An information bottleneck approach to optimize the dictionary of visual data

Shyju Wilson, *Member, IEEE,* and C. Krishna Mohan, *Member, IEEE,*

***Abstract*—In this paper, we propose a novel information theoretic approach to obtain compact and discriminative dictionary of visual data. This approach squeezes discriminative information from dictionary for efficient representation using *information bottleneck*. The dictionary is optimized from the initial sparse dictionary which is learned from action data. In this, a constraint information optimization problem is formulated in which mutual information between initial and optimized dictionary is minimized while maximizing mutual information between optimized dictionary and class labels. We use an effective similarity measure, *Jensen-Shannon divergence* with adaptive weightages, for class distributions of each dictionary atom.These adaptive weightages are obtained based on the usage of dictionary atom among different classes. The resultant dictionary becomes discriminative and compact, while retaining maximum information with fewer atoms. Using simple reconstruction error, we test computational efficiency of the proposed method without compromising classification accuracy on popular benchmark datasets. It is further demonstrated how efficiently discriminative information is retained by comparing the classification performance of the dictionary before and after the removal of redundant dictionary atoms.**

***Index Terms*—Dictionary learning, Sparse representation, Information bottleneck, Mutual Information**

I. Introduction

THE evolving field of visual data concerns large volume and complexity of emerging data in the digital world. The availability of capturing devices and data storage capability not only resulted in enormous amount of visual data, but also escalates continuous growth of the data around us. In this era of Big Data, how to handle the data and how to get information from the data being an open conundrum. Finding discriminative and compact representation from codebook or dictionary is widely addressed and relevant in this time [1]–[5]. According to new census, 42.2% of world population is using Internet. Video data is one of the fast growing data day by day as many video sharing sites like YouTube, metacafe, flickr, vimeo, dailymotion etc. contribute large quantity of data everyday. Also, large amount of videos emanate from social media sites like facebook, Google+, Twitter etc. These facts clearly indicate the exponential growth of videos in the digital world. So, an efficient way of representing video data is vital now. Our work is to optimize video data with minimum loss of discriminative information to recognize videos efficiently. In this paper, we propose two level optimization of action data. The first level is to learn input data via sparse coding based approach for initial dictionary. In the next level, the well representative dictionary atoms are extracted from the learned dictionary using information bottleneck approach.

Shyju Wilson and C. Krishna Mohan are with the Visual Learning and Intelligence Labs, Department of Computer Science and Engineering, Indian Institute of Technology Hyderabad, Telangana, India, 502285.
E-mail: cs10p006@iith.ac.in, ckm@iith.ac.in

Sparse coding has been widely used in many signal processing applications [6], [7]. It reconstructs the signal using linear combination of basic building blocks which are called atoms. These atoms, d_i, are grouped into dictionary, so it is also called as *dictionary atoms*. The dictionary $D = \{d_1, d_2 \dots d_K\}$ is over complete dictionary which has infinite number of solutions. In sparse coding, it always looks for sparse solution. The dictionary D is composed of K dictionary atoms and y represents an input signal. We can approximate y as the linear combination of few atoms in the dictionary D,

$$y \approx Dx,$$

where $x \in R^K$ is called sparse vector which is to be found using any of the standard sparse coding algorithm like basis pursuit, matching pursuit etc. LASSO (Least Absolute Shrinkage and Selection Operator) is the variant form of basis pursuit which uses l_1 norm whereas orthogonal matching pursuit (OMP) is variant form of matching pursuit which uses l_0 norm to find sparse vector. Sparse coding gives sparse vector which provides information regarding most correlated atoms in D to reconstruct the input signal y. Dictionary learning algorithms such as method of optimal direction (MOD) [8] online dictionary learning [9] and K-SVD dictionary learning [10], learn adptively the input data into dictionary D and guarantee to converge at local optimum [10]. The dictionary learning alternates between sparse coding and dictionary update. This dictionary learning is very successful in signal reconstruction. By learning discriminative dictionary, we can use this powerful tool efficiently for machine learning purpose.

In fact, the large-sized dictionary leads computational inefficiency and more memory requirement in many machine learning applications. In our work, the first level optimization is carried out by learning the input video data $Y = \{y_1, y_2, \dots y_N\}$ into a dictionary D which is further optimized in the second level of optimization using information bottleneck approach. The discriminative dictionary atoms can be obtained by removing redundant dictionary atoms in the dictionary. This redundancy removal is often a challenging task due to degradation of recognition performance by loosing discriminative information. In order to remove this redundant dictionary atoms, we use information bottleneck approach [11] which minimizes the loss of discriminative information while removing redundancy in the dictionary D. This is a constraint information optimization problem in which mutual information between initial dictionary and optimized dictionary is

minimized while maximizing mutual information between optimized dictionary and class labels. To achieve this, we adopt an efficient approach using Jenson-Shannon divergence [12] with adaptive weightages to remove redundant dictionary atoms. This approach removes redundant dictionary atoms while retaining disriminative dictionary atoms. We evaluate the performance of our approach in the following aspects: (1) first, the removal of redundant dictionary atoms does not affect the classification accuracy; (2) we demonstrate the proposed optimization approach indeed improves the classification time when compared to other existing classification approaches; (3) we compare runtime of the proposed optimization method with other similar methods; (4) we show the adaptive weights in JS divergence affects the classification accuracy compared to fixed weights.

The main contributions of this paper are:

- A novel information theoretic approach for visual data recognition is proposed.
- We combine dictionary learning and information bottleneck to obtain compact and discriminative dictionary of visual data.
- Adaptive weightages for class distributions of each dictionary atom has been used in the similarity measure *Jensen-Shannon divergence*.
- We utilize atom contribution in sparse decomposition to label and share dictionary atoms to find reconstruction error.

The organization of the paper is as follows. Section II discusses related works and section III details proposed approach for optimization in which initial dictionary learning and optimization using information bottleneck are described. Section IV briefs about atom contribution to label and share dictionary atoms which helps to obtain reconstruction error. The detailed experimental study is discussed in section V. Finally, section VI concludes summary of the paper and presents future directions.

II. Related Work

This work proposes two levels of optimization, namely, sparsity based optimization and information theoretic optimization. This provides compact and discriminative dictionary atoms for efficient action recognition. In [5], twenty one binary descriptors are tested to obtain compact and discrminative representation of visual data and they showed that gradient based approaches claim comparatively more discriminative ability. Chen et al. [13] suggest discriminative visual phrase selection for mobile land recognition. This reduces disriminative information loss and removes those that are common across various categories. The sparse based approach is very powerful and widely applied in many machine learning applications. In [14], a set of dictionaries are learned separately from different classes of images and atoms, which contains common features, are shared among classes. The excessive coherent atoms among different classes are discarded. Instead of learning separate dictionaries, sparse based discriminative dictionaries [15] [16] [17] are learned for image classification in which each dictionary atom to be labeled. We use the amount of class information contained in each dictionary atom to label it.

Wright et al. [18] used sparse coding for face recognition and reconstruction error for classification which yielded better result. Yang et al. [19] introduce Fisher discrimination criterion to get discriminative dictionary atoms. The extension of [19] presents a support vector based discriminative dictionary learning model [20]. In [21], Fisher discriminative dictionary learning is exploited to map data from various modalities to common subspace in which inherent relationship between different modalities can be more evident. In [22], Mairal et al. add discriminative term to the dictionary learning which optimize the dictionary. The dictionary learning is tuned to specific task like semi-supervised learning [23] by adding more discriminative terms. This exploits unlabelled data by sparse representation and solves specific task like classification. Pham and Venkatesh [24] learned linear classifier from dictionary and then update the dictionary from the learned classifier. This process will alternate until convergence to get discriminative sparse representation for face recognition. In [25], discriminative comapct vocabulary of context information are obtained for mobile landmark recognition. In [26], reconstruction error and projection of test vector on to the dictionary are used as classification measure. Nguyen et al. [27] applies kernel trick to improve discriminative information, but this needs high computational power and storage which is addressed by linearized kernel dictionary learning [28].

