*Article*

# Sparsity Based Locality–Sensitive Discriminative Dictionary Learning for Video Semantic Analysis

**Ben-Bright Benuwa [1, 2, a]*, Yongzhao Zhan [1, b], Benjamin Ghansah [2, c], Ernest K. Ansah [2, d] and Andriana Sarkodie [2, e]**

[1] School of Computer Science and Telecommunication Engineering, Jiangsu University, Xuefu Road 301 Jingkou District Zhenjiang province Jiangsu City, Zhenjiang 212013, China
[2] School of Computer Science, Data Link Institute P. O Box 2481 Tema Ghana
* [a] benuwa778 @gmail.com; +8615606101927, [b] yzzhang@ujs.edu.cn, [c] ben@datalink.edu.gh, [d] mayopia202@yahoo.com and [e] andrianaluv@yahoo.co.uk

**Abstract:** Dictionary Learning (DL) and Sparse Representation (SR) based Classifier have impacted greatly on the classification performance and has had good recognition rate on image data. In Video Semantic Analysis (VSA), the local structure of video data contains more vital discriminative information needed for classification. However, this has not been fully exploited by the current DL based approaches. Besides, similar coding findings are not being realized from video features with the same video category. Based on the issues stated afore, a novel learning algorithm, called Sparsity based Locality-Sensitive Discriminative Dictionary Learning(SLSDDL) for VSA is proposed in this paper. In the proposed algorithm, a discriminant loss function for the category based on sparse coding of the sparse coefficients is introduced into structure of Locality-Sensitive Dictionary Learning (LSDL) algorithm. Finally, the sparse coefficients for the testing video feature sample are solved by the optimized method of SLSDDL and the classification result for video semantic is obtained by minimizing the error between the original and reconstructed samples. The experiment results show that, the proposed SLSDDL significantly improves the performance of video semantic detection compared with the comparative state-of-the-art approaches. Moreover, the robustness to various diverse environments in video is also demonstrated, which proves the universality of the novel approach.

**Keywords:** Sparse Representation; locality information; Dictionary Learning; Video Semantic Analysis; Discriminative Function

## 1. Introduction

SR is an active research hotspot, particularly its application for signal reconstruction [1] and it performs well with video analysis and image classification based on experimental outcomes discussed in [2, 3]. DL on the other hand is to learn a good dictionary from training samples so that a given signal could be well represented hence the quality of the dictionary is very crucial for efficient sparse representation [4]. The dictionary could be determined by either using all the training samples as the dictionary to code the test samples (e.g. Locality Constrained Linear Coding (LLC) in [5]) or adopt a learned dictionary for the sparse representation for each training sample in the set (e.g. KSVD in [6], Fisher Discriminative Dictionary Learning (FDDL) [7]). Besides, Group centered sparse coding likened to rank minimization problem is used to measure the sparse coefficient of each group by estimating the values of each grouping in [8]. All the methods that adopts the first strategy use

training samples as the dictionary. Although they show good classification performance the dictionary might not be effective enough to represent the samples well because of noisy information that could have accompanied the original training samples and it may not also wholly exploit the discrimination information hidden in the training samples. The second category is also not suitable for recognition because it only requires that the dictionary is best expressed in the training samples with strict sparse representation. The above stated issues were addressed by the LSDL approach that incorporated a locality constraint into the objective function of the DL which ensures that over complete dictionary leaned is more representative. The problem with the traditional sparse representation methods is that, they cannot produce the same or similar results when the input features are from the same categorization.

The elementary approach to building the dictionary is by exhausting all the training samples which may result in the size of the dictionary being huge, and is subsequently unfavorable to the sparse solver in [9]. A lot of techniques such as the Method of Optimal Direction (MOD) in [10] which updates all the atoms simultaneously with fixed sparse codes using least square methods and an orthogonal matching pursuit for sparse coding, the K-SVD algorithm in [11] for learning a handful sized dictionary for sparse coding from the training data have recently been proposed. [12] also tried to resolve this issue by further iteratively updating the KSVD (which focuses on only the representational power or efficiency of the dictionary but not its discrimination abilities) trained dictionary grounded on the result of a linear classifier, by obtaining a dictionary that may be good for classification with representational power.

