Sparse Dictionary-based Representation and Recognition of Action Attributes

Qiang Qiu, Zhuolin Jiang, Rama Chellappa
Center for Automation Research, UMIACS
University of Maryland, College Park, MD 20742
qiu@cs.umd.edu,{zhuolin, rama}@umiacs.umd.edu

Abstract

We present an approach for dictionary learning of action attributes via information maximization. We unify the class distribution and appearance information into an objective function for learning a sparse dictionary of action attributes. The objective function maximizes the mutual information between what has been learned and what remains to be learned in terms of appearance information and class distribution for each dictionary item. We propose a Gaussian Process (GP) model for sparse representation to optimize the dictionary objective function. The sparse coding property allows a kernel with a compact support in GP to realize a very efficient dictionary learning process. Hence we can describe an action video by a set of compact and discriminative action attributes. More importantly, we can recognize modeled action categories in a sparse feature space, which can be generalized to unseen and unmodeled action categories. Experimental results demonstrate the effectiveness of our approach in action recognition applications.

1. Introduction

Describing human actions using attributes is closely related to representing an object using attributes [8]. Several studies have investigated the attribute-based approaches for object recognition problems [12, 9, 8, 21, 30]. These methods have demonstrated that attribute-based approaches can not only recognize object categories, but can also describe unknown object categories. In this paper, we propose a dictionary-based approach for learning human action attributes which are useful to model and recognize known action categories, and also describe unknown action categories.

Dictionary learning is an approach to learn attributes (i.e., dictionary items) from a set of training samples. In [1], a promising dictionary learning algorithm, K-SVD, is introduced to learn an over-complete dictionary. Input signals can then be represented as a sparse linear combination of dictionary items. K-SVD only focuses on minimizing the reconstruction error. Discriminative K-SVD in [31] extends K-SVD by incorporating the classification error into the objective function to obtain a more discriminative dictionary.

In this paper, we propose an approach for dictionary learning of human action attributes via information maximization. In addition to using the appearance information between dictionary items, we also exploit the class label information associated with dictionary items to learn a compact and discriminative dictionary for human action attributes. The mutual information for appearance information and class distributions between the learned dictionary and the rest of the dictionary space are used to define the objective function, which is optimized using a Gaussian Process (GP) model [22] proposed for sparse representation. The property of sparse coding naturally leads to a GP kernel with compact support, i.e., zero values for a most portion, for significant speed-ups. The representation and recognition of actions are through sparse coefficients related to learned attributes. A compact and discriminative attribute dictionary should encourage the signals from the same class to have very similar sparse representations. In other words, the signals from the same class are described by a similar set of dictionary items. As shown in Fig. 1, our approach produces consistent sparse representations for the same class signals. The main contributions of this paper are:

- We propose a novel probabilistic model for sparse representation.
- We learn a compact and discriminative dictionary for sparse coding via information maximization.
- We describe and recognize human actions, including unknown actions, via a set of human action attributes in a sparse feature space.

1.1. Related Work

Discriminative dictionary learning is gaining widespread attention in many disciplines. Some examples include LDA-based basis selection [7], combined dictionary learning and classifier training [31, 28, 20], distance matrix learning [2], hierarchical pairwise merging of visual words [26], maximization of mutual information (MMI) [14, 24, 17], and sparse coding-based dictionary learning [18, 19, 10].

Recent dictionary-based approaches for learning action attributes include agglomerative clustering [25], forward se-