The mutual information has been used as a similarity measure in many machine learning applications [3], [29]. Similar to our approach, Qiu et.al. [3] learnt input data by K-SVD dictionary learning and then select atoms by maximizing mutual information between selected and unselected atoms. They also maximize mutual information between classes to ensure enough representation of all classes in the optimized dictionary. But they have used *Gaussian Process* (GP) model for sparse representation, so the inverse calculation of matrix claims more computational time. In [30] [31], Krause et al. maximize mutual information for optimal placement of sensors based on Gaussian process (GP) which ultimately helps to reduce communication cost. For the selection of compact and discriminative dictionary atoms, Chellapa et.al. [2] maximize mutual information between selected and unselected atoms, between sparse codes and class labels, between input signals and selected atoms and then update dictionary using gradient ascent algorithm.

To learn human actions, Jingen Liu and Mubarak Shah [32] extract 3D interest points called video words and optimize these video words by maximizing mutual information. Lee et al. [29] use mutual information to measure similaity between two activity vectors which are obtained from different cameras. In [33], for image classification and segmentation, codebooks are learned by minimizing the loss of information. Information theoretic approach is an effective tool to determine how much information retains after learning data. In [34], Tishby et.al. systematically analyze information loss while learning through each layer in the deep neural network. Lobel et al. [1] encode mid level representation from different regions of an image using dictionary of linear classifiers. These classifiers

are applied to feature descriptor and max pooling strategy makes total energy of an image as linear combination of max functions to obtain compact and discriminative visual words. Liu et al. [4] developed probabilistic framework for merging criteria to produce well representative codebook.

In this work, we propose information bottleneck method to remove redundancy in the learned dictionary. The information bottleneck method was introduced in late 90's by Tishby et.al [35] [11]. It was an attempt to semantic application of information theoretic approach apart from information theoretic application in the field of communication which was proposed in the midst of 20th century. In our work, we use adaptive weightages in *Jensen-Shannon divergence* [12] to find similar dictionary atoms . This similarity measure is computationally effective and easy to implement when compared to other existing dictionary optimization approaches [2] [3] [30] [31] in which the matrix inverse is involved which claims more computational complexity. The next section details entire optimization procedure.

III. Two level optimization for compact discriminative dictionary

Our goal is to extract well representative information from the input data. In this two level optimization of action data, initially the data is learned or optimized by dictionary learning and further the learned dictionary is optimized by using information bottleneck. The details are discussed below.

A. Learning initial dictionary from action data

This optimization problem can be posed in two ways: sparsity based and error based. For sparsity based minimization, the optimization is given by,

$$\underset{D,X}{\operatorname{argmin}} \|Y - DX\|_F^2 \quad \text{subject to} \quad \forall i \;\; \|x_i\|_p \leq T, \tag{1}$$

and for error based minimization,

$$\underset{D,X}{\operatorname{argmin}} \sum_i \|x_i\|_p \quad \text{subject to} \quad \|Y - DX\|_F^2 \leq \epsilon, \tag{2}$$

where $X = \{x_1, x_2, \ldots x_N\} \in R^{K \times N}$ is sparse matrix, each sparse vector x_i corresponds to input sample y_i. The notation $\|.\|_F$ and $\|.\|_p$ denote frobenius norm and l_p norm, respectively and the l_p norm can be l_0 or l_1. In this work, we deal with dictionary learning with sparsity based minimization as in (1). The K-SVD dictionary learning is used for initial optimization which will give dictionary of K dictionary atoms. Generally, dictionary learning alternates between two steps, sparse coding and dictionary update, to learn initial data $Y \in R^{m \times N}$ into dictionary $D \in R^{K \times N}$ as follows:

$$\underset{D,X}{\operatorname{argmin}} \|Y - DX\|_F^2 \quad \text{subject to} \quad \forall i \;\; \|x_i\|_0 \leq T, \tag{3}$$

where $\|.\|_0$ denotes l_0 norm and number of non zeros values in sparse vector x_i restricted to constraint T. Orthogonal matching pursuit is used to obtain sparse vector in K-SVD dictionary learning. This Sparse coding finds sparse matrix X that minimizes squared error $\|Y - DX\|_F^2$ with fixed D.

To update the dictionary D, every column of D to be updated and X is fixed during updation. Each dictionary atom d_k to be updated seperately, so the updation procedure has to run *K* times. For updation, reconstruction error function can be rewritten as,

$$\begin{aligned} \|Y - DX\|_F^2 &= \|E_k - d_k x^k\| \\ E_k &= Y - \sum_{j \neq k} d_j x^j. \end{aligned} \tag{4}$$

The matrix E_k denotes error matrix which is the error for all N samples when dictionary atom d_k is removed. The row vector x^j is j^{th} row of X, which indicates the usage of dictionary atom d_j by input samples. After removing zeros from x^k and corresponding columns from E_k, SVD is applied to update d_k and x^k. This learned dictionary D is not an optimal one in machine learning perspective, so we can optimize dictionary further by removing redundant dictionary atoms. This dictionary D becomes the input dictionary to the next level of optimization, which is discussed below.

B. Information bottleneck for optimization

In this phase, our goal is to remove the redundancy in the dictionary obtained by k-svd dictionary learning discussed in section III-A. More clearly, we want to optimize the signal $d \in D$ which provides information about another signal $c \in C$. The Notations D and C denote dictionary and class labels respectively. Here our aim is to compress D into $\tilde{D}$ while retaining maximum information about C. In other words, prediction of C from $\tilde{D}$ should be as close as possible the prediction of C from D, so $D \rightarrow \tilde{D}$ and $\tilde{D} \rightarrow C$ are the rules to be optimized. Let $\mathbb{D}$, $\tilde{\mathbb{D}}$, and $\mathbb{C}$ be random variable notation for D, $\tilde{D}$, and C, respectively. We denote probability mass function by $\mathsf{p}(\mathsf{d})$ rather than $\mathsf{p}_{\mathbb{D}}(\mathsf{d})$ for ease of use.

In this optimization problem, we try to minimize mutual information between $\mathbb{D}$ and $\tilde{\mathbb{D}}$ with constraint of mutual information between $\tilde{\mathbb{D}}$ and $\mathbb{C}$. The Shannon's entropy $H(\mathbb{D})$ of discrete random variable $\mathbb{D}$ on alphabet $\mathcal{D}$ is defined by,

$$H(\mathbb{D}) = -\sum_{\mathsf{d} \in \mathcal{D}} \mathsf{p}(\mathsf{d}) \log \mathsf{p}(\mathsf{d}), \tag{5}$$

and the conditional entropy $H(\mathbb{D}|\tilde{\mathbb{D}})$ is defined by,

$$H(\mathbb{D}|\tilde{\mathbb{D}}) = -\sum_{\mathsf{d}} \sum_{\tilde{\mathsf{d}}} \mathsf{p}(\mathsf{d}, \tilde{\mathsf{d}}) \log \mathsf{p}(\mathsf{d}|\tilde{\mathsf{d}}). \tag{6}$$

Mutual information is the amount of information contains in one random variable about another. In other words, it is the reduction in uncertainty of one random variable by knowing another one. $I(\mathbb{D}; \tilde{\mathbb{D}})$ denotes the mutual information between $\mathbb{D}$ and $\tilde{\mathbb{D}}$, which is defined as,

$$\begin{aligned} I(\mathbb{D}; \tilde{\mathbb{D}}) &= H(\mathbb{D}) - H(\mathbb{D}|\tilde{\mathbb{D}}) \\ &= -\sum_{\mathsf{d}} \sum_{\tilde{\mathsf{d}}} \mathsf{p}(\mathsf{d}, \tilde{\mathsf{d}}) \log \frac{\mathsf{p}(\mathsf{d}, \tilde{\mathsf{d}})}{\mathsf{p}(\mathsf{d})\mathsf{p}(\tilde{\mathsf{d}})} \\ &= -\sum_{\mathsf{d}} \sum_{\tilde{\mathsf{d}}} \mathsf{p}(\mathsf{d})\mathsf{p}(\tilde{\mathsf{d}}|\mathsf{d}) \log \frac{\mathsf{p}(\tilde{\mathsf{d}}|\mathsf{d})}{\mathsf{p}(\tilde{\mathsf{d}})} \end{aligned} \tag{7}$$

The redundancy among the dictionary atoms can be effectively removed by information bottleneck approach which squeezes the information that $\mathbb{D}$ contains about $\mathbb{C}$ through *bottleneck* formed by well representative dictionary atoms in $\tilde{\mathbb{D}}$. This problem is analogous to *rate distortion function, R(D)*, [36] which provides trade-off between rate and distortion. Here the distortion D means compression which also depends on relevant features for better representation, i.e., rate R. Then the important question arises, what is the affordable distortion for achievable rate. This is a constrained information optimization problem which is addressed in [11].

