims at understanding information that a signal ensemble shares and varies upon. While studies on joint sparsity are prevalent in multi-receiver communication settings [2], they have also been applied for vision problems[1] involving image sets such as expression-invariant face recognition [27], face recognition under lighting and occlusion [49], multimodal image fusion [47] and target detection in hyperspectral imagery [6]. However there hasn't been much work on understanding the utility of joint sparsity for video analysis, which is one of the main motivations behind this paper.

2. Proposed Approach

2.1. Problem Description

Let $\mathcal{V} = \{V_i\}_{i=1}^{N}$ denote the set of N videos in the system belonging to m activity classes. For each video V_i, we first decompose it into several spatially and/or temporally distributed segments $\{V_i^j\}_{j=1}^{N'}$ and obtain its joint sparse representation $J(V_i)$ consisting of a set of common and innovative components, C_i and I_i respectively. With the collection of such intermediate representations $\mathcal{J} = \{J(V_i)\}_{i=1}^{N}$ corresponding to all videos in the system, we perform subsequent video analysis by learning $f(\bar{C}, \bar{I})$ where $\bar{C} = \{C_i\}_{i=1}^{N}$, $\bar{I} = \{I_i\}_{i=1}^{N}$ and f is modeled using strategies based on Grassmann manifolds. We address both classification and clustering scenarios, by adapting f to account for activity labels $l_i \in \{1, 2, ..., m\}$ accompanying V_i or otherwise, and perform experiments on several standard unconstrained activity datasets. Figure 2 presents an overview of our approach, and the details are provided in the following sub-sections.

2.2. Joint Sparse Representation of a Video

Given a video V_i we first extract twenty four segments $\{V_i^j\}_{j=1}^{24} = V_i^s \cup V_i^t \cup V_i^{st}$ that are grouped into spatial V_i^s, temporal V_i^t and spatio-temporal V_i^{st} ensembles. V_i^s consists of four spatially distributed segments that are created by dividing a frame into four quadrants and then considering the information in each quadrant across all other frames in the video. V_i^t is made up of four temporally distributed

[1] We'd like to point out that while [48] performs joint sparsity-based visual classification, they refer to 'joint sparsity' in the context of multi-task learning and not as discussed in this paper (which is along the lines of [2]).

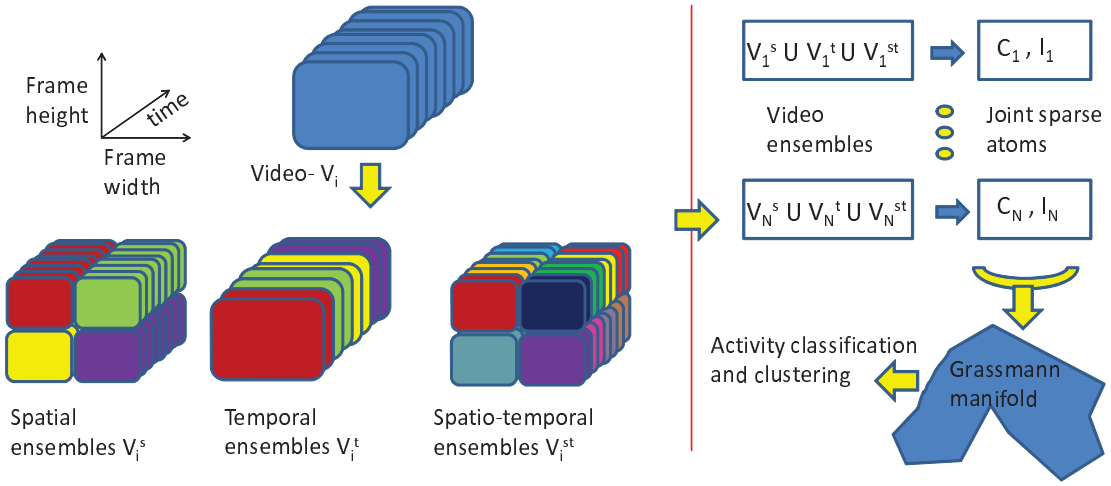

Figure 2. An overview of the proposed approach. Step-1: Extracting spatial V_i^s, temporal V_i^t and spatio-temporal V_i^{st} ensembles (containing 4,4, and 16 video segments V_i^j respectively) from a video V_i and utilizing joint sparse models to obtain an intermediate representation $J(V_i)$ consisting of common C_i and innovative I_i atoms. Step-2: Modeling these joint sparse atoms across all N videos in the system, using techniques on the Grassmann manifold, to perform subsequent video analysis such as activity classification and clustering. *This figure is best viewed in color*.