2011 IEEE International Conference on Computer Vision
978-1-4577-1102-2/11/$26.00

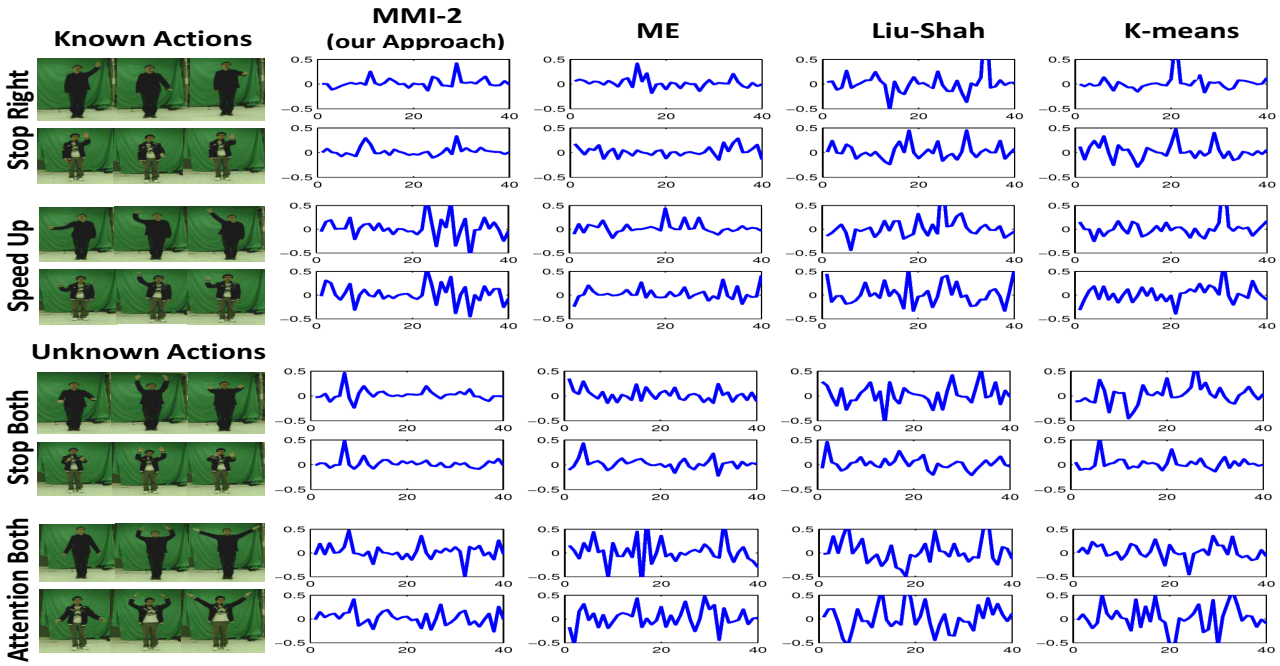


Figure 1: Sparse representations of four actions (two are known and two are unknown to the attribute dictionary) using attribute dictionaries learned by different methods. Each action is performed by two different humans. For visualization purpose, each waveform shows the average of the sparse codes of all frames in an action sequence. We learned attribute dictionaries using several methods including our approach, the Maximization of Entropy approach (ME), the *Liu-Shah* approach [17] and the K-means approach. A compact and discriminative attribute dictionary should encourage actions from the same class to be described by a similar set of attributes, i.e., similar sparse codes. The attribute dictionary learned by our approach provides similar waveforms, which shows consistent sparse representations, for the same class action sequences.

lection [27] and probabilistic graphical model [6]. [15] proposes an unsupervised approach and uses l_1 minimization to find basic primitives to represent human motions.

Our approach adopts the rule of Maximization of Mutual Information to obtain a compact and discriminative dictionary. The dictionary items are considered as attributes in our approach. Compared to previous methods, our approach maximizes the mutual information for both the appearance information and class distribution of dictionary items to learn a dictionary while [24] and [17] only maximize the mutual information for class distribution. Thus, we can expect improved dictionary compactness from our approach. Both [24] and [17] obtain a dictionary through merging of two visual words, which can be time-consuming when the dictionary size is large. Besides, our approach is efficient because the dictionary is learned in the sparse feature space so we can leverage the property of sparse coding to use kernel locality for speeding up the dictionary learning process.

2. Action Features and Attributes

Human action features are extracted from an action interest region for representing and describing actions. The action interest region is defined as a bounded region around the human performing the activity, which is obtained using background subtraction and/or tracking.

2.1. Basic Features

The human action attributes require feature descriptors to represent visual aspects. We introduce basic features, including both local and global features, used in our work.

Global Features: Global features encode rich information from an action interest region, so they generally perform better than local features in recognition. When cameras and backgrounds are static, we use the silhouette-based feature descriptor presented in [16] to capture shape information, while we use the Histogram of Oriented Gradient (HOG) descriptors used in [4] for dynamic backgrounds and moving cameras. For encoding motion information, we use the optical-flow based feature descriptors as in [5].

Local Features: Spatio-temporal local features describe a video as a collection of independent patches or 3D cuboids, which are less sensitive to viewpoint changes, noise and partial occlusion. We first extract a collection of space-time interest points (STIP) introduced in [13] to represent an action sequence, and then use HOG and histogram of flow to describe them.

2.2. Human Action Attributes

Motivated by [25, 27, 6], an action can be represented as a set of basic action units. We refer to these basic action units as human action attributes. In order to effectively describe human actions, we need to learn a representative and semantic set of action attributes. Given all the base features from the training data, we aim to learn a compact and discriminative dictionary where all the dictionary items can be used as human action attributes. The final learned dictionary can be used as a "Thesaurus" of human action attributes.

3. Learning Attribute Dictionary

We first obtain an initial dictionary via K-SVD [1]. Then we learn a compact and discriminative dictionary from the initial dictionary via information maximization.

3.1. Dictionary Initialization

As the optimal dictionary size is rarely known in advance, we obtain through K-SVD an initial dictionary D^o of a large size K. K-SVD [1] is a method to learn an over-complete dictionary for sparse coding. Let Y be a set of N input signals in a n-dimensional feature space $Y = [y_1...y_N],\ y_i \in \mathbb{R}^n$. In K-SVD, a dictionary with a fixed number of K items is learned by finding a solution iteratively to the following problem:

$$\arg\min_{D,X} \|Y - DX\|_2^2 \ \ s.t. \forall i, \|x_i\|_0 \leq T \qquad (1)$$

where $D = [d_1...d_K],\ d_i \in \mathbb{R}^n\ (K > n)$ is the learned dictionary, $X = [x_1, ..., x_N],\ x_i \in \mathbb{R}^K$ are the sparse codes of input signals Y, and T specifies the sparsity that each signal has fewer than T items in its decomposition. Each dictionary item d_i is l_2-normalized. The initial dictionary D^o from (1) only minimizes the reconstruction error, and is not optimal in terms of compactness and discriminability.