Other efforts along similar direction include [13] and [14], which use more sophisticated objective functions in dictionary optimization for the training phase so as to gain some discriminative power for the dictionary. The other discriminative dictionary learning method that uses discriminative and reconstructive error to learn the dictionary. [15] proposed a structured based Fisher Discriminative Dictionary Learning (FDDL) that improves the performance of pattern classification whose dictionary has a relation with the class labels and further exploited the discriminability in both representative coefficient and its residual in [16]. Recently [17] introduced the discriminative structured dictionary learning with hierarchical group sparsity that reduces the linear predictive error and discriminability of sparse codes using hierarchical group sparsity. [18] also proposed the discriminative structured dictionary learning for image classification that combines into the objective function, the discriminative properties of the reconstruction error, representative error and the classification error terms. In the same vein of improving the discriminability of dictionary learning algorithms, [19] proposed the structure adaptive dictionary leaning for sparse representation based classification, [20] proposed the learning of dictionary discriminatively for group sparse representation, and discriminative dictionary learning for face recognition with low ranked regularization by [21]. [5]proposed the Locality-constrained Linear Coding for Image Classification (LLC) built on the theories of Linear Spatial Pyramid Matching using Sparse Coding for Image Classification [22]. LLC could achieve a smaller reconstruction error with multiple code reconstruction and local smoothness with respect to sparsity. More recently, there has been an advancement in sparse dictionary learning for video semantic as proposed by [2], a video semantic detection method based on Locality-sensitive Discriminant Sparse Representation and Weighted KNN (LSDSR-WKNN), to have better category discrimination on the sparse representation of video semantic concepts. [27] proposed a Latent Label Consistent DL (LLCDL) that solves a unified

framework that minimizes simultaneously, the structured discrimination sparse codes approximation, the classification error and the structured sparse reconstruction to boost representation and classification performances. The authors of LLCDL algorithm, previously introduced the joint label consistence Embedded and Dictionary Learning (JEDL) algorithm [28] based on LC-KSVD. The technique concurrently learns discriminative sparse codes, sparse reconstruction, classification errors and code approximation that enable linear combination of signals for the sparse code auto-extractor with sparse codes during classification. More so, vital similarity information of signals is achieved by enforcing sparse coefficients to be discriminative amongst different classes. However, it implemented the KSVD computation approach to obtain the dictionary columns which renders it computationally expensive.

Based on the aforementioned, this paper seeks to enhance the power of discrimination of sparse representation features by encoding group sparse codes of samples from the same category. Thus proposes a Sparsity Based Locality –Sensitive Discriminative Dictionary Learning for (SLSDDL) for VSA, which is more efficient and has superior numeric stability. The SLSDDL algorithm enables the adjustment of the dictionary adaptively using the differences between the reconstructed dictionary atoms and the training samples with the locality adaptor. More so, the implementation of the locality adaptor at both DL and Sparse Coding stages increases the efficiency of the algorithm of the locality and similarity of the dictionary atoms. The dictionary information could also be exploited by the proposed algorithm with the introduction of a discriminant loss function during the dictionary learning stage, to obtain a more discriminative information and enhance the classification ability of spares representation features. The dictionaries learned could realistically represent video semantic more, and thus gives a better representation of the samples belonging the sparse coefficients.

The main contributions of this paper are as enumerated below:

(1) A discriminative loss function is utilized, giving an optimal dictionary for addressing sparse representation issues and optimized sparse coefficients for reconstructing the dictionary, so as to enhance the discriminative classification of sparse representation features.

(2) The introduction of the group sparsity also enables efficient reconstruction of samples most especially when the size gets large.

(3) A locality sensitive discriminative dictionary learning and sparse representation based on group sparsity is developed for video semantic detection that encodes video features originally. Thus features are sparsely encoded with video semantic detection for better preservation of the dictionary hence an improvement in the accuracy of video concept detection.

The rest of the paper is organized into the following sections; Section 2 presents a review of related works, section, Section 3 discusses the proposed algorithms, Experimental results are presented in section 4. Finally, section 5 outlines the main conclusions and recommendations.