segments obtained by dividing the video into four equal intervals along the temporal dimension. We then consider each spatial segment pertaining to each temporal interval to make up 16 spatio-temporal segments that represent V_i^{st}. For each of these segments we obtain the spatio-temporal bag-of-features representation, a d-dimensional histogram, using the method of [22]. For sake of clarity, we defer more details on feature extraction until the experiment section.

We now recover common and innovative components pertaining to each of these three ensembles using the principles of joint sparsity [2], to obtain the intermediate representation $J(V_i)$ of a video V_i. Three joint sparse models (JSM) have been investigated in [2] in the context of multi-receiver communication settings by imposing assumptions on the sparsity of common and innovative components of the signal ensemble. We first present a qualitative analogy of these models in terms of information contained in a video, to facilitate intuitions on the applicability of JSM to video analysis. JSM-1 represents the case where both the common and innovative components are sparse. This could translate to a video with a fixed background and few foreground motions such as a camera stream in an office entrance observing people entering the building between 9am and 6pm on a weekend. While there could be sparse changes in the global background due to variations in the daylight intensity, the innovations will also be sparse since number of people at work on a weekend is relatively less than that during weekdays. JSM-2 corresponds to cases where the common component is ideally zero whereas the innovations are sparse with similar supports. This could correspond to video highlights of a day's events at the Olympics. While the video may cover a range of sports that could be very different (no common component, temporally across different sports), the innovations in them share similar support such as athletes competing, crowds cheering and so on, and at the same time they are sparse since a large portion within each sport may portray repetitive content such as running/swimming patterns. JSM-3 deals with the case where the common component is not sparse while the innovations are sparse. This could correspond to a video covering the daily routine of a mailman. While the places he travels to deliver mails will have large variations (non-sparse common component), the action he performs at those places will be mostly similar with few innovations (interacting with customer, delivering mail in the mailbox or dropping off near the front door etc).

In pursuit of such joint sparse information contained in spatial and/or temporal ensembles of a video, we begin with the d-dimensional features extracted from segments V_i^j and use the recovery algorithms[2] presented in [2] to obtain the intermediate representation $J(V_i)$. We used all three JSM's since we are dealing with unconstrained videos and any of these models could be representative of the activities occurring in different video segments. Hence for JSM-1 and JSM-3, we obtain 1 common and four innovations each for ensembles V_i^s and V_i^t, and 1 common and 16 innovations for the ensemble V_i^{st}, whereas for JSM-2 we obtain the same number of innovations as mentioned above but without any common component. Each of these components are of d dimensions. The collection of 6 common components C_i and 72 innovative components I_i represent the

[2]The algorithmic details from [2] are provided in the supplementary material.

joint sparse representation $J(V_i) \in \mathbb{R}^{d \times 78}$ of the video V_i. In the following we refer to C_i and I_i as joint sparse *atoms* of a video V_i, and let $\bar{C}$ and $\bar{I}$ refer to collection of such atoms obtained from all videos in the system.

2.3. Modeling Jointly Sparse Atoms

We now perform activity analysis by modeling information contained in $\bar{C}$ and $\bar{I}$. While the focus of [2] was to derive a linear combination of the common component (if available, depending on the JSM) and an innovation component to represent each segment V_i^j of a signal ensemble, our first goal is to extrapolate interactions between C_i and I_i to obtain different possible descriptions of activities contained in a video V_i. Towards this end we consider the subspace S_i spanned by the columns of (orthonormalized) matrix $J(V_i)$, which includes the set of all linear combinations of the joint sparse atoms from video V_i.