We need to compress variable $\mathbb{D}$ into $\tilde{\mathbb{D}}$ which retains maximum information about $\mathbb{C}$. This yields Markov chain $\tilde{\mathbb{D}} \rightarrow \mathbb{D} \rightarrow \mathbb{C}$ and based on data processing inequality [36], amount of information in $\tilde{\mathbb{D}}$ about $\mathbb{C}$ is given by,

$$I(\tilde{\mathbb{D}};\mathbb{C}) \leq I(\mathbb{D};\mathbb{C}). \tag{8}$$

Optimization will be carried in such a way that it minimizes the mutual information $I(\tilde{\mathbb{D}};\mathbb{D})$ while maximizing the constraint $I(\tilde{\mathbb{D}};\mathbb{C})$ as high as possible. This optimization can be achieved by minimizing the following function:

$$\underset{p(\tilde{d}),p(\tilde{d}|d)}{\operatorname{argmin}} \quad I(\tilde{\mathbb{D}};\mathbb{D}) - \beta I(\tilde{\mathbb{D}};\mathbb{C}), \tag{9}$$

where β is the Lagrange multiplier. The self consistent equations $p(\tilde{d})$ and $p(\tilde{d}|d)$ can be obtained by minimizing (9). There is a well known iterative procedure called *Blahut-Arimoto Algorithm* [37] to solve this problem. We can find $p(\tilde{d})$ and $p(\tilde{d}|d)$ which minimize mutual information subject to distortion $\text{dist}(d,\tilde{d})$. These iterative steps, $(t+1)^{th}$ update, are given by,

$$\begin{cases} p_{t+1}(\tilde{d}) & = \sum_d p(d)p_t(\tilde{d}|d) \\ p_{t+1}(\tilde{d}|d) & = \frac{p_t(\tilde{d})\exp(-\beta\,\text{dist}(d,\tilde{d}))}{\sum_{\tilde{d}} p_t(\tilde{d})\exp(-\beta\,\text{dist}(d,\tilde{d}))}. \end{cases} \tag{10}$$

These iterations converge to a unique minimum in the convex set of two distributions [36] [37].

The optimal assignment, which minimizes (9), satisfies the following equation,

$$p(\tilde{d}|d) = \frac{p_t(\tilde{d})}{\mathcal{N}(d,\beta)}\exp\Big[-\beta \sum_c p(c|d)\log\frac{p(c|d)}{p(c|\tilde{d})}\Big], \tag{11}$$

where $\mathcal{N}(d,\beta)$ is normalization function. The detailed proof is given in [35]. The distribution $p(c|\tilde{d})$ is given by Baye's rule and Markov chain $\tilde{\mathbb{D}} \rightarrow \mathbb{D} \rightarrow \mathbb{C}$,

$$\begin{aligned} p(c|\tilde{d}) &= \sum_d p(c|d)p(d|\tilde{d}) \\ &= \frac{1}{p(\tilde{d})}\sum_d p(c|d)p(\tilde{d}|d)p(d) \end{aligned} \tag{12}$$

and,

$$p(\tilde{d}) = \sum_d p(\tilde{d}|d)p(d). \tag{13}$$

The relative entropy or Kullback-Leibler divergence [38] is the well known distance measure between two probability distributions. The relative entropy between two probability mass functions $p(x)$ and $q(x)$ is defined as,

$$D(p||q) = \sum_x p(x)\log\frac{p(x)}{q(x)}. \tag{14}$$

Then, the equation (11) becomes,

$$p(\tilde{d}|d) = \frac{p_t(\tilde{d})}{\mathcal{N}(d,\beta)}\exp\Big[-\beta\, D\big(p(c|d)||p(c|\tilde{d})\big)\Big]. \tag{15}$$

The Kullback-Leibler divergence becomes distortion measure in (10). This makes sense, because it is a natural distortion measure to find distance between distributions $p(c|d)$ and $p(c|\tilde{d})$. Here, we use Jensen-Shannon divergence instead of Kullback-Leibler divergence because in the Jensen-Shannon divergence, we can weigh the distribution of class given dictionary atom for better comparison and the information loss, δI_c, can be computed in an efficient way which are explained in the next section.

1) JS divergence using adaptive weightages: Jensen-Shannon divergence is based on *Jensen's inequality* and *Shannon's entropy*. In this, we can assign weights (prior probabilities) to different probability distributions which helps decision problems and it provides both lower and upper bound for the Bayes' probability of misclassification error [12]. For *I* directed divergence [38] and it's symmetric measure *J* divergence, both distributions should be *absolutely continuous* with respect to each other. This is not an issue in Jensen-Shannon divergence. Unlike other divergence measures, this can be generalized for more than two distributions. Let $p_1, p_2 \ldots, p_n$ be n probability distributions with weightages $\pi_1, \pi_2, \ldots, \pi_n$, respectively, and $\sum_i \pi_i = 1$. The generalized Jensen-Shannon is defined by,

$$JS_\pi(p_1, p_2, \ldots, p_n) = H\Big(\sum_i \pi_i p_i\Big) - \sum_i \pi_i H(p_i). \tag{16}$$

These properties of Jensen-Shannon divergence are very helpful in our context. In this work, we efficiently merge similar dictionary atoms using Jenson-Shannon divergence and these merging steps are explained in section III-B2. The best possible merge is determined by the loss of mutual information, δI_c, i.e.,

$$\delta I_c = I(\mathbb{Z}_m;\mathbb{C}) - I(\mathbb{Z}_{m-1};\mathbb{C}) \tag{17}$$

the loss of information is evaluated for every possible pair in $\mathbb{Z}_m$ ($\mathbb{Z}_m$ be the current m-partition and $\mathbb{Z}_{m-1}$ be the partition after merging a pair). This is a greedy approach ie., for every pair, it looks for best possible merge. For each pair, $O(m.|\mathbb{C}|)$ operations are needed. Using Jensen-Shannon divergence, loss of mutual information due to merge can be calculated in $O(|\mathbb{C}|)$ operations. The loss of mutual information, δI_c, can be defined [11] as,

$$\delta I_c = \big(p(z_i) + p(z_j)\big) JS_\pi\big(p(c|z_i), p(c|z_j)\big), \tag{18}$$

where $\pi = [\pi_i, \pi_j]$. In this, we have given adaptive weightages to both distributions of $p(c|z_i)$ and $p(c|z_j)$ based on the presence of dictionary atom among different classes. Here

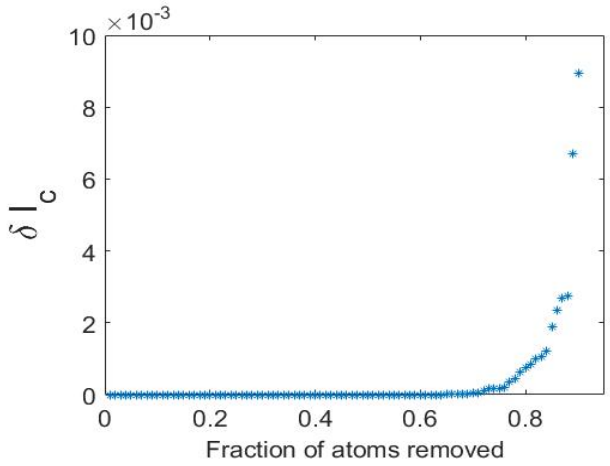


Fig. 1: Loss of information while removing dictionary atoms in KTH dataset.