## 2. Related Works

### *2.1 Sparse Representation Algorithms*

SR for signal acquirement and the overall techniques included coding, sampling, compression, transmission, and decoding, and is one of the utmost essential ideologies in the field of signal

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 January 2018 doi:10.20944/preprints201801.0101.v1



processing theorem which tells us that, a signal that has been sampled can be impeccably reconstructed from an arrangement of samples if the sampling rate surpasses twice the maximum frequency of the original signal [23]. It utilizes an over-complete dictionary to linearly reconstruct a data case for a given dictionary. Suppose that the example is from space $R^d$, and thus all the examples concatenated to form a matrix, denoted as $X \epsilon R^{dxm}$. If any sample can be approximately represented by a linear amalgamation of dictionary $D$ and the number of the samples is larger than the dimension of samples in $D$, i.e. $m > d$, dictionary $D$ is referred to as an over-complete dictionary. A signal is said to be compressible if it is a sparse signal in the original or transformed domain when there is no information or energy loss during the course of transformation. Generally, from the matrix stated above, if d<m, and y $\epsilon R^d$ then we define the linear system of equations as

$$y = Z\beta \quad (1)$$

The sparsest representation solution can be acquired by solving equation (1) with the $l_0$-norm minimization constraint [24]. Thus problem (1) can be converted to the following optimization problem:

$$\hat{\beta} = \text{argmin} ||\beta||_0 \qquad s.t. y = Z\beta \quad (2)$$

Problem (2) is called the $k$-sparse approximation problem. Because real data always contains noise, representation noise is unavoidable in most cases. Thus the original model (1) can be revised to a modified model with respect to small possible noise by denoting

$$y = Z\beta + s \quad (3)$$

Where $s \epsilon R^d$ refers to representation noise and is bounded as$||s||_2 \leq \varepsilon$. With the presence of noise, the sparse solutions of problems (2) can be approximately obtained by resolving the following optimization problems:

$$\hat{\beta} = \text{argmin} ||\beta||_0 \qquad s.t. ||y - Z\beta||_2^2 \leq \varepsilon \quad (4)$$

Or

$$\hat{\beta} = \text{argmin}||y - Z\beta||_2^2 \qquad s.t. \ ||\beta||_0 \leq \varepsilon \quad (5)$$

Furthermore, according to the Lagrange multiplier theorem, a proper constant $\lambda$ exists such that equations (4) and (5) are equivalent to the following unconstrained minimization problem with a proper value of $\lambda$.

$$\hat{\beta} = L(\beta, \lambda) = \text{argmin}||y - Z\beta||_2^2 + \lambda||\beta||_0 \quad (6)$$

where $\lambda$ refers to the Lagrange multiplier associated with $||\beta||_0$.

*2.2. Locality-sensitive Sparse Representation*

Test examples in SR are usually represented by the dictionary atoms that may actually not be neighboring it for signal or data reconstructions. This may render it incongruous for holding data locality, and hence could lead to poor recognition rates. In the LSDL method, the locality constraints of data locality were added in the phase of the dictionary learning and sparse coding, so that the test sample can be well represented by the neighboring dictionary atoms. The optimized function for LSDL is stated as follows:

$$\min_{D,X} ||A - DX||_F^2 + \lambda \sum_{i=1}^{N} ||p_i \ \Theta x_i||_2^2 \quad (7)$$

$$s.t. 1^T x_i = 1 \ \forall i = 1, \dots, N$$

where $p_i \epsilon R^{Kx1}$ is locality adapter and $\lambda$ is the scalar parameters which measures the locality constraint. The LSDL method does not consider the class information, and therefore we design the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 January 2018 doi:10.20944/preprints201801.0101.v1



discriminating loss function using class information to enhance the sparse representation based classification, in order to improve the accuracy of image data representation or video analysis.