The problem of performing activity analysis then translates to that of 'comparing' subspaces $\mathcal{S} = \{S_i\}_{i=1}^N$ corresponding to all videos in the system, for which we pursue a geometrically meaningful Grassmann manifold interpretation[3] of the space spanned by $\mathcal{S}$. The Grassmannian $\mathcal{G}_{n,d}$, a nonlinear analytical manifold, is the space of all n-dimensional subspaces in $\mathbb{R}^d$ (i.e. column span of $d \times n$ orthonormal matrices) containing the origin, with $d > n$. S_i maps onto a 'point' in $\mathcal{G}_{n,d}$. There have been several works addressing geometric properties [10] and related statistical techniques [7] on this manifold, and we now utilize some of these tools f for analyzing the point cloud $\mathcal{S}$ to facilitate both activity classification and clustering.

2.3.1 Activity Classification

We first consider the case where each video V_i is accompanied by an activity label $l_i \in \{1, 2..., m\}$. Let $\tilde{V}$ denote the test video whose activity label $\tilde{l}$ is to be inferred. We pursue two statistical techniques for this supervised scenario namely, intrinsic ($f_1 \in f$) and extrinsic ($f_2 \in f$). While the extrinsic methods embed nonlinear manifolds in higher dimensional Euclidean spaces and perform computations in those larger spaces, the intrinsic methods are completely restricted to the manifolds themselves and do not rely on any Euclidean embedding.

We first pursue the intrinsic method of [40] that learns parametric class conditional densities pertaining to the labeled point cloud $\mathcal{S}$. We outline the general methodology of this technique below. For the set of points $\mathcal{S}_i \subset \mathcal{S}$ corresponding to ith activity label, we first estimate its 'mean' M_i using the Karcher mean algorithm [18]. Then we define a 'tangent plane' at the mean, which is a locally Euclidean representation of the non-linear space around M_i. We then map $\mathcal{S}_i$ onto the tangent plane and fit a Gaussian distribution to these points to obtain the mean μ_i and covariance Σ_i of the Gaussian. The class conditional $\mathcal{C}_i$ for the ith activity class is then completely represented by the tuple $\mathcal{C}_i = \{M_i, \mu_i, \Sigma_i\}$. The same process is repeated for points in $\mathcal{S}$ corresponding to all m activity labels. The activity label $\tilde{l}$ of a test point $\tilde{S}$ (column span of the orthonormalized version of $J(\tilde{V})$ derived from test video $\tilde{V}$) is determined by evaluating the m class conditional densities at the test point, by mapping $\tilde{S}$ to the tangent plane defined at each of the means M_i's, and then selecting the one with maximum likelihood (f_1).

Next we pursue an extrinsic method proposed by [14] that performs kernel discriminant analysis on the labeled point cloud $\mathcal{S}$ using a projection kernel k_P, which is a positive definite kernel well-defined for points on $\mathcal{G}_{n,d}$. More specifically, given a pair of $d \times n$ orthonormal matrices $\bar{J}(V_1)$ and $\bar{J}(V_2)$ obtained from a compact singular value decomposition of $J(V_1)$ and $J(V_2)$ respectively, the Mercer kernel $k_P(\bar{J}(V_1), \bar{J}(V_2)) = ||\bar{J}(V_1)^T \bar{J}(V_2)||_F^2 = trace[(\bar{J}(V_1)\bar{J}(V_1)^T)(\bar{J}(V_2)\bar{J}(V_2)^T)]$ implicitly computes the inner product between $\bar{J}(V_i)$'s in the space obtained using the embedding $\omega_P : \mathcal{G}_{n,d} \rightarrow \mathbb{R}^{d \times d}, span(\bar{J}(V_i)) \rightarrow \bar{J}(V_i)\bar{J}(V_i)^T$. In the above, the superscript T denotes matrix transpose, and $||.||_F$ denotes Frobenius norm. We then use k_P to create kernel matrices from training and test data, and perform test video activity classification f_2 in the standard discriminant analysis framework. We provide more details of this method in the supplementary material.