3.2. Probabilistic Model for Sparse Representation

Before we present our dictionary learning framework, we first suggest a novel probabilistic model for sparse representation motivated by [11].

3.2.1 A Gaussian Process

Given a set of input signals Y, $Y = [y_1...y_N], y_i \in \mathbb{R}^n$, there exists an infinite dictionary space $\mathcal{D} \subseteq \mathbb{R}^n$. Each dictionary item $d_i \in \mathcal{D}$ maps the set of input signals to its corresponding sparse coefficients $x_{d_i} = [x_{i,1}...x_{i,N}]$ in X, which can be viewed as its observations to the set of input signals. When two dictionary items d_i and d_j are similar, it is more likely that input signals will use them simultaneously in their sparse decomposition [21]. Thus the similarity of two dictionary items can be assessed by the correlation between their observations (i.e., sparse coefficients). Such correlation properties of sparse coefficients has been used in [21] to cluster dictionary items.

With the above formulation, we obtain a problem which is commonly modeled as a *Gaussian Process* (GP). A GP is specified by a mean function and a symmetric positive-definite covariance function $\mathcal{K}$. Since we simplify our problem by assuming an initial dictionary D^o, we only need to specify entries in the covariance function $\mathcal{K}$ for items existing in D^o, and leave the rest undefined. For each pair of dictionary items $\forall d_i, d_j \in D^o$, we define the corresponding covariance function entry $\mathcal{K}(i,j)$ as the covariance between their associated sparse coefficients $cov(x_{d_i}, x_{d_j})$. For simplicity, we use the notation $\mathcal{K}_{(d_i,d_j)}$ to refer to the covariance entry at indices d_i, d_j. Similarly, we use $\mathcal{K}_{(D^*,D^*)}$ to denote the covariance matrix for a set of dictionary items D^*.

The GP model for sparse representation gives us the following useful property: given a set of dictionary items D^* and the associated sparse coefficients X_{D^*}, the distribution $P(X_{d^*}|X_{D^*})$ at any given testing dictionary item d^* is a Gaussian with a closed-form conditional variance [22].

$$\mathbb{V}(d^*|D^*) = \mathcal{K}_{(d^*,d^*)} - \mathcal{K}^T_{(d^*,D^*)}\mathcal{K}^{-1}_{(D^*,D^*)}\mathcal{K}_{(d^*,D^*)} \qquad (2)$$

where $\mathcal{K}_{(d^*,D^*)}$ is the vector of covariances between d^* and each item in D^*.

3.2.2 Dictionary Class Distribution

When the set of input signals Y is labeled with one of M discrete class labels, we can further derive class related distributions over the sparse representation.

As mentioned, each dictionary item d_i maps the set of input signals to its corresponding sparse coefficients $x_{d_i} = [x_{i,1}...x_{i,N}]$ in X. Since each coefficient $x_{i,j}$ here corresponds to an input signal y_j, it is associated with a class label. If we aggregate x_{d_i} based on class labels, we obtain a M sized vector. After normalization, we have the conditional probability $P(L|d_i),\ L \in [1, M]$, where $P(L|d_i)$ represents the probability observing a class given a dictionary item.

3.3. Dictionary Learning for Attributes

Given the initial dictionary D^o obtained from (1), we aim to compress it into a dictionary D^* of size k, which encourages the signals from the same class to have very similar sparse representations, as shown in Fig. 1. In other words, the signals from the same class are described by a similar set of attributes, i.e., dictionary items. Therefore, a compact and discriminative dictionary is more desirable.

An intuitive heuristic is to start with $D^* = \emptyset$, and iteratively choose the next best item d^* from $D^o \backslash D^*$ which

provides a maximum increase for the entropy of D^*, i.e., $\arg\max_{d^*} H(d^*|D^*)$, until $|D^*| = k$, where $D^o \backslash D^*$ denotes the remaining dictionary items after D^* has been removed from the initial dictionary D^o. With our GP modeling, we can evaluate $H(d^*|D^*)$ as a closed-form Gaussian conditional entropy,

$$H(d^*|D^*) = \frac{1}{2} log(2\pi e \mathbb{V}(d^*|D^*)) \tag{3}$$

where $\mathbb{V}(d^*|D^*)$ is defined in (2). This heuristic is a good approximation to the *maximization of joint entropy* (ME) criteria, i.e., $\arg\max_{D^*} H(D^*)$.

With the ME rule, as items in the learned dictionary are less correlated to each other due to their high joint entropy, the learned dictionary is compact. However, the maximal entropy criteria will favor attributes associated with the beginning and the end of an action, as they are least correlated. Such a phenomenon is shown in Fig. 3b and Fig. 3d in the experiment section. Thus we will expect high reconstruction error and weak discriminability. To mitigate this in our dictionary learning framework, we adopt Maximization of Mutual Information (MMI) as the criteria for ensuring dictionary compactness and discriminability.