## 3. Proposed Method

The proposed algorithm for SLSDDL based on the locality-sensitive adaptor, and a discriminative lost function centered on group SR is incorporating into the objective function of locality sensitive discriminative dictionary learning method as stated in equation (7) and detailed as explained below. The proposed method could achieve optimal dictionary and further enhance the power of discriminability of sparse codes as well as computational cost. As a result, features from representation coefficient might not be capable of achieving the same coding results hence the assumption is made that samples should be encoded as similar sparse coefficients in the SR based data detection with the objective of enhancing the power of discrimination in SR based classifiers. Assume that $A = [a_1, a_2, \dots, a_n] = [A_1, A_2, \dots, A_k] \epsilon R^{dxn}$ denotes n training samples from k classes where the column vector $a_1$is the sample $i$ $(i = 1, \dots, n)$, and the submatrix $A_j$ consisting of column vectors from class $j$ $(j = 1, \dots, k)$. If there are M atoms each in the dictionary $D = [D_1, D_2, \dots, D_M] \epsilon R^{dxM}$, an over complete dictionary for SR with $(M \leq n)$, then coding coefficients of $A$ over $D$ could be denoted by $X = [x_1, x_2, \dots, x_L] \epsilon R^{Mxn}$. Now considering image or video features of the same category of the representation coefficient or the representation solution with a class information, which is very imperative for classification, our SLSDDL framework is modeled as:

$$\min_{D,X} || A_i^j - DX||_F^2 + \lambda \frac{N}{2} ||P_i^j \Theta x_i^j||_2^2 + g(x_i^j) \quad (8)$$

$$s.t. 1^T x_i^j = 1 \ \forall i = 1, \dots, N$$

where we have the locality adaptor as $P \epsilon R^{Kx1}$ , $A \epsilon R^{dxn}$ is the training sample, $D \epsilon R^{dxM}$ is the dictionary, $x_i \epsilon R^{Mx1}$ is the sparse coefficient vector and the discriminative lost function for group sparsity is $g(x_i^j)$. Here the locality adaptor $P_i^j$ uses the l2-norm function and its kth element in $P_i^j$ is given by $P_{ij}^j = \left|\left|a_i^j - d_i^j\right|\right|$ and the symbol $\Theta$ represents element-wise multiplication, $g(x_i^j)$ is the proposed discriminant loss function to enforce group sparsity discriminability as defined in equation (9) below, the shift variant constant $1^T x_i^j$ enforces the coding result of x to remain the same although the origin of the data coordinate may be shifted as indicated in [25] and the regularization parameter controlling the reconstruction error and the sparsity is $\lambda$.

Bearing in mind cases where sample features from the same category could be having similar sparse codes, we propose a discriminant loss function based on a group sparse codding of the sparse coefficients for the purposes of enhancing the power of discrimination of input signals or samples in SR. The group structure enforced on the coefficient vector, ensures that the variables in the same group either tend to be zero or nonzero simultaneously. Thus enhances classification hence better classification results. consequently, samples from same category are compacted and the ones from different categories are separated. The discriminant loss function, $g(x_i^j)$ based on sparse coefficients is explained below:

$$g(x_i^j) = \left|\left|\left((1 - \frac{1}{N_j}) x_i^j - \frac{1}{N_j} \sum_{l \epsilon (1, N_j) \Lambda l \neq i} x_i^j\right)\right|\right|_2^2 + ||x_i^j||_2^2 \quad (9)$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 January 2018 doi:10.20944/preprints201801.0101.v1



Where $\left|\left|\left(\left(1-\frac{1}{N_j}\right)x_i^j-\frac{1}{N_j}\sum_{l\epsilon(1,N_j)\Lambda l\neq i}x_i^j\right)\right|\right|_2^2+||x_i^j||_2^2$ is the within class-similar term enforcing group sparsity by computing the representations whose nonzero coefficients are further divided into groups, $N_j$ is the number of samples of the representation coefficient $x_i^j$ belonging to the class $j$ and $N$ is the number of training samples. The term $||x_i^j||_2^2$ combined with $||X||_2^2$ could make equation (8) more stable based on theorem of [26]. It is worth noting that, group sparsity is enforced in the way $x_i^j$ is calculated. The encoding of $x_i^j$ using the proposed method significantly, is computationally less expensive, compared with the existing SR classification approaches as a result of the smaller dictionary size. Thus $x_i^j$ is much more easier to obtain than $x_i$, and it enforces the capture of dependencies amongst video samples of the same categorization. With $\eta$ set to 1 for simplicity, equation (8) can be reformulated as

$$\min_{D,X}||A_i^j-DX||_F^2+\lambda_1\sum_{i=1}^N||P_i^j\,\Theta x_i^j||_2^2+\lambda_2\left|\left|\left(\left(1-\frac{1}{N_j}\right)x_i^j-\frac{1}{N_j}\sum_{l\epsilon(1,N_j)\Lambda l\neq i}x_i^j\right)\right|\right|_2^2+||x_i^j||_2^2 \quad (10)$$