2.3.2 Activity Clustering

We then consider cases where the videos $\mathcal{V}$ do not have activity labels associated to them. In such cases we perform clustering ($f_3 \in f$) on the unlabeled point cloud $\mathcal{S}$. We pursue k-means [40] where given a set of points $\mathcal{S} = (S_1, S_2, ..., S_N)$ on $\mathcal{G}_{n,d}$, we seek to estimate k clusters $\mathbb{C} = (\bar{c}_1, \bar{c}_2, ..., \bar{c}_k)$ with cluster centers $(\bar{\mu}_1, \bar{\mu}_2, ..., \bar{\mu}_k)$ so that the sum of geodesic-distance squares, $\sum_{i=1}^{k} \sum_{S_j \in \bar{c}_i} d_G^2(S_j, \bar{\mu}_i)$ is minimized. d_G can be computed using the arc length metric d_{arc} [10]. Given a pair of points S_1 and S_2 on $\mathcal{G}_{n,d}$, $d_G(S_1, S_2) = d_{arc}^2(S_1, S_2) = \sum_{i=1}^{n} \theta_i^2$ is a function of principal angles θ_i between the two subspaces spanned by the columns of $d \times n$ orthonormal matrices $\bar{J}(V_1)$ and $\bar{J}(V_2)$ respectively. As is the case with standard (Euclidean) k-means, we can solve this problem using an EM-based approach. We initialize the algorithm with a random selection

[3] While there has been some work [46] on performing compressed sensing on Grassmann manifolds, they focus on obtaining sharp bounds for *recovering* approximate sparse signals using null-space Grassmann-angle characterization, whereas the focus of this work is on *representing* the column space of *jointly sparse* atoms and performing subsequent video analysis using Grassmannian tools.

of k points as the cluster centers. In the E-step, we assign each of the points of the dataset $\mathcal{S}$ to the nearest cluster center. Then in the M-step, we recompute the cluster centers using the Karcher mean computation algorithm described in the supplementary material.

In our experiments, we hide the activity labels of the videos $\mathcal{V}$ in the dataset and obtain their grouping f_3. We then evaluate the clustering accuracy with the widely used method of [45], which labels each of the resulting clusters with the majority activity class according to the original ground truth labels l_i, and finally measures the number of misclassifications in all clusters.

3. Experiments

We now evaluate our approach on four standard activity analysis datasets. We first experiment with the UCF-50 dataset [1] that consists of real-world videos taken from YouTube. We then consider the Human Motion DataBase (HMDB) [21] which is argued to be more challenging that UCF-50 since it contains videos from multiple sources such as YouTube, motion pictures etc. We then focus on the Olympic Sports dataset [28] that contains several sports clips which makes it interesting to analyze the performance of the method within a specific theme of activities. Finally we perform evaluations on the KTH dataset [37] which, though older and has more constrained actions, provides a benchmark on which many techniques have reported results. Figure 3 provides a sample of activity classes from these datasets. Before getting into the results, we discuss some design issues involved in our approach.

3.1. Feature Extraction

Our basic feature representation of the video segments $V_i^j, \forall j = 1\ to\ 24, \forall i = 1\ to\ N$ is a $d-$dimensional histogram pertaining to spatio-temporal bag-of-features obtained using the method of [22], which has shown good empirical performance on many datasets. We followed the protocol of [22] and constructed a 4000-size codebook ($d = 4000$) based on histogram of oriented gradient (HOG) descriptors. Although our method is independent of choice of features (with the constraint $d > n$), we used this feature for all four datasets to study the generalizability of the approach. We then obtain the intermediate joint sparse representation $J(V_i) \in \mathbb{R}^{4000\times 78}$ for all videos in the system with which the subsequent modeling (f_1, f_2, f_3) is done on $\mathcal{G}_{78,4000}$.