we give more priority to distributions which included more dictionary atoms. These weights π_i and π_j are assigned as,

$$\pi_i = \frac{\mathsf{p}(\mathsf{z_i})}{\mathsf{p}(\mathsf{z_i}) + \mathsf{p}(\mathsf{z_j})}$$
$$\pi_j = \frac{\mathsf{p}(\mathsf{z_j})}{\mathsf{p}(\mathsf{z_i}) + \mathsf{p}(\mathsf{z_j})} \tag{19}$$

In this way, closest dictionary atoms can be determined while computing distance between distributions $\mathsf{p}(\mathsf{c}|\mathsf{z_i})$ and $\mathsf{p}(\mathsf{c}|\mathsf{z_j})$ by effectively using *Jensen-Shannon divergence*. The figure 1 shows loss of mutual information, δI_c, while removing dictionary atoms. It can be seen that the loss of information increases rapidly after a particular point where we can stop the removal of redundant dictionary atoms. From the figure 1, we can approximate the optimal number of dictionary atoms to be retained.

2) Removal of redundant dictionary atoms: We merge similar dictionary atoms using *Jensen-Shannon divergence* to remove redundant dictionary atoms. This section explains, how the merging process carried out to remove redundancy. To avoid confusion, we use one more variable Z and $\tilde{Z}$. Initially, Z is equal to D and the relation between Z and $\tilde{Z}$ is just one step away in the merging process, i.e., after merging dictionary atoms in Z, we will get new compressed dictionary $\tilde{Z}$. For merging, first we need to initialize the following:

$$\mathbb{Z} = \mathbb{D}, \;\; \mathsf{z_i} = \mathsf{d_i} \tag{20}$$
$$\mathsf{p}(\mathsf{c}|\mathsf{z_i}) = \mathsf{p}(\mathsf{c}|\mathsf{d_i}) \;\; \text{for every } \mathsf{c} \in \mathbb{C}, \tag{21}$$
$$\mathsf{p}(\mathsf{z_i}|\mathsf{d_j}) = \begin{cases} 1 & \text{if } \mathsf{j{=}i} \\ 0 & \text{otherwise} \end{cases} \tag{22}$$

and compute distance for every $i, j \in \{1, \ldots, N\}, i < j$

$$S_{i,j} = \big(\mathsf{p}(\mathsf{z_i}) + \mathsf{p}(\mathsf{z_j})\big)\, \mathsf{JS}_\pi\big[\mathsf{p}(\mathsf{c}|\mathsf{z_i}), \mathsf{p}(\mathsf{c}|\mathsf{z_j})\big] \tag{23}$$

The distance matrix S is a lower triangular matrix with zeros in diagonal and it is used to find most similar atoms for merging process. After each merge, the probability of dictionary atoms, which are included in merging, to be updated. Here, we merge two similar atoms at a time instead of merging more than two. In this way, we can understand loss of information at every merge and take decision about optimal merge which gives minimum information loss between D and C. By merging process, we remove redundant atoms with minimum loss of discriminative information. The atoms which are having minimum distance can be computed from S,

$$< u, v >= \underset{i,j}{\operatorname{argmin}}(S_{i,j}). \tag{24}$$

Then merge $(z_u, z_v) \rightarrow \tilde{z}$. So, the probabilities of dictionary atoms, which are currently merged, to be updated after merging as shown below:

$$\mathsf{p}(\tilde{\mathsf{z}}) = \mathsf{p}(\mathsf{z_u}) + \mathsf{p}(\mathsf{z_v}) \tag{25}$$
$$\mathsf{p}(\mathsf{c}|\tilde{\mathsf{z}}) = \frac{1}{\mathsf{p}(\tilde{\mathsf{z}})}\big(\mathsf{p}(\mathsf{z_u}, \mathsf{c}) + \mathsf{p}(\mathsf{z_v}, \mathsf{c})\big) \tag{26}$$
$$\mathsf{p}(\tilde{\mathsf{z}}|\mathsf{d_j}) = \begin{cases} 1 & \text{if } \mathsf{d_j} \in \tilde{\mathsf{z}} \\ 0 & \text{otherwise} \end{cases}$$
$$Z = \Big\{Z - \{z_u, z_v\}\Big\} \cup \Big\{\tilde{z}\Big\} \tag{27}$$

Finally, the distance between $\tilde{z}$ and all other z_i are updated in the distance matrix S. This method merges similar dictionary atoms and the merging process will be stopped at the point where the information loss, δI_c, is minimum. In this way, we can approximate the optimal number of dictionary atoms to be retained without loosing discriminative information. From each merged group, one representative dictionary atom is selected as the mean of similar dictionary atoms in the group. Next we use simple reconstruction error to see how good this optimized dictionary is.

IV. Atom Contribution and Sharing

In this section, we discuss effective utilization of the distribution of dictionary atoms to test the input data. The optimized dictionary $\tilde{D}$ is obtained after the removal of redundant atoms in the second phase of optimization. In this work, we use simple reconstruction error, $y_i - Dx_i$, to evaluate the performance of the optimized dictionary $\tilde{D}$. Prior to this, each dictionary atom is to be labelled to find reconstruction error of test data. The label of the dictionary atom $\tilde{d_k}$ is determined from sparsity coefficients in the sparse matrix X i.e. the contribution of dictionary $\tilde{d_k}$ to the particular class of input vectors while learning the dictionary. So, the label is assigned based on the maximum contribution of $\tilde{d_k}$ among different classes in C, i.e.,

$$\underset{c}{\operatorname{argmax}} \sum_{i=1}^{C_t} |x_{k,i}| \quad , \qquad \forall c \in C \tag{28}$$

where $x_{k,i}$ denotes k^{th} element of sparse vector x_i and C_t is number of input vectors in class t. In other way, we can say maximum amount of class information contained in the dictionary atom determines the label of the dictionary atom. This is a maximum a posterior probability of $\mathsf{p}(\mathsf{c}|\tilde{\mathsf{d_k}})$. These dictionary atoms can be shared among different classes if it contributes equally to more than one class which ultimately helps overall recognition task.

These labeled dictionary atoms can be used to find reconstruction error for test vector y_i. The sparse vector x_i can be obtained in two ways by Batch OMP [39]: one way is to use all dictionary atoms while the second way uses only atoms belong to particular class c, ie.,

$$x_i = \text{OMP}(y_i, D, T_1) \tag{29}$$

$$x_i^c = \text{OMP}(y_i, D^c, T_2) \tag{30}$$

where D^c and x_i^c represent dictionary atoms and sparse coefficients corresponds to class c, respectively. T_1 and T_2 are sparsity constraints and T_1 is always larger than T_2. Then we find the reconstruction error of y_i based on each class and the class of minimum error will be assigned to test vector y_i as,

$$\min_c(\|y_i - D^c x_i^c\|^2 + \|y_i - D^c x_i^{c'}\|^2), \tag{31}$$

$x_i^{c'}$ denotes coefficients belong to classs c in sparse vector x_i in equation (29). Each class has it's own sparse decomposition which eventually helps to determine class of the test input.

V. Experimental Results

The performance of the proposed optimization approach is evaluated using different benchmark datasets. For the experiment, we have used USPS digit database [40], AR face database [41] and three action datasets, namely, UCF sports [42], KTH [43] and HMDB51 [44]. Action datasets are represented by action bank features which are used by Sadanand and Corso in their work [45]. The action bank features comprise of many individual action detectors which constitute mid-level representation of action data and carry rich semantic information. For all databases, feature vectors are stacked as matrix. Moreover, each feature vector is mean extracted and normalized to unit l_2 norm. Initially, the input data matrix is learned by K-SVD dictionary learning. In this experiment, we have performed 20 dictionary learning iterations and the sparsity constraint T is determined empirically.