3.3.1 MMI for Unsupervised Learning (MMI-1)

The rule of maximization of entropy only considers the entropy of dictionary items. Instead we choose to learn D^* that most reduces the entropy about the rest of dictionary items $D^o \backslash D^*$.

$$\arg\max_{D^*} I(D^*; D^o \backslash D^*) \tag{4}$$

It is known that maximizing the above criteria is NP-complete. Fortunately, a similar problem has been studied in the machine learning literature [11]. We can use a very simple greedy algorithm here. We start with $D^* = \emptyset$, and iteratively choose the next best dictionary item d^* from $D^o \backslash D^*$ which provides a maximum increase in mutual information, i.e.,

$$\begin{aligned} \arg\max_{d^* \in D^o \backslash D^*} & I(D^* \cup d^*; D^o \backslash (D^* \cup d^*)) - I(D^*; D^o \backslash D^*) \\ &= H(d^*|D^*) - H(d^*|\bar{D}^*); \end{aligned} \tag{5}$$

where $\bar{D}^*$ denotes $D^o \backslash (D^* \cup d^*)$. Intuitively, the ME criteria only considers $H(d^*|D^*)$, i.e., force d^* to be most different from already selected dictionary items D^*, now we also consider $-H(d^*|\bar{D}^*)$ to force d^* to be most representative among the remaining items.

It has been proved in [11] that the above greedy algorithm serves a polynomial-time approximation that is within $(1 - 1/e)$ of the optimum. Based on similar arguments in [11], the near-optimality of our approach can be guaranteed if the initial dictionary size $|D^o|$ is sufficiently larger than $2|D^*|$.

With our proposed GP model, the objective function in (5) can be written in a closed form using (2) and (3).

$$\arg \max_{d^* \in D^o \backslash D^*} \frac{\mathcal{K}_{(d^*,d^*)} - \mathcal{K}^T_{(d^*,D^*)} \mathcal{K}^{-1}_{(D^*,D^*)} \mathcal{K}_{(d^*,D^*)}}{\mathcal{K}_{(d^*,d^*)} - \mathcal{K}^T_{(d^*,\bar{D}^*)} \mathcal{K}^{-1}_{(\bar{D}^*,\bar{D}^*)} \mathcal{K}_{(d^*,\bar{D}^*)}} \tag{6}$$

Given the initial dictionary size $|D^o| = K$, each iteration requires $\mathcal{O}(K^4)$ to evaluate (6). Such an algorithm seems to be computationally infeasible for any large initial dictionary size. The nice feature of this approach is that we model the covariance kernel $\mathcal{K}$ over sparse codes X, which entitles $\mathcal{K}$ a compact support, i.e., most portion of $\mathcal{K}$ has zero or very tiny value. After we ignore those zero valued entries while evaluating (6), the actual computation becomes very efficient.

3.3.2 MMI for Supervised Learning (MMI-2)

The objective functions in (4) and (5) only consider the appearance information of dictionary items, hence D^* is not optimized for classification. For example, attributes to distinguish a particular class can possibly be missing in D^*. So we need to use appearance information and class distribution to find a dictionary that also causes minimal loss information about labels.

Let L denote the labels of M discrete values, $L \in [1, M]$. In Sec. 3.2.2, we discussed how to obtain $P(L|d^*)$, which represents the probability of observing a class given a dictionary item. Give a set of dictionary item D^*, we define $P(L|D^*) = \frac{1}{|D^*|} \sum_{d_i \in D^*} P(L|d_i)$. For simplicity, we denote $P(L|d^*)$ as $P(L_{d^*})$, and $P(L|D^*)$ as $P(L_{D^*})$.

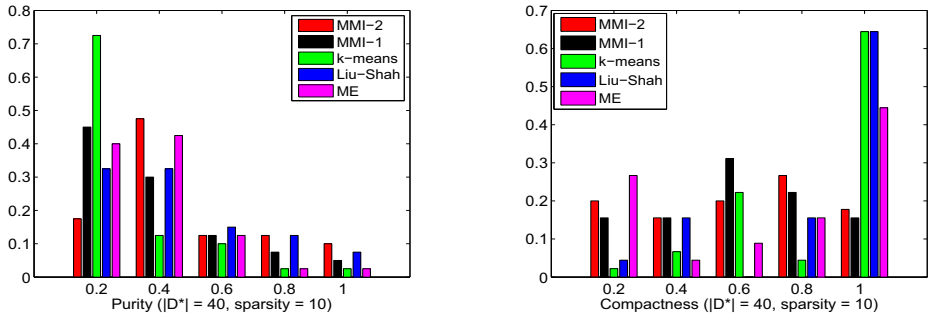
To enhance the discriminative power of the learned dictionary, we propose to modify the objective function (4) to

$$\arg\max_{D^*} I(D^*; D^o \backslash D^*) + \lambda I(L_{D^*}; L_{D^o \backslash D^*}) \tag{7}$$

where $\lambda \geq 0$ is the parameter to regularize the emphasis on appearance or label information. When we approximate (7) as

$$\begin{aligned} \arg \max_{d^* \in D^o \backslash D^*} & [H(d^*|D^*) - H(d^*|\bar{D}^*)] \\ & + \lambda [H(L_{d^*}|L_{D^*}) - H(L_{d^*}|L_{\bar{D}^*})] \end{aligned} \tag{8}$$

we can easily notice that we also force the classes associated with d^* to be most different from classes already covered by selected items D^*; and at the same time, the classes associated with d^* should be most representative among classes covered by the remaining items. Thus the learned dictionary is not only compact, but also covers all classes to maintain the discriminability. It is interesting to note that MMI-1 is a special case of MMI-2 with $\lambda = 0$.