3.2. UCF-50 Dataset

A recent real-world dataset is the UCF-50 [1] that has 50 activity classes with atleast 100 videos per class. These videos taken from YouTube has a range of activities from general sports to daily-life exercises, and there are 6618 videos in total. Each activity class is divided into 25 homogenous groups with atleast 4 videos per activity in each of these groups. The videos in the same group may share some common features, such as the same person, similar background or similar viewpoint. We evaluate our intrinsic f_1 and extrinsic f_2 modeling strategies using the Leave-one-Group-out (LoGo) cross-validation scheme suggested in the dataset website and report the average classification accuracy across all activity classes in Table 1. To enable comparison with some existing approaches we also conducted a 10-fold cross validation test, where each time 9 random groups out of the 25 were used as test and the remaining 16 as training, and report the average classification accuracy in Table 1. For the clustering experiment, we report clustering results in two settings: (i) Case-A where only the test videos used in the classification experiment are clustered, and (ii) Case-B where both training and test videos in the classification experiment are clustered. We ran the clustering algorithm f_3 ten times, and the average clustering accuracy (for LoGo setting, but without the activity labels) is as follows, Case-A: 53.41%, Case-B: 57.8%

Method	Classification Accuracy (%)	
	Splits	LoGo
Laptev et al. [22]	47.9	-
Sadanand & Corso [35]	57.9	-
Kliper-Gross et al. [19]	68.51	72.68
Wang et al. [43]	-	85.6
Ours - intrinsic (f_1)	73.46	84.17
Ours - extrinsic (f_2)	71.24	80.63

Table 1. Performance comparison on UCF-50 dataset [1].

3.3. HMDB

Another recent dataset is the HumanMotion DataBase (HMDB) [21] that has 51 distinct activity classes with atleast 101 videos per class. The total 6766 videos were extracted from a wide range of sources, including YouTube and motion pictures, and has 10 overlapping activity classes with the UCF-50 dataset. Each video was validated by atleast two human observers to ensure consistency. The evaluation protocol for this dataset consists of three distinct training and testing splits, each containing 70 training and 30 testing videos per activity class, and the splits are selected in such way to display a representative mix of video quality and camera motion attributes. The dataset contains both original videos and their stabilized version, and we report classification results on both of these sets in Table 2. Our clustering algorithm got the following accuracy, Case-A: 12.41% (original), 13.2% (stabilized) and Case-B: 14.66% (original), 18.72% (stabilized).

Figure 3. Sample frames corresponding to different activity classes in datasets - UCF-50 (first row), HMDB (second), KTH (third), and Olympic Sports (last). These datasets have a good mix of realistic activities from YouTube, motion pictures, sports programs etc.

Method	Classification Accuracy (%)	
	Original videos	Stabilized videos
Laptev et al. [22]	20.44	21.96
Jhuang et al. [17]	22.83	23.18
Sadanand & Corso [35]	26.90	-
Cao et al. [5]	-	27.84
Kliper-Gross et al. [19]	29.17	-
Ours - intrinsic (f_1)	34.13	37.5
Ours - extrinsic (f_2)	32.89	35.42

Table 2. Performance comparison on HMDB [21].

3.4. Olympic Sports Dataset

The Olympic Sports dataset [28] consists of athletes practicing different sports, which are collected from YouTube and annotated using Amazon Mechanical Turk. There are 16 sports events: high-jump, long-jump, triple-jump, pole-vault, basketball lay-up, bowling, tennis-serve, platform, discus, hammer, javelin, shot-put, springboard, snatch, clean-jerk and vault, represented by a total of 783 video sequences. This dataset has rich scene context information, which is very helpful for recognizing sports actions, and we used the training/testing split provided in the dataset for evaluation. The mean average precision over all activity classes is reported in Table 3. The clustering accuracy for this dataset is: Case-A: 59.6%, Case-B: 64.2%

3.5. KTH Dataset

The KTH dataset [37] is a relatively old dataset comprising of six human action classes: walking, jogging, running, boxing, waving and clapping. Each action is performed several times by 25 subjects. The sequences were recorded in four different scenarios: outdoors, outdoors with scale variation, outdoors with different clothes and indoors. The background is homogeneous and static in most sequences. We follow the original experimental setup [37] by dividing the samples into test set with 9 subjects and training set with 16 subjects. Average classification accuracy over all action classes is given in Table 4 and the clustering results for this dataset are Case-A: 71.4%, Case-B: 72.8%.