$$s.t.\,1^Tx_i^j=1\ \ \forall i=1,\dots,N$$

The proposed method as stated equation (10) is implemented by enforcing data locality with the locality adaptor being used to measure the distance between the test sample $y$ and each column of $D$. Note that $D=[D_1,D_2,\dots,D_M]$ represents all the training samples in the feature space. The vector P, the dissimilarity vector in equation (10) is implemented to suppress the corresponding weight and also penalizes the distance between the test and each training samples in the feature space. Furthermore, it should however be made known that the resulting coefficients in our SLSDDL formulation may not be fully sparse with regards to l2 – norm, but is seen as sparse because the representation solutions only have few significant values with most being zero. The test samples and their neighboring training samples in feature space are encoded when the problem of equation (9) is being minimized and the resulting coefficients X still sparsed because as $P_{ij}^j$ gets large, $x_{ij}^j$ shrink to be zero. Hence most coefficients get zero with just some few having significant values.

*3.1 Optimization*

The objective function in equation (10) could be divided into two sub-problems by updating X with D fixed and updating D with X fixed since it is not convex as a whole. To find the desired dictionary D and the sparse coefficient X, an alternative optimization is implemented iteratively adopting the theories of [15, 27-29].

3.1.1 Updating X with D Fixed, Sparse Coding Stage

When the coefficient matrix X is being updated with the dictionary D being fixed, the objection function of equation (10) is reduced to the coding problem below:

$$\langle X\rangle=\min_X||A_i^j-DX||_F^2+\lambda_1\sum_{i=1}^N||P_i^j\,\Theta x_i^j||_2^2+\lambda_2\sum_{i=1}^N g(x_i) \quad (11)$$

The X in equation (12) could be addressed on class bases. Thus $x_i^j$ corresponding to class $i$ could be derived as:

$$\langle x_i^j\rangle=\min_{x_i^j}||A_i^j-Dx_i^j||_2^2+\lambda_1\sum_{i=1}^N||P_i^j\,\Theta x_i^j||_2^2+\lambda_2\left|\left|\left(\left(1-\frac{1}{N_j}\right)x_i^j-\frac{1}{N_j}\sum_{l\epsilon(1,N_j)\Lambda l\neq i}x_i^j\right)\right|\right|_2^2+||x_i^j||_2^2$$

$$s.t.\,1^Tx_i^j=1 \quad (12)$$

Eq. (13) can be rewritten by adopting Lagrange function $L(x_i^j,\mu)$ as follows:

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 January 2018 doi:10.20944/preprints201801.0101.v1



$$L(x_i,\mu) = ||\,A_i - Dx_i||_2^2 + \lambda_1 \sum_{i=1}^{N} ||P_i^j\ \Theta x_i^j||_2^2 + \lambda_2 \left|\left|\left((1-\frac{1}{N_j})\,x_i^j - \frac{1}{N_j}\sum_{l\epsilon(1,N_j)\Lambda l\neq i}\ x_i^j\right)\right|\right|_2^2 + ||x_i^j||_2^2 + \mu(1^T x_i^j - 1) \quad (13)$$

Eq. (14) can be converted as:

$$L(x_i^j,\mu) = x_i^{jT} C x_i + \lambda_1 x_i^{jT}\, diag\ (P_i^j)^2 x_i^j + \lambda_2((x_i^j)^T Y x_i^j + (x_i^j)^T I\ x_i^j) + \mu(1^T x_i^j - 1) \quad (14)$$

where $C = (a_i^j 1^T - D)^T\ (a_i^j 1^T - D)$, $diag\ (P_i^j)^2$ is a diagonal matrix whose nonzero elements are the square of the entries of $P_i^j$, $Y = ((\left(I - \frac{1}{N_j}\right)I - \frac{1}{N_j}\sum_{l\epsilon(1,N_j)\Lambda l\neq i}\ x_i^j\ (x_i^j 1^T)^T\left((I - \frac{1}{N_j})I - \frac{1}{N_j}\sum_{l\epsilon(1,N_j)\Lambda l\neq i}\ x_i^j\ (x_i^j 1^T)\right)$.