The learned dictionary is further optimized by information bottleneck approach as described in section III-B and this optimized dictionary is used in the experimental evaluation. In [3], the learned dictionary is optimized by comparing sparse decompositions in terms of mutual information using Guassian process. In this, inverse of covariance of sparse matrix is to be determined which is computationally expensive. In our method, instead of computing inverse of the matrix, we used computationally efficient Jensen-Shannon divergence to compare distributions as explained earlier. The recognition accuracies are determined based on the minimum reconstruction error as discussed in the section IV. We also compare our approach with traditional classifiers such as KNN (K nearest neighbor), SVM (support vector machine) etc. All experiments are conducted on the same machine and execution time of classification and dictionary optimization are determined to compare with other similar approaches.

A. Evaluation on the USPS digit dataset

The USPS database consists of handwritten digits of 0-9 which constitute 10 classes. There are 7291 training and

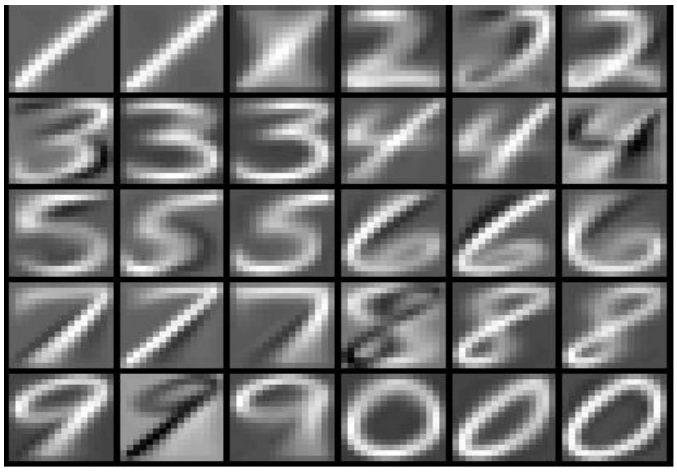


Fig. 2: USPS digit dataset: dictionary atoms obtained after applying K-SVD dictionary learning

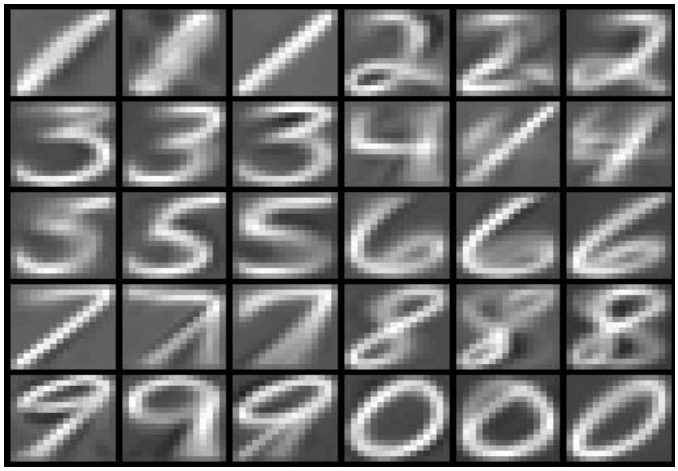


Fig. 3: USPS digit dataset: dictionary atoms obtained after applying proposed approach

TABLE I: Comparison of time (measured in seconds) taken to optimize dictionary from initial dictionary with other approaches viz. MMI, MMI-2.

	Initial dictionary size	Optimized dictionary size	MMI	MMI-2	Our method
UCF	100	50	0.74	0.70	0.67
KTH	200	100	5.85	6.64	1.90
KTH	300	150	14.43	15.95	4.32
USPS	400	300	15.21	16.27	4.15

TABLE II: Performance comparison of UCF sports action classification with existing methods.

Method	Average performance (%)
Proposed method	**95.6**
Sadanand et.al [45]	95.0
Yao et al. [46]	86.6
Qiu et al. [3]	83.6
Rodriguez et al. [47]	69.2
Yeffet Wolf [48]	79.2

TABLE III: Comparing recognition accuracy (%) and testing time (measured in seconds) of our proposed approach with KNN and linear-SVM classifier.

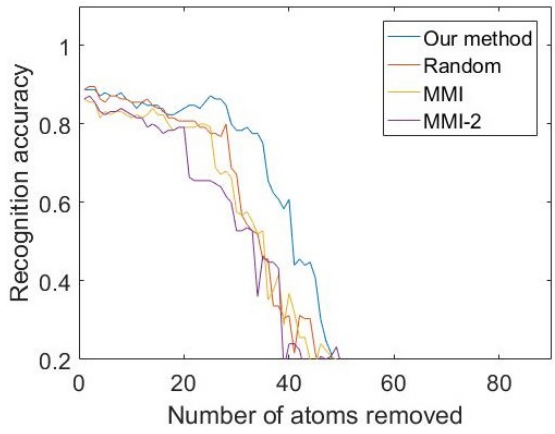
	KNN		SVM		**Proposed Method**	
	Acc.	Time	Acc.	Time	Acc.	Time
USPS	91.00	1.212	95.00	1.961	**97.60**	**0.512**
AR Face	85.00	0.204	91.00	0.295	**94.60**	**0.193**
UCF10	88.00	0.312	95.00	0.486	**95.60**	**0.203**
KTH	78.95	1.950	97.15	3.121	**97.60**	**1.942**
HMDB 51	26.59	190.12	26.91	450.61	**35.32**	**188.190**

2007 test images of digits of size 16×16 which become feature vector of dimension 256. The figures 2 and 3 compare dictionary atoms obtained directly and proposed approach. The figure 2 gives visualization of dictionary atoms obtained using the direct application of K-SVD dictionary learning on USPS data. Whereas figure 3 visualizes dictionary atoms obtained after removing dictionary atoms using proposed approach from the initial dictionary of size 100. It can be observed that atoms in figure 3 are more discriminative than figure 2 which shows our optimization method tries to retain maximum discriminative atoms than direct approach.

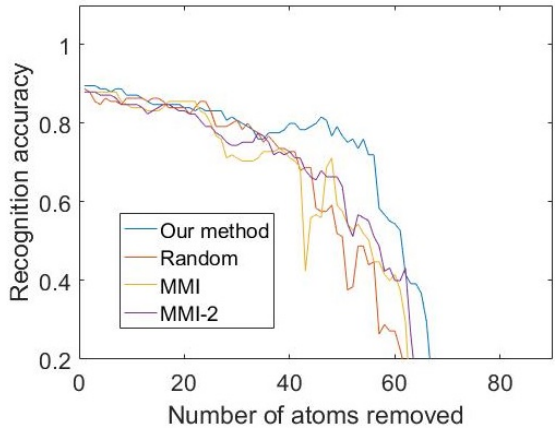
First, we evaluate the removal of dictionary atoms does not affect classification accuracy. For the experiment 40, 30 and 10 dictionary atoms are learned from each class which constitute dictionary of size 400, 300, 100 respectively. The sparsity constraint T is taken as 5. Table IV shows classification accuracy and time of the initial dictionary and optimized ditionary in which it preserves the accuracy even after removing redundant dictionary atoms. The maximum performance we achieved is 97.2% which is comparable to other approaches [28]. Table III compares the classification accuracy and time with other traditional approaches. Our approach shows good computational efficiency in classification when compared to SVM and KNN. Another impact of our approach is the time taken for the optimization process. We compare our method with other similar methods MMI, MMI-1 in [3], Table I shows proposed approach clearly outperforms other methods in computational aspects. Table V indicates adaptive weightages help to merge similar dictionary atoms compared to equal weghtages (at a time only two distributions are compared, so weights are 0.5 and 0.5) and this adaptive weights improve overall accuracy.