Method	Classification Accuracy (%)
Niebles et al. [28]	72.1
Brendel et al. [4]	77.3
Wang et al. [43]	77.2
Ours - intrinsic (f_1)	78.61
Ours - extrinsic (f_2)	77.12

Table 3. Performance comparison on Olympic Sports dataset [28].

Method	Classification Accuracy (%)
Gilbert et al. [12]	94.5
Kovashka & Grauman [20]	94.5
Wu et al. [44]	94.5
Sadanand & Corso [35]	98.2
Ours - intrinsic (f_1)	97.31
Ours - extrinsic (f_2)	97.2

Table 4. Performance comparison on KTH dataset [37].

3.6. Discussion

We now empirically analyze the utility of modeling using joint sparsity principles. Given d-dimensional bag-of-feature histograms for each of the 24 segments correspond-

Figure 4. Sample classification results on HMDB (dive, run) and UCF-50 (biking, golfswing) datasets. Top three videos classified into each of these classes are displayed here, with a representative frame corresponding those videos, where the first row pertains to results using joint sparse modeling and the second row to that of PCA modeling. Mis-classification results are given within a red bounding box.

ing to a video V_i, instead of extracting jointly sparse atoms $J(V_i)$, we perform principal component analysis (PCA) to obtain an orthonormal matrix using which we perform subsequent computations for activity classification and clustering (Section 2.3). As before, the number of columns of this matrix must be less than d, and we varied it such that the matrix contains atleast 80%, and upto 99% of the original energy. The highest results obtained using this method (with intrinsic classification f_1) for each of the four datasets are as follows, UCF-50 - 62.34% (splits), 75.6% (LoGo), HMDB - 22.58% (original videos), 31.45% (stabilized videos), Olympic Sports - 76.1%, and KTH - 95.38%. We see that the joint sparse modeling yields better results, and some illustrations on activity (mis-)classification are shown in Figure 4.

From these results we make the following observations. (i) The classification performance our approach is comparable to, and in many cases better than, the existing methods. Given the amount of variation in activity patterns across the datasets, these results demonstrate the generalizability of our joint sparsity based manifold modeling. (ii) Intrinsic classification outperforms extrinsic on all cases. This sheds light on the advantages of learning class-specific distributions rather than focusing on just the discriminative information. One pitfall however is that the intrinsic method is computationally expensive than the extrinsic approach. For a 100 frame test video $\tilde{V}$, it takes about 10 seconds to obtain its intermediate joint sparse representation $J(\tilde{V})$. Then to infer its activity label it takes around 8 seconds using the intrinsic method f_1 and 3 seconds using the extrinsic algorithm f_2. All these computational times correspond to a 2GHz processor with 2 GB of RAM. (iii) Clustering performance is inferior to that of classification, which reinforces the conventional wisdom of advantages provided by labels (or supervision). An interesting observation is that clustering under Case B is better than Case A. This shows that more data, even if it is unlabeled, helps and our modeling process is able to extract meaningful information from the data.

4. Conclusion

Through this work we studied the utility of joint sparsity models for representing common and diverse content within a video and a subsequent manifold interpretation for performing video analysis under both supervised and unsupervised settings. We demonstrated the generalizability of the approach across several video activity datasets that portrayed varying degree of event complexity, without resorting to feature tuning, and achieved competitive results on many counts. It is an interesting future work to explore integrating feature selection mechanisms with our model and study their utility for more general video understanding problems.