(a) Purity (b) Compactness

Figure 2: Purity and compactness of learned dictionary D^*: purity is the histograms of the maximum probability observing a class given a dictionary item, and compactness is the histograms of $D^{*T}D^*$. At the right-most bin of the respective figures, a discriminative and compact dictionary should exhibit high purity and small compactness. MMI-2 dictionary is most "pure" and second most compact (MMI-1 is most compact but much less pure.)

4. Experimental Evaluation

This section presents an experimental evaluation on three public datasets: the Keck gesture dataset [16], the Weizmann action dataset [3] and the UCF sports action dataset [23]. On the Keck gesture dataset, we thoroughly evaluate the basic behavior of our two proposed dictionary learning approaches MMI-1 and MMI-2, in terms of dictionary compactness and discriminability, by comparing with other three alternatives. Then we further evaluate the discriminability of our learned action attributes over the popular Weizmann action dataset and the challenging UCF sports action dataset.

4.1. Comparison with Alternative Approaches

The Keck gesture dataset consists of 14 different gestures, which are a subset of the military signals. These 14 classes include turn left, turn right, attention left, attention right, flap, stop left, stop right, stop both, attention both, start, go back, close distance, speed up, come near. Each of the 14 gestures is performed by three subjects. Some sample frames from this dataset are shown in Fig. 1.

4.1.1 Alternative Methods for Comparison

For comparison purposes, in addition to the maximization of entropy (ME) method discussed before, we also implemented two additional action attributes learning approaches. The first approach is similar to the approach presented in [17] to obtain a compact and discriminative human action model. We revise it for sparse representation and refer to it as the *Liu-Shah* method.

Liu-Shah: In this approach, we start with an initial dictionary D^o obtained from K-SVD. At each iteration, for each pair of dictionary items, d_1 and d_2, we compute the MI loss if we merge these two into a new dictionary item d^*, and pick the pair which gives the minimum MI loss. We continue the merging process till the desired dictionary size. The MI loss is defined as,

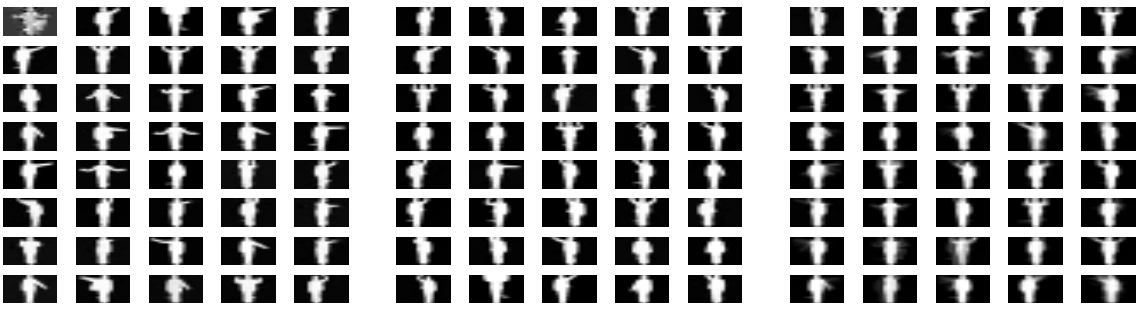
$$\begin{aligned}\triangle I(d_1,d_2) = & \sum_{L\in[1,M],i=1,2} p(d_i)p(L|d_i)\log p(L|d_i) \\ & -p(d_i)p(L|d_i)\log p(L|d^*) \end{aligned} \quad (9)$$

where $p(L|d^*) = \frac{p(d_1)}{p(d^*)}p(L|d_1) + \frac{p(d_2)}{p(d^*)}p(L|d_2)$ and $p(d^*) = p(d_1) + p(d_2)$

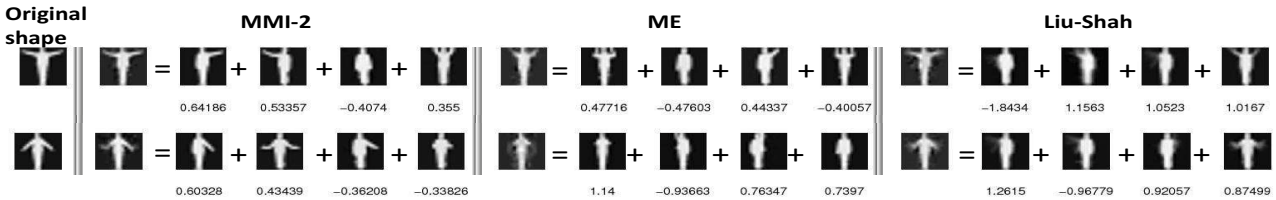
k-means: The second approach is to simply perform k-means over an initial dictionary D^o from K-SVD to obtain a desired size dictionary.