B. Evaluation on the AR face dataset

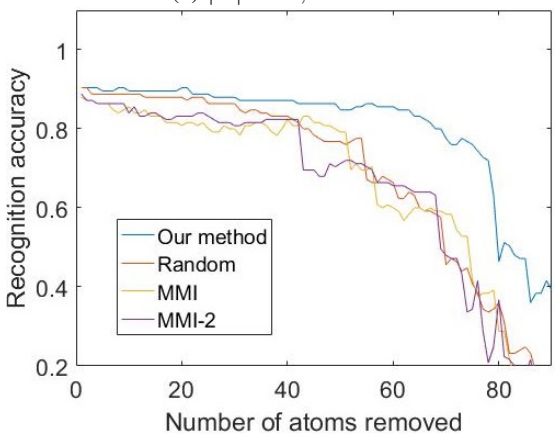
The original AR Face database contains 4000 color images of faces from 126 people, namely, 70 men and 56 women. The frontal view face images are taken based on different facial expressions, illumination conditions, occlusions etc. Following the experiment in [16], 2600 images were chosen from first 50 classes of males and first 50 classes of females, so total 100 classes for the experiment. Each class has 26 images in which 20 for training and remaining for testing. Table III gives performance comparison of the proposed method with KNN and SVM. It can be observed that the proposed dictionary learning method perform better than KNN and SVM in terms of both classification accuracy and time. As you can see in Table IV, dictionary is learned 1500 atoms because the number



(a) $|D| = 60$, $T = 15$



(b) $|D| = 80$, $T = 10$



(c) $|D| = 100$, $T = 2$

Fig. 4: UCF action data: Performance comparison of the proposed method (for different dictionary sizes) with other approaches, viz., random removal of atoms, MMI and MMI-2.

TABLE IV: Comparing recognition accuracy and time (%) of initial dictionary and optimized dictionary.

	Initial	accuracy	time	Optimized	accuracy	time
USPS	$\|D\| = 400$	97.20	0.124	$\|D\| = 300$	96.80	0.118
	$\|D\| = 300$	95.50	0.119	$\|D\| = 200$	95.20	0.094
	$\|D\| = 100$	92.20	0.086	$\|D\| = 90$	92.60	0.069
AR	$\|D\| = 1500$	94.60	1.526	$\|D\| = 1400$	93.00	1.420
	$\|D\| = 1000$	92.10	1.350	$\|D\| = 900$	90.50	1.263
	$\|D\| = 800$	89.00	1.116	$\|D\| = 700$	82.83	1.031
UCF10	$\|D\| = 100$	95.60	0.194	$\|D\| = 70$	95.00	0.130
	$\|D\| = 80$	87.20	0.166	$\|D\| = 70$	88.00	0.120
	$\|D\| = 60$	84.00	0.154	$\|D\| = 50$	84.20	0.117
KTH	$\|D\| = 300$	96.30	0.708	$\|D\| = 200$	97.60	0.542
	$\|D\| = 200$	94.51	0.555	$\|D\| = 100$	94.53	0.344
	$\|D\| = 100$	94.41	0.343	$\|D\| = 50$	94.26	0.269
HMDB 51	$\|D\| = 900$	36.70	195.068	$\|D\| = 600$	32.30	87.550
	$\|D\| = 650$	33.32	90.253	$\|D\| = 590$	32.57	85.931

TABLE V: Comparing recognition accuracy (%) when we use equal weightage and adaptive weigtage.

	Equal wts.	Adaptive wts.
USPS	96.30	97.20
AR Face	92.10	94.55
UCF10	94.10	95.60

of classes are high and we got 94.6% accuracy which is comparable to [16] [28]. The atom removal from dictionary of size 800 causes much performance degradation due to loss of more discriminative information. As shown in Table V, the adaptive weightages improve the classification performance significantly.

C. Evaluation on the UCF sports action data

UCF sports action dataset has 10 different classes of sports viz. diving, golfing, kicking, weight lifting, horse riding, running. skate boarding, swinging bench, swinging side angle and walking. Experiments have been done with five fold cross validation, ie., four folds were used for training and remaining one for testing. We experiment different initial dictionaries of size 100,80, 60 with sparsity of 3, 10, 15, respectively. The dictionary of size 60 learned with sparsity $T = 15$, this includes more dictionary atoms while learning and improves overall recognition performance. The atoms are removed in each iteration and our results are compared with random removal, MMI, MMI-2 shown in Figure 4. Whenever it reaches smaller and smaller dictionary size, our method clearly outperforms other methods. After removing 50% of atoms from the initial dictionary, proposed method still maintain good performance. The computational efficiency of our approach is also better than MMI and MMI-2 as shown Table I. The performance of our proposed approach with other state of art approach is shown in Table II and we achieved comparable result with [45], but dominate performances in other methods [48] [47] [46] [3]. In addition, this optimization tremendously reduces classification time compared to other traditional approaches such as SVM, KNN as shown in Table III. Our approach shows better performance in both recognition accuracy and testing time compared to SVM and KNN classifier.

The figure 6(a) shows mutual information between optimized dictionary $\tilde{\mathbb{Z}}$ and class $\mathbb{C}$, ie.,$I(\tilde{\mathbb{Z}};\mathbb{C})$. It can be observed that, our optimization problem tries to maximize $I(\tilde{\mathbb{Z}};\mathbb{C})$. In contrast to $I(\tilde{\mathbb{Z}};\mathbb{C})$, the mutual information between optimized dictionary $\tilde{\mathbb{Z}}$ and initial dictionary $\mathbb{D}$, $I(\tilde{\mathbb{Z}};\mathbb{D})$, to be minimized which can be seen in figure 6(b). The recognition accuracies of initial dictionary and optimized dictionary are shown in Table IV which indicate our method could remove the redundant dictionary atoms without degrading recognition performance. This resulted in better classification time. The dictionaries of size 80 and 60 slightly improve the recognition accuracy after removing the redundancy.

D. Evaluation on the KTH action dataset

In this dataset, 25 different subjects performing 6 different actions, which are walking, jogging, running, boxing, hand waving and hand clapping. We partitioned data into 3 folds and 2 folds used as training data, remaining one as testing data. Here, three different initial dictionaries of sizes 300, 200, 100 are learned with sparsity 3, 7, 3, respectively. As shown in Table I, computational time of our optimization is better than other approaches which suffer computational burden of inverse calculation of the matrix. We achieved recognition accuracy of 97.60% which is comparable to 98.20% in [45]. In table III, testing time is comparable to KNN but recognition accuracy is far better when compared to KNN classifier but in case of SVM, we got better testing time. Figure 5 shows comparison of our result with random removal, MMI and MMI-2. In this dataset, performance of all methods differs slightly, because this is comparatively easy dataset and feature vectors are well represented. Still the clear difference is evident at smaller dictionary sizes as seen in Figure 5. The table IV compares recognition accuracies of initial and optimized dictionaries on different dictionary sizes. Consider the dictionary of size 200, after removing half of the dictionary still it shows good accuracy. Two confusion matrices of dictionary of size 100 and it's optimized dictionary of size 50 using our method are shown in the Table VI and VII, respectively. It can be observed that there is a minute variation in the recognition performance which clearly indicates that this proposed method retains maximum discriminative information while optimizing.

E. Evaluation on the HMDB action data

Here we conducted experiment with very challenging dataset discussed in previous sections. There are 51 actions

TABLE VI: Confusion matrix of KTH dataset using initial dictionary of size 100

	boxing	clapping	handwaving	jogging	running	walking
boxing	1.0	0	0	0	0	0
clapping	0	0.92	0.08	0	0	0
handwaving	0	0.03	0.97	0	0	0
jogging	0	0	0	1.0	0	0
running	0	0	0	0	1.0	0
walking	0	0	0	0	0	1.0

TABLE VII: Confusion matrix of KTH dataset using optimized dictionary of size 50.

	boxing	clapping	handwaving	jogging	running	walking
boxing	1.0	0	0	0	0	0
clapping	0	0.92	0.06	0.02	0	0
handwaving	0	0.06	0.94	0	0	0
jogging	0	0	0	1.0	0	0
running	0	0	0	0	1.0	0
walking	0	0	0	0	0	1.0

categories in this dataset. In this experiment, the dataset is divided into 10 folds in which 9 folds are used for training and remaining one for testing. We achieved recognition accuracy of 36.70% compared to 26.9% [45] which is benchmark result using action bank features. The Table III compares the proposed method with KNN and SVM in which the time taken for SVM classifier is more than double of testing time of our method because of the large input data. In case of KNN, it got only 26.59% compared to our accuracy of 35.32%. Recognition accuracy and computational time of initial and optimized dictionaries are shown in Table IV. We have learned dictionaries of size 900 and 650 with sparsity T=10. The dictionary of size 650 is optimized into 590 dictionary by removing 60 atoms, but recognition accuracy only vary from 35.32% to 35.17%. There are 300 atoms removed from the dictionary of size 900 and it can be seen that recognition accuracy reduced 4.4% in the optimized dictionary, but computational time reduced drastically. There is more information loss in this compared to previous dataset because of the high variability and large number of classes in the dataset, but still it gives comparable performance.