References

[1] Ucf-50. http://server.cs.ucf.edu/ vision/data/ucf50.rar. 1, 5

[2] D. Baron, M. Duarte, M. Wakin, S. Sarvotham, and R. Baraniuk. Distributed compressive sensing. *arXiv preprint arXiv:0901.3403*, 2009. 1, 2, 3, 4

[3] A. Bobick and J. Davis. The recognition of human movement using temporal templates. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 23(3):257–267, 2001. 1

[4] W. Brendel and S. Todorovic. Learning spatiotemporal graphs of human activities. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 778–785. IEEE, 2011. 6

[5] L. Cao, Y. Mu, A. Natsev, S. Chang, G. Hua, and J. Smith. Scene aligned pooling for complex video recognition. In *ECCV*, 2012. 1, 2, 6

[6] Y. Chen, N. Nasrabadi, and T. Tran. Simultaneous joint sparsity model for target detection in hyperspectral imagery. *Geoscience and Remote Sensing Letters, IEEE*, 8(4):676–680, 2011. 2

[7] Y. Chikuse. *Statistics on special manifolds*. Springer Verlag, 2003. 4

[8] L. Duan, D. Xu, and S. Chang. Exploiting web images for event recognition in consumer videos: A multiple source domain adaptation approach. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 1338–1345. IEEE, 2012. 2

[9] L. Duan, D. Xu, I. Tsang, and J. Luo. Visual event recognition in videos by learning from web data. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(9):1667–1680, 2012. 2

[10] A. Edelman, T. Arias, and S. Smith. The geometry of algorithms with orthogonality constraints. *SIAM Journal on Matrix Analysis and Applications*, 20(2):303–353, 1998. 4

[11] Y. Fu, T. Hospedales, T. Xiang, and S. Gong. Attribute learning for understanding unstructured social activity. In *European Conference on Computer Vision*, 2012. 1, 2

[12] A. Gilbert, J. Illingworth, and R. Bowden. Fast realistic multi-action recognition using mined dense spatio-temporal features. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 925–931. IEEE, 2009. 6

[13] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29(12):2247–2253, 2007. 1

[14] J. Hamm and D. Lee. Grassmann discriminant analysis: a unifying view on subspace-based learning. In *Proceedings of the 25th international conference on Machine learning*, pages 376–383. ACM, 2008. 4

[15] Z. Han, Z. Xu, and S. Zhu. Video primal sketch: A generic middle-level representation of video. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 1283–1290. IEEE, 2011. 2

[16] M. Hoai and F. De la Torre. Max-margin early event detectors. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2863–2870. IEEE, 2012. 2

[17] H. Jhuang, T. Serre, L. Wolf, and T. Poggio. A biologically inspired system for action recognition. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8. IEEE, 2007. 6

[18] H. Karcher. Riemannian center of mass and mollifier smoothing. *Communications on pure and applied mathematics*, 30(5):509–541, 1977. 4

[19] O. Kliper-Gross, Y. Gurovich, T. Hassner, and L. Wolf. Motion interchange patterns for action recognition in unconstrained videos. In *ECCV*, 2012. 2, 5, 6

[20] A. Kovashka and K. Grauman. Learning a hierarchy of discriminative space-time neighborhood features for human action recognition. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2046–2053. IEEE, 2010. 1, 2, 6

[21] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. HMDB: a large video database for human motion recognition. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2011. 1, 5, 6

[22] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008. 1, 2, 3, 5, 6

[23] R. Li and T. Zickler. Discriminative virtual views for cross-view action recognition. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2855–2862. IEEE, 2012. 2

[24] J. Liu, B. Kuipers, and S. Savarese. Recognizing human actions by attributes. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 3337–3344. IEEE, 2011. 1

[25] J. Liu, J. Luo, and M. Shah. Recognizing realistic actions from videos in the wild. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1996–2003. IEEE, 2009. 2

[26] J. Liu, S. McCloskey, and Y. Liu. Local expert forest of score fusion for video event classification. *Computer Vision–ECCV 2012*, pages 397–410, 2012. 1, 2

[27] P. Nagesh and B. Li. A compressive sensing approach for expression-invariant face recognition. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1518–1525. IEEE, 2009. 2

[28] J. Niebles, C. Chen, and L. Fei-Fei. Modeling temporal structure of decomposable motion segments for activity classification. *Computer Vision–ECCV 2010*, pages 392–405, 2010. 5, 6

[29] S. Oh, A. Hoogs, A. Perera, N. Cuntoor, C. Chen, J. Lee, S. Mukherjee, J. Aggarwal, H. Lee, L. Davis, et al. A large-scale benchmark dataset for event recognition in surveillance video. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 3153–3160. IEEE, 2011. 1, 2