4.1.2 Dictionary Purity and Compactness

Through K-SVD, we start with an initial 500 size dictionary using the shape feature (sparsity 30 is used). We then learned a 40 size dictionary D^* from D^o using 5 different approaches. We let $\lambda = 1$ in (8) throughout the experiment. To evaluate the discriminability and compactness of these learned dictionaries, we evaluate the *purity* and *compactness* measures as shown in Fig. 2. The purity is assessed by the histograms of the maximum probability observing a class given a dictionary item, i.e., $max(P(L|d_i))$, and the compactness is assessed by the histograms of $D^{*T}D^*$. As each dictionary item is l_2-normalized, $d_i^t d_j \in [0,1]$ and indicates the similarity between the dictionary items d_i and d_j. Fig. 2a shows MMI-2 is most "pure", as around 25% of dictionary items learned by MMI-2 have 0.6-above probability to only associate with one of the classes. *Liu-Shah* shows comparable purity to MMI-2 as the MI loss criteria used in *Liu-Shah* does retain the class information during dictionary learning. However, as shown in Fig. 2b, MMI-2 dictionary is much more compact, as only about 20% MMI-2 dictionary items have 0.80-above similarity. As expected, comparing to MMI-2, MMI-1 shows better compactness but much less purity.

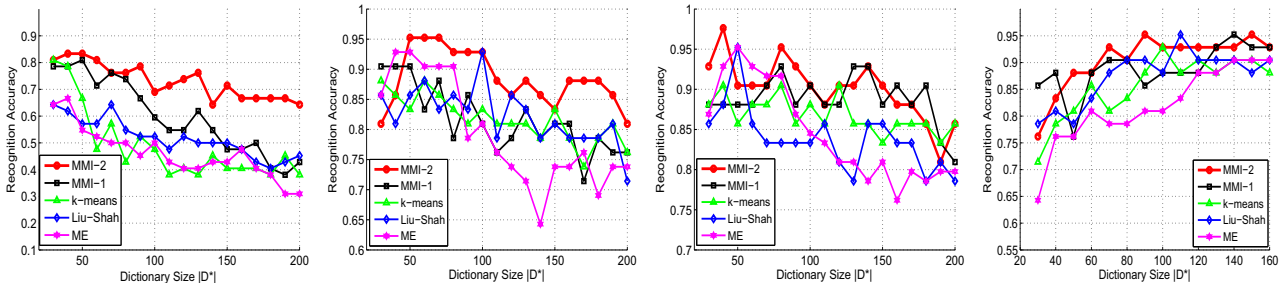


(a) MMI-2 shape attributes (b) ME shape attributes (c) *Liu-Shah* shape attributes



(d) Description to two example frames in an unknown action *flap* using attribute dictionaries (Sparsity 10 is used and top-4 attributes are shown.)

Figure 3: Learned attribute dictionaries on shape features ("unseen" classes: *flap*, *stop both* and *attention both*)



(a) Shape ($|D^o| = 600$) (b) Motion ($|D^o| = 600$) (c) Shape and Motion ($|D^o| = 1200$) (d) STIP ($|D^o| = 600$)

Figure 4: Recognition accuracy on the Keck gesture dataset with different features and dictionary sizes (*shape* and *motion* are global features. *STIP* [13] is a local feature.). The recognition accuracy using initial dictionary D^o: (a) 0.23 (b) 0.42 (c) 0.71 (d) 0.81. In all cases, the proposed MMI-2 (red line) outperforms the rest.

4.1.3 Describing Unknown Actions

We illustrate here how unknown actions can be described through a learned attribute dictionary. We first obtain a 500 size initial shape dictionary D^o using 11 out of 14 gesture classes, and keep *flap*, *stop both* and *attention both* as unknown actions. We would expect nearly perfect description to these unknown actions, as we notice these three classes are composed by attributes observed in the rest classes. For example, *flap* is a two-arm gesture "unseen" by the attribute dictionary, but its left-arm pattern is similar to *turn left*, and right-arm is similar to *turn right*.

As shown in Fig. 3, we learned 40 size dictionaries using MMI-2, ME and *Liu-Shah* respectively from D^o. Through visual observation, ME dictionary (Fig. 3b) is most compact as dictionary items look less similar to each other. However, different from the MMI-2 dictionary (Fig. 3a), it contains shapes mostly associated with the action start and end as discussed in Sec. 3.3, which often results in high reconstruction errors shown in Fig. 3d. *Liu-Shah* dictionary (Fig. 3c) only concerns about the discriminability, thus obvious redundancy can be observed in its dictionary. We see from Fig. 3d that, though the action *flap* is unknown to the dictionary, we still obtain a nearly perfect reconstruction through MMI-2, i.e., we can perfectly describe it using attributes in dictionary with corresponding sparse coefficients.