VI. Conclusion

In this paper, we proposed well discriminative and computationally efficient dictionary optimization method. Dictionary learning is the fastest way to get initial dictionary rather than clustering approach used in previous approaches [4] [32]. In this work, we formulated constraint information optimization problem where we minimized mutual information between optimized dictionary and initial dictionary while maximizing mutual information between optimized dictionary and class labels. The discriminative dictionary is obtained by removing redundant atoms using *Jensen-Shannon divergence* which is simple and computationally effective way to find similar distribution in atoms among classes. Hence, this proposed approach can be applied to large amount of data. Experiments on three benchmark datasets proved that the proposed approach not only retain discriminative information, but computationally efficient when compared to other similar kind dictionary optimization. In the future work, we concentrate on updating representative dictionary atom of similar group with respect to removal of atoms in order to minimize loosing discriminative information.

References

[1] H. Lobel, R. Vidal, and A. Soto, "Learning shared, discriminative, and compact representations for visual recognition," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 37, no. 11, pp. 2218–2231, Nov 2015.

[2] Q. Qiu, V. Patel, and R. Chellappa, "Information-theoretic dictionary learning for image classification," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 36, no. 11, pp. 2173–2184, Nov 2014.

[3] 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*, Nov 2011, pp. 707–714.

[4] L. Liu, L. Wang, and C. Shen, "A generalized probabilistic framework for compact codebook creation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 2, pp. 224–237, Feb 2016.

[5] S. Madeo and M. Bober, "Fast, compact, and discriminative: Evaluation of binary descriptors for mobile applications," *IEEE Transactions on Multimedia*, vol. 19, no. 2, pp. 221–235, Feb 2017.

[6] C. T. Lee, Y. H. Yang, and H. H. Chen, "Multipitch estimation of piano music by exemplar-based sparse representation," *IEEE Transactions on Multimedia*, vol. 14, no. 3, pp. 608–618, June 2012.

[7] T. Guha and R. K. Ward, "Image similarity using sparse representation and compression distance," *IEEE Transactions on Multimedia*, vol. 16, no. 4, pp. 980–987, June 2014.

[8] K. Engan, S. O. Aase, and J. H. Husoy, "Method of optimal directions for frame design," in *Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on*, vol. 5, 1999.

[9] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online dictionary learning for sparse coding," in *Proceedings of the 26th Annual International Conference on Machine Learning*, ser. ICML '09. New York, NY, USA: ACM, 2009, pp. 689–696.

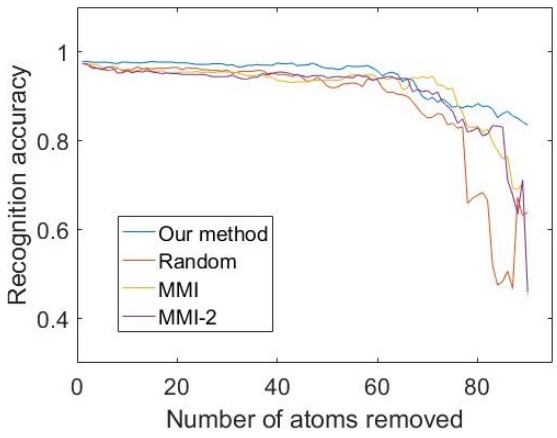
[10] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *Signal Processing, IEEE Transactions on*, vol. 54, no. 11, pp. 4311–4322, 2006.

[11] N. Slonim and N. Tishby, "Agglomerative information bottleneck," in *Advances in Neural Information Processing systems (NIPS-12)*. MIT Press, 1999, pp. 617–623.

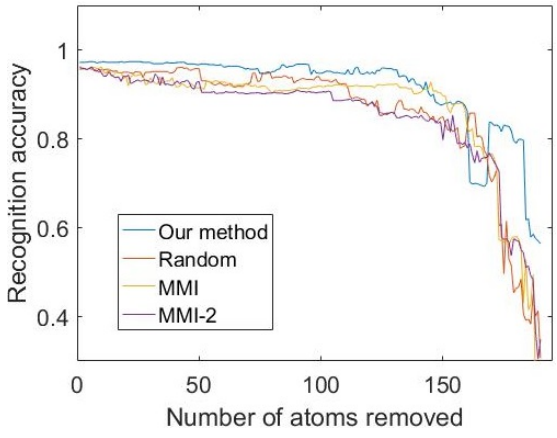
[12] J. Lin, "Divergence measures based on the shannon entropy," *Information Theory, IEEE Transactions on*, vol. 37, no. 1, pp. 145–151, Jan 1991.

[13] T. Chen, K. H. Yap, and D. Zhang, "Discriminative soft bag-of-visual phrase for mobile landmark recognition," *IEEE Transactions on Multimedia*, vol. 16, no. 3, pp. 612–622, April 2014.

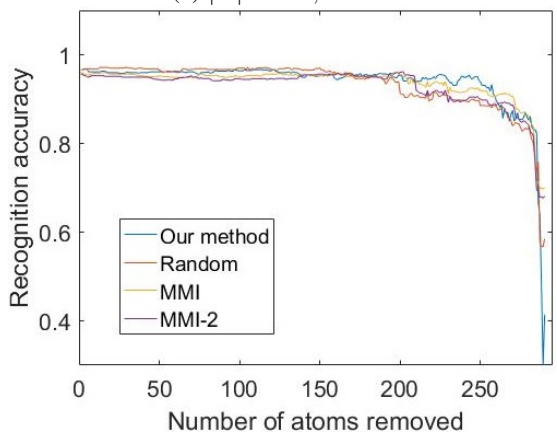
[14] I. Ramirez, P. Sprechmann, and G. Sapiro, "Classification and clustering via dictionary learning with structured incoherence and shared features," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, June 2010, pp. 3501–3508.



(a) $|D| = 100$, $T = 3$



(b) $|D| = 200$, $T = 7$



(c) $|D| = 300$, $T = 1$

Fig. 5: KTH action data: Performance comparison of the proposed method (for different dictionary sizes) with other approaches, viz., random removal of atoms, MMI and MMI-2.

[15] J. Mairal, J. Ponce, G. Sapiro, A. Zisserman, and F. R. Bach, "Supervised dictionary learning," in *Advances in Neural Information Processing Systems 21*, D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, Eds. Curran Associates, Inc., 2009, pp. 1033–1040. [Online]. Available: http://papers.nips.cc/paper/3448-supervised-dictionary-learning.pdf

[16] Z. Jiang, Z. Lin, and L. Davis, "Label consistent k-svd: Learning a discriminative dictionary for recognition," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, no. 11, pp. 2651–2664, Nov 2013.

[17] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, "Discriminative learned dictionaries for local image analysis," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, June 2008, pp. 1–8.

[18] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 210–227, Feb 2009.

[19] M. Yang, L. Zhang, X. Feng, and D. Zhang, "Fisher discrimination dictionary learning for sparse representation," in *2011 International Conference on Computer Vision*, Nov 2011, pp. 543–550.