Let $\partial\frac{L(x_i^j,\mu)}{\partial x_i^j} = 0$, we have

$$\theta x_i^j + \mu 1 = 0 \quad (15)$$

where $\theta = 2(C + \lambda_1 diag\left(P_i^j\right)^2 + \lambda_2(Y+1))$, Once Eq. (16) is pre-multiplied by $1^T\,\theta^{-1}$, we get $\mu = -(1^T\,\theta^{-1}1)^{-1}$. Then through substituting $\mu$ into (15), the analytical solution of (13) is obtained as

$$\widetilde{x_i} = (C + \lambda_1 diag\left(P_i^j\right)^2 + \lambda_2(Y+1)^{-1}1 \quad (16)$$

Furthermore, the normalized $x_i^j$ of $\widetilde{x_i^j}$ is given as

$$x_i^j = \widetilde{x_i^j}/(1^T\widetilde{x_i^j}) \quad (17$$

3.1.2. Updating D with the Sparse Coefficient Matrix X fixed; the dictionary update stage.

During the dictionary update stage, the objective function in equation (10) is reduced to equation (18) as stated below:

$$\langle D\rangle = \min_D ||\,A - DX||_F^2 + \lambda_1 \sum_{i=1}^{N} ||P_i^j\ \Theta x_i^j||_2^2 + \lambda_2 \sum_{i=1}^{N} g(x_i^j) \quad (18)$$

Let $F(D) = \min_D ||\,A - DX||_F^2 + \lambda_1 \sum_{i=1}^{N} ||P_i^j\ \Theta x_i^j||_2^2 + \lambda_2 \sum_{i=1}^{N} g(x_i)$, we take the partial derivatives of F(**D**) with respect to $\mathbf{d}_k$ for $k = 1,2\ldots K$, which gives

$$\frac{\partial F}{\partial d_k} = \sum_{I=1}^{N} -2x_{ik}^j\left(a_i^j - Dx_i^j\right) - 2\lambda_1(x_{ik}^j)^2(x_i^j - {d_k}^j)\Theta\left(\frac{\partial {d_k}^j}{\partial d_k}\right) \quad (19)$$

Let $\frac{\partial F}{\partial d_k} = 0$, and $\gamma = \left(\frac{\partial {d_k}^j}{\partial d_k}\right)$, we obtain $UD^T = V$, where the matrices $U \in R^{KxK}$ and $V \in R^{Kxm}$ are as detailed below

$$U = \sum_{i=1}^{N}\begin{pmatrix} (1+\lambda_1||\,\gamma_i^j||_1){x_i^j}_1^2 & x_{i1}^j x_{i2}^j & \cdots & x_{i1}^j x_{iK}^j \\ x_{i1}^j x_{i2}^j & (1+\lambda_1||\,\gamma_i^j||_1){x_i^j}_2^2 & \cdots & x_{i2}^j x_{iK}^j \\ \vdots & \vdots & \ddots & \vdots \\ x_{i1}^j x_{iK}^j & x_{i2}^j x_{iK}^j & \cdots & (1+\lambda_1||\,\gamma_i^j||_1){x_i^j}_K^2 \end{pmatrix} \quad (20)$$

$$V = \sum_{i=1}^{N}\begin{pmatrix} x_{i1}^j(1+\lambda_1 x_{i1}^j)((a_i^j)^T) \\ x_{i2}^j(1+\lambda_1 x_{i2}^j)((a_i^j)^T) \\ \vdots \\ x_{iK}^j(1+\lambda_1 x_{iK}^j)((a_i^j)^T) \end{pmatrix} \quad (21)$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 January 2018 doi:10.20944/preprints201801.0101.v1



The optimal X and D could be gotten by alternating between sparse coding and dictionary learning until an optimal iteration, or error of reconstruction less than a threshold value is gotten. The algorithm for SLSDDL L is summarized in Algorithm 1. As an extension of the LSDL method, the proposed SLSDDL method seeks to obtain a discriminative information of all training samples. Therefore, the LSDL method is espoused to initialize the sparse coefficients of all the training samples.