[30] H. Pirsiavash and D. Ramanan. Detecting activities of daily living in first-person camera views. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2847–2854. IEEE, 2012. 2

[31] Q. Qiu, Z. Jiang, and R. Chellappa. Sparse dictionary-based representation and recognition of action attributes. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 707–714. IEEE, 2011. 2

[32] C. Rao and M. Shah. View-invariance in action recognition. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, volume 2, pages II–316. IEEE, 2001. 1

[33] M. Raptis, I. Kokkinos, and S. Soatto. Discovering discriminative action parts from mid-level video representations. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 1242–1249. IEEE, 2012. 2

[34] M. Rohrbach, M. Regneri, M. Andriluka, S. Amin, M. Pinkal, and B. Schiele. Script data for attribute-based recognition of composite activities. *Computer Vision–ECCV 2012*, pages 144–157, 2012. 2

[35] S. Sadanand and J. Corso. Action bank: A high-level representation of activity in video. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 1234–1241. IEEE, 2012. 2, 5, 6

[36] S. Savarese, A. DelPozo, J. Niebles, and L. Fei-Fei. Spatial-temporal correlatons for unsupervised action classification. In *Motion and video Computing, 2008. WMVC 2008. IEEE Workshop on*, pages 1–8. IEEE, 2008. 1

[37] 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. IEEE, 2004. 1, 5, 6

[38] A. Tamrakar, S. Ali, Q. Yu, J. Liu, O. Javed, A. Divakaran, H. Cheng, and H. Sawhney. Evaluation of low-level features and their combinations for complex event detection in open source videos. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 3681–3688. IEEE, 2012. 2

[39] P. Turaga, R. Chellappa, V. Subrahmanian, and O. Udrea. Machine recognition of human activities: A survey. *Circuits and Systems for Video Technology, IEEE Transactions on*, 18(11):1473–1488, 2008. 1, 2

[40] P. Turaga, A. Veeraraghavan, A. Srivastava, and R. Chellappa. Statistical computations on grassmann and stiefel manifolds for image and video-based recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(11):2273–2286, 2011. 4

[41] S. Vitaladevuni, P. Natarajan, and R. Prasad. Efficient orthogonal matching pursuit using sparse random projections for scene and video classification. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 2312–2319. IEEE, 2011. 2

[42] H. Wang, A. Klaser, C. Schmid, and C. Liu. Action recognition by dense trajectories. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 3169–3176. IEEE, 2011. 1, 2

[43] H. Wang, A. Kläser, C. Schmid, C. Liu, et al. Dense trajectories and motion boundary descriptors for action recognition. In *Research report, INRIA*, 2012. 5, 6

[44] X. Wu, D. Xu, L. Duan, and J. Luo. Action recognition using context and appearance distribution features. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 489–496. IEEE, 2011. 6

[45] L. Xu, J. Neufeld, B. Larson, and D. Schuurmans. Maximum margin clustering. *Advances in neural information processing systems*, 17:1537–1544, 2004. 5

[46] W. Xu and B. Hassibi. Compressed sensing over the grassmann manifold: A unified analytical framework. In *Communication, Control, and Computing, 2008 46th Annual Allerton Conference on*, pages 562–567. IEEE, 2008. 4

[47] H. Yin and S. Li. Multimodal image fusion with joint sparsity model. *Optical Engineering*, 50(6):067007–067007, 2011. 2

[48] X. Yuan and S. Yan. Visual classification with multi-task joint sparse representation. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 3493–3500. IEEE, 2010. 2

[49] Q. Zhang and B. Li. Joint sparsity model with matrix completion for an ensemble of face images. In *Image Processing (ICIP), 2010 17th IEEE International Conference on*, pages 1665–1668. IEEE, 2010. 2

[50] B. Zhou, X. Wang, and X. Tang. Understanding collective crowd behaviors: Learning a mixture model of dynamic pedestrian-agents. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2871–2878. IEEE, 2012. 2