[20] S. Cai, W. Zuo, L. Zhang, X. Feng, and P. Wang, "Support vector guided dictionary learning," in *Computer Vision – ECCV 2014*, ser. Lecture Notes in Computer Science, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds., vol. 8692. Springer, 2014, pp. 624–639.

[21] C. Deng, X. Tang, J. Yan, W. Liu, and X. Gao, "Discriminative dictionary learning with common label alignment for cross-modal retrieval," *IEEE Transactions on Multimedia*, vol. 18, no. 2, pp. 208–218, Feb 2016.

[22] J. Mairal, J. Ponce, G. Sapiro, A. Zisserman, and F. R. Bach, "Supervised dictionary learning," in *Advances in Neural Information Processing Systems 21*, D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, Eds. Curran Associates, Inc., 2009, pp. 1033–1040. [Online]. Available: http://papers.nips.cc/paper/3448-supervised-dictionary-learning.pdf

[23] J. Mairal, F. Bach, and J. Ponce, "Task-driven dictionary learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 4, pp. 791–804, April 2012.

[24] D.-S. Pham and S. Venkatesh, "Joint learning and dictionary construction for pattern recognition," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, June 2008, pp. 1–8.

[25] T. Chen and K. H. Yap, "Context-aware discriminative vocabulary learning for mobile landmark recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 23, no. 9, pp. 1611–1621, Sept 2013.

[26] S. Wilson, M. Srinivas, and C. Mohan, "Dictionary based action video classification with action bank," in *Digital Signal Processing (DSP), 2014 19th International Conference on*, Aug 2014, pp. 597–600.

[27] H. V. Nguyen, V. M. Patel, N. M. Nasrabadi, and R. Chellappa, "Kernel dictionary learning," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2012, pp. 2021–2024.

[28] A. Golts and M. Elad, "Linearized kernel dictionary learning," *IEEE Journal of Selected Topics in Signal Processing*, vol. 10, no. 4, pp. 726–739, June 2016.

[29] S. Y. Lee, J. Y. Sim, C. S. Kim, and S. U. Lee, "Correspondence matching of multi-view video sequences using mutual information based similarity measure," *IEEE Transactions on Multimedia*, vol. 15, no. 8, pp. 1719–1731, Dec 2013.

[30] A. Krause, C. Guestrin, A. Gupta, and J. Kleinberg, "Near-optimal sensor placements: maximizing information while minimizing communication cost," in *Information Processing in Sensor Networks, 2006. IPSN 2006. The Fifth International Conference on*, 2006, pp. 2–10.

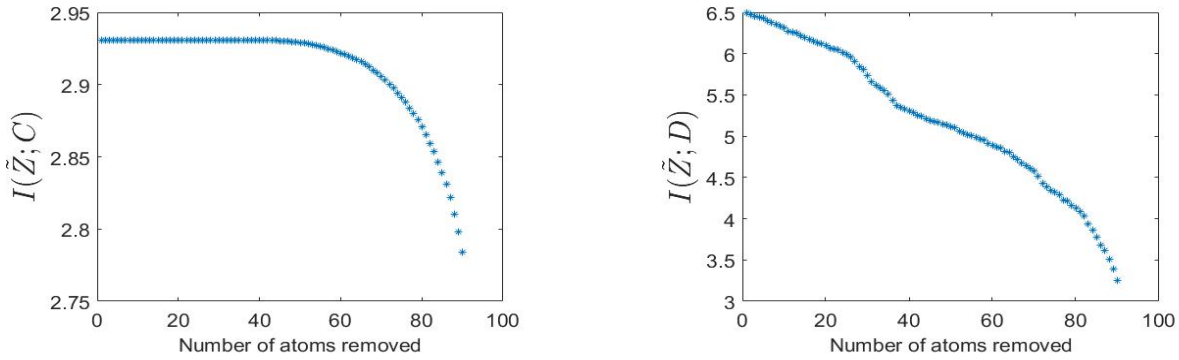
[31] A. Krause, A. Singh, and C. Guestrin, "Near-optimal sensor placements in gaussian processes: Theory, efficient algorithms and empirical studies," *J. Mach. Learn. Res.*, vol. 9, pp. 235–284, Jun. 2008.

[32] J. Liu and M. Shah, "Learning human actions via information maximization," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, June 2008, pp. 1–8.

[33] S. Lazebnik and M. Raginsky, "Supervised learning of quantizer codebooks by information loss minimization," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 7, pp. 1294–1309, July 2009.

[34] N. Tishby and N. Zaslavsky, "Deep learning and the information bottleneck principle," in *Information Theory Workshop (ITW), 2015 IEEE*, April 2015, pp. 1–5.

[35] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," in *Communication, Control, and Computing, The 37'th Allerton Conference on*, 1999, pp. 368–377.



(a) (b)

Fig. 6: $I(\tilde{\mathbb{Z}};\mathbb{C})$ and $I(\tilde{\mathbb{Z}};\mathbb{D})$ while removing dictionary atoms in UCF10.

[36] T. M. Cover and J. A. Thomas, *Elements of Information Theory, 2nd edition*. John Wiley & Sons, Inc., New Jersey, 2006.
[37] R. Blahut, "Computation of channel capacity and rate-distortion functions," *Information Theory, IEEE Transactions on*, vol. 18, no. 4, pp. 460–473, July 1972.
[38] S. Kullback and R. A. Leibler, "On information and sufficiency," *Ann. Math. Statist.*, vol. 22, no. 1, pp. 79–86, 03 1951.
[39] R. Rubinstein, M. Zibulevsky, and M. Elad, "Efficient implementation of the k-svd algorithm using batch orthogonal matching pursuit." Technical Report - CS, Technion, April 2008.
[40] J. J. Hull, "A database for handwritten text recognition research," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 16, no. 5, pp. 550–554, May 1994.
[41] A. Martinez and R. Benavente, "The AR face database," *CVC Technical Report*, no. 24, June 1998.
[42] M. D. Rodriguez, J. Ahmed, and M. Shah, "Action mach a spatio-temporal maximum average correlation height filter for action recognition," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*, June 2008, pp. 1–8.
[43] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: A local svm approach," in *Proceedings of the Pattern Recognition, 17th International Conference on (ICPR'04) Volume 3 - Volume 03*, ser. ICPR '04. Washington, DC, USA: IEEE Computer Society, 2004, pp. 32–36. [Online]. Available: http://dx.doi.org/10.1109/ICPR.2004.747
[44] 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.
[45] S. Sadanand and J. J. Corso, "Action bank: A high-level representation of activity in video," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, June 2012, pp. 1234–1241.
[46] A. Yao, J. Gall, and L. Van Gool, "A hough transform-based voting framework for action recognition," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, June 2010, pp. 2061–2068.
[47] M. Rodriguez, J. Ahmed, and M. Shah, "Action mach a spatio-temporal maximum average correlation height filter for action recognition," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, June 2008, pp. 1–8.
[48] L. Yeffet and L. Wolf, "Local trinary patterns for human action recognition," in *Computer Vision, 2009 IEEE 12th International Conference on*, Sept 2009, pp. 492–497.

Shyju Wilson received bachelor degree in computer science and engineering from Institution of Engineers (India), India and master degree in computer science and engineering from National Institute of Technology Rourkela, India in 2006 and 2009, respectively. He is currently pursuing PhD in computer science and engineering from Indian Institute of Technology Hyderabad, India. His research interests include computer vision and machine learning.

Dr. C. Krishna Mohan received PhD Degree in Computer Science and Engineering from Indian Institute of Technology Madras, India in 2007. He received the Master of Technology in System Analysis and Computer Applications from National Institute of Technology Surathkal, India in 2000. He received the Master of Computer Applications degree from S. J. College of Engineering, Mysore, India in 1991 and the Bachelor of Science Education (B.Sc.Ed) degree from Regional Institute of Education, Mysore, India in 1988. He is currently an Associate Professor with the Department of Computer Science and Engineering, Indian Institute of Technology Hyderabad, India. His research interests include video content analysis, pattern recognition, and neural networks.