**Algorithm 1: Sparsity Locality –sensitive Discriminative Dictionary Learning Algorithm.**

Input: Training samples A, Initial dictionary D and parameters $\lambda_1, \lambda_2$

Output: The learned dictionary D

(a) Initialize the coding coefficient matrix X and i ($i \leftarrow 1$)

(b) Optimize the solution by updating the sparse coding matrix X and the dictionary D. Update X by fixing D and updating each $x_i = (i = 1, \dots, k)$ according to equation (18) on class bases by solving for $x_i^j$ and then do $X`(:, i) \leftarrow x_i^j$

(c) If $i = N$, update X, which means $X \leftarrow X`$ else do $i = i + 1$ and go to (b)

(d) Initialize $k_1$ and $k_2$ respectively with $k_1 \leftarrow 1$ and $k_2 \leftarrow 1$

(e) Solve equation (20) for the solution of $U(k_1, k_2)$

(f) Go to (g) if $k_2 = K$ else do $k_2 = k_2 + 1$ and return to (e)

(g) Compute $V(k_1, :)$ from equation (21)

(h) Fix X and update D using equation (18)

(i) If $k_2 = K$, execute $D` = (U/V)^T$ and go to (j) otherwise execute $k_1 \leftarrow k_1 + 1$, $k_2 \leftarrow 1$ and return to (e)

(j) Compute the reconstruction error. If the error value is less than the current minimum values, update the dictionary D`` on the minimum error which is $D`` \leftarrow D`$.

(k) Iteratively return to step (b) until convergence or the reconstruction error is less than the threshold error, update D and the algorithm ends or go to (a).

*3.2 Classification Scheme*

For a given test sample y, the representation coefficients on dictionary D is given as:

$$x = arg \min_x ||y - Dx||_2^2 + \lambda ||P_i^j \, \Theta x_i^j||_2^2 \quad (22)$$

The test sample can finally be classified after obtaining *D*,

Where $\lambda$ is a scalar constant that measures sparsity, $\Theta$ represents element wise multiplicative symbol, $\boldsymbol{p} = [p_1, p_2, \dots, p_k]^T$ and every entry is represented $p_k = \|\mathbf{y} - \boldsymbol{d}_k\|_2$ . We utilized the Lagrange multiplier to obtain the solution for equation (24) as explained below;

$$L(x, \mu) = \min_x ||y - Dx||_2^2 + \lambda ||P_i^j \, \Theta x_i^j||_2^2 + \mu(1^T x_i^j - 1) \quad (23)$$

Equation (23) is easily simplified as

$$L(x, \mu) = x^T C` x_i + \lambda x^T diag\,(P)^2 x + \mu(1^T x - 1) \quad (24)$$

where $C` = (y1^T - D)^T (y1^T - D)$

Let $\partial \frac{L(x,\mu)}{\partial x} = 0$, which gives us

$$\theta x + \mu 1 = 0 \quad (25)$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 January 2018 doi:10.20944/preprints201801.0101.v1



where $\theta = 2(C` + \lambda\, diag\,(P)^2)$ , Once Eq. (25) is pre-multiplied by $1^T\,\theta^{-1}$ , we get $\mu = -(1^T\,\theta^{-1}1)^{-1}$. Then through substituting $\mu$ into (26), the analytical solution of (26), $\tilde{x}$ is obtained as

$$\tilde{x} = (C + \lambda\, diag\,(P)^2 +)^{-1}1 \tag{26}$$

Sparse coding coefficient $x$is the then obtained by normalizing $\tilde{x}$ and is determined as

$$x = \tilde{x}/(\,1^T\tilde{x}) \tag{27}$$

Using the analysis in [30], we get the sparse coding efficient associated with test sample y by solving equation (26) and (27), after which the residual $r_j(y) = ||\,y - D_i x_i||_2^2$ is calculated. Where $x_i$ is the sparse coding coefficient on the $i^{th}$ class and the test sample $y$ is finally classified into the class that has the minimum residual as formulated below;