4.1.4 Recognition Accuracy

In all of our experiments, we use the following classification schemes: when the global features, i.e., *shape* and *motion*, are used for attribute dictionaries, we first adopt dynamic time warping (DTW) to align and measure the distance between two action sequences in the sparse code domain; then

Figure 5: Sample frames from the UCF sports action dataset. The actions include: diving, golfing, kicking, weight-lifting, horse-riding, running, skateboarding, swinging-1 (on the pommel horse and on the floor), swinging-2 (at the high bar), walking.

a k-NN classifier is used for recognition. When the local feature *STIP* [13] is used, DTW becomes not applicable, and we simply perform recognition using a k-NN classifier based on the sparse code histogram of each action sequence.

In Fig. 4, we present the recognition accuracy on the Keck gesture dataset with different dictionary sizes and over different global and local features. We use a leave-one-person-out setup, i.e., sequences performed by a person are left out, and report the average accuracy. We choose an initial dictionary size $|D^o|$ to be twice the dimension of an input signal and sparsity 10 is used in this set of experiments. In all cases, the proposed MMI-2 outperforms the rest. The sparse code noise has more effects on the DTW methods than the histogram method, thus, MMI-2 brings more improvements on global features over local features. The peak recognition accuracy obtained from MMI-2 is comparable to 92.86% (motion), 92.86% (shape), 95.24% (shape and motion) reported in [16].

As discussed, the near-optimality of our approach can be guaranteed if the initial dictionary size $|D^o|$ is sufficiently larger than $2|D^*|$. We usually choose a size for D^* to keep $|D^o|$ be 10 to 20 times larger. As shown in Fig. 4, such dictionary size range usually produces good recognition performance. We can also decide $|D^*|$ when the MI increase in (8) is below a predefined threshold, which can be obtained via cross validation from training data.

4.2. Discriminability of Learned Action Attributes

In this section, we further evaluate the discriminative power of learned action attributes using MMI-2.

4.2.1 Recognizing Unknown Actions

The Weizmann human action dataset contains 10 different actions: bend, jack, jump, pjump, run, side, skip, walk, wave1, wave2. Each action is performed by 9 different people. We use the shape and the motion features for attribute dictionaries. In the experiments on the Weizmann dataset, we learn a 50 size dictionary from a 1000 size initial dictionary and the sparsity 10 is used. When we use a leave-one-person-out setup, we obtain 100% recognition accuracy over the Weizmann dataset.

To evaluate the recognition performance of attribute representation for unknown actions, we use a leave-one-action-out setup for dictionary learning, and then use a leave-one-person-out setup for recognition. In this way, one action class is kept unknown to the learned attribute dictionary, and its sparse representation using attributes learned from the rest classes is used for recognition. The recognition accuracy is shown in Table 1.

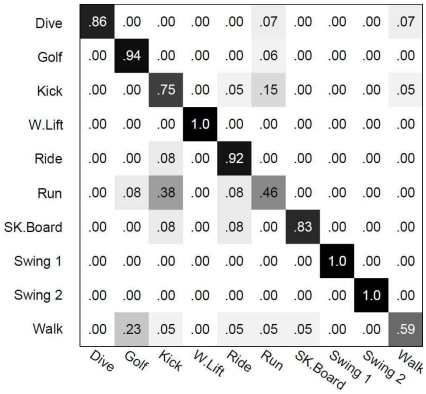


Figure 6: Confusion matrix for UCF sports dataset

It is interesting to notice from the second row of Table 1 that only *jump* can not be perfectly described using attributes learned from the remaining 9 actions, i.e., *jump* is described by a set of attributes not completely provided by the rest actions. By examining the dataset, it is easy to notice that *jump* does exhibit unique shapes and motion patterns.

As we see from the third row of the table, omitting attributes of the *wave2*, i.e., the *wave-two-hands* action, brings down the overall accuracy most. Further investigation shows that, when the *wave2* attributes are not present, such accuracy loss is caused by 33% *pjump* being misclassified as *jack*, which means the attributes contributed by *wave2* is useful to distinguish *pjump* from *jack*. This makes great sense as *jack* is very similar to *pjump* but *jack* contains additional *wave-two-hands* pattern.

4.2.2 Recognizing Realistic Actions

The UCF sports dataset is a set of 150 broadcast sports videos and contains 10 different actions shown in Fig. 5. It is a challenging dataset with a large scenes and viewpoints variability. As the UCF dataset often involves multiple people in the scene, we use tracks from ground-truth annotations. We use the HOG and the motion features for attribute dictionaries. We learned a 60 size dictionary from a 1200 size initial dictionary and the sparsity 10 is used. We adopt

Unknown Action	bend	jack	jump	pjump	run	side	skip	walk	wave1	wave2
Action Accuracy	1.00	1.00	0.89	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Overall Accuracy	1.00	1.00	0.98	0.98	1.00	1.00	1.00	0.99	0.97	0.94

Table 1: Recognition accuracy on the Weizmann dataset using a leave-one-action-out setup for dictionary learning. The second row is the recognition accuracy on the unknown action, and the third row is the overall average accuracy over all classes given the unknown action. The second row reflects the importance of attributes learned from the rest actions to represent the unknown action, and the third row reflects the importance of attributes from the unknown action to represent the rest actions.

a five-fold cross-validation setup. With such basic features and a simple k-NN classifier, we obtain 83.6% average recognition accuracy over the UCF sports action dataset, and the confusion matrix is shown in Fig. 6. Though we use a much simpler classifier, our accuracy is comparable to the 86.6% reported in [29]. The accuracy numbers reported here show how the learned attribute representation of human actions can improve the recognition even with basic features and a simple classifier.

5. Conclusion

We presented an attribute dictionary learning approach via information maximization for action recognition. By formulating the mutual information for appearance information and class distributions between the learned dictionary and the rest of dictionary space into an objective function, we learned a dictionary that is both representative and discriminative. The objective function is optimized through a GP model proposed for sparse representation. The sparse representations for signals enable the use of kernels locality in GP to speed up the optimization process. An action sequence is described through a set of action attributes, which enable both modeling and recognizing actions, even including "unseen" human actions. Our future work includes how to automatically update the learned dictionary for a new action category.

Acknowledgements

This work was supported by a MURI grant N00014-10-1-0934 from the Office of Naval Research and the DARPA Mind's Eye program.

References

[1] M. Aharon, M. Elad, and A. Bruckstein. K-svd: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Trans. on Signal Process.*, 54(1):4311–4322, 2006.

[2] M. Bilenko, S. Basu, and R. J. Mooney. Integrating constraints and metric learning in semi-supervised clustering, 2004. *ICML*.

[3] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes, 2005. *ICCV*.

[4] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection, 2005. *CVPR*.

[5] A. Efros, A. Berg, G. Mori, and J. Malik. Recognizing action at a distance, 2003. *ICCV*.

[6] A. Elgammal, V. Shet, Y. Yacoob, and L. Davis. Learning dynamics for exemplar-based geature recognition, 2003. *CVPR*.

[7] K. Etemad and R. Chellappa. Separability-based multiscale basis selection and feature extraction for signal and image classification. *IEEE Trans. on Image Process.*, 7(10):1453–1465, 1998.

[8] A. Farhadi, I. Endres, and D. Hoiem. Attribute-centric recognition for cross-category generatlization, 2010. *CVPR*.

[9] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes, 2009. *CVPR*.

[10] Z. Jiang, Z. Lin, and L. Davis. Learning a discriminative dictionary for sparse coding via label consistent k-svd, 2011. *CVPR*.

[11] A. Krause, A. Singh, and C. Guestrin. Near-optimal sensor placements in gaussian processes: Theory, efficient algorithms and empirical studies. *JMLR*, (9):235–284, 2008.

[12] C. Lampert, H. Nickisch, and S. Harmerling. Learning to detect unseen object classes by between-class attribute transfer, 2009. *CVPR*.

[13] I. Laptev, M. Marszałek, C. Schmid, and B. Ronfeld. Learning realistic human actions from movies, 2008. *CVPR*.

[14] S. Lazebnik and M. Raginsky. Supervised learning of quantizer codebooks by information loss minimization, 2009. *TPAMI*.

[15] Y. Li, C. Fermuller, and Y. Aloimonos. Learning shift-invariant sparse representation of actions, 2010. *CVPR*.

[16] Z. Lin, Z. Jiang, and L. Davis. Recognizing actions by shape-motion prototype trees, 2009. *ICCV*.

[17] J. Liu and M. Shah. Learning human actions via information maximization, 2008. *CVPR*.

[18] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman. Discriminative learned dictionaries for local image analysis, 2008. *CVPR*.

[19] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman. Supervised dictionary learning, 2008. *NIPS*.

[20] D. Pham and S. Venkatesh. Joint learning and dictionary construction for pattern recognition, 2008. *CVPR*.

[21] I. Ramirez, P. Sprechmann, and G. Sapiro. Classification and clustering via dictionary learning with structured incoherence and shared features, 2010. *CVPR*.

[22] C. Rasmussen and C. Williams. Gaussian processes for machine learning, 2006. *the MIT Process*.

[23] M. D. Rodriguez, J. Ahmed, and M. Shah. Action mach: a spatio-temporal maximum average correlation height filter for action recognition, 2008. *CVPR*.

[24] N. Slonim and N. Tishy. Document clustering using word clusters via the information bottleneck method, 2000. *ACM SIGIR*.

[25] C. Thurau and V. Hlavac. Pose primitive based human action recognition in videos or still images, 2008. *CVPR*.

[26] L. Wang, L. Zhou, and C. Shen. A fast algorithm for creating a compact and discriminative visual codebook, 2008. *ECCV*.

[27] D. Weinland and E. Boyer. Action recognition using exampler-based embedding, 2008. *CVPR*.

[28] L. Yang, R. Jin, R. Sukthankar, and F. Jurie. Unifying discriminative visual codebook genearation with classifier training for object category recognition, 2008. *CVPR*.

[29] A. Yao, J. Gall, and L. V. Gool. A hough transform-based voting framework for action recognition, 2010. *CVPR*.

[30] X. Yu and Y. Aloimonos. Attribute-based transfer learning for object categorization with zero/one training example, 2010. *ECCV*.

[31] Q. Zhang and B. Li. Discriminative k-svd for dictionary learning in face recognition, 2010. *CVPR*.