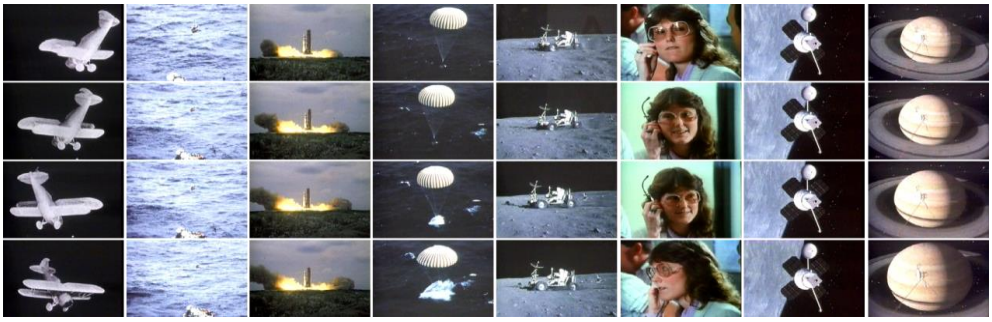
$$class\,(y) = arg\,\min_j r_j\,(y) \tag{28}$$

## 4. Experimental Results and Analysis

This section details the experimental results and their analysis to demonstrate the effectiveness of the proposed method.

### *4.1. Video shot preprocessing*

In the experiments, ordered samples clustering centered on artificial immune in [31] was adopted to extract the static image frames from each original video data utilizing the Video key-frame extraction approach. Afterwards, features consisting of the 5-dimensional radial Tchebichef moment [32], 6-dimensional gray level co-occurrence matrix (GLCM) [33], 81-dimensional HSV color histogram [34], and 30-dimensional multi-scale local binary pattern (LBP) [35] features from the key-frame were extracted. The details of these features could be referred from the paper [2]. Figure 1 below show some key-frames of the three video datasets implemented.



(a) Some key-frames from OV video set


(b). Part key-frames from the TRECVID 2012 video set

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 January 2018 doi:10.20944/preprints201801.0101.v1



(c). Part key-frames from the YouTube video set

**Figure 1**: Some key-frames from video shots for the respective datasets

***4.2.*** *Database Selection and Video Analysis*

Experiments are performed on natural images to demonstrate the efficient performance of our methods in this section. We analyze and compare the performance of our proposed algorithm with most typical algorithms such as the KSVD [11], FDDL, LLC[5] and LSDL[3] algorithms. Out of the many collections of datasets for video categorization and classification, three of them are being used in our experimental set up or evaluations. The public datasets being used are the TRECVID 2012 video dataset, OV video dataset, and YouTube video dataset.

**The TRECVID 2012 videos dataset** has airplane, baby, building, car, dog, flower, instrumental_musician, mountain, scene_text, and speech as its video semantic models with each class containing 60 data samples, 50 of the samples were randomly as training samples, and the remaining as test samples. **The YouTube video dataset** comprise of basketball, biking, diving, golf_swing, horse_riding, soccer_juggling, swing, tennis, trampoline_jumping, volleyball and walking as semantic videos with each class containing 70 data samples, out of which 60 were randomly selected as training samples, and the rest as test samples, and the semantic concepts of **the OV dataset** contain aircraft, sea, rocket, parachute, road, face, satellite and star. Each class consists of 70 samples. For each class, 60 samples are randomly selected as training samples, and the rest ones are testing samples. The proposed approach considered the L2 norm locality adaptor. The experimental results were realized by twentyfold cross validation in which training samples are selected randomly and the remaining as the testing samples from the various video datasets.

*4.3 Parameter Selection*

Several parameters including $\lambda$, the classification parameter, $\lambda_1$, a positive weighing parameter for the locality sensitive constraint and $\lambda_2$ for the discriminative constraint were utilized by the proposed SLSDDL and the classification error scheme, with varied values assigned depending on the dataset being used. More so, the classification parameter, $\lambda$ made a little or no effect on the output of the experimental results hence was set to 0.01 in all experiments. The recognition rates of the proposed SLSDDL approach (for the optimization phase) with varying values of $\lambda_1$ and $\lambda_2$ on TRECVID dataset are as depicted by figure 2 below. It could be seen from the figure that, the optimum results were obtained for recognition when $\lambda_1$ is 0.01 and $\lambda_2$ is 0.007 respectively with $\lambda_2$ and $\lambda_1$ values being fixed for each dataset by twentyfold cross validation.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 January 2018 doi:10.20944/preprints201801.0101.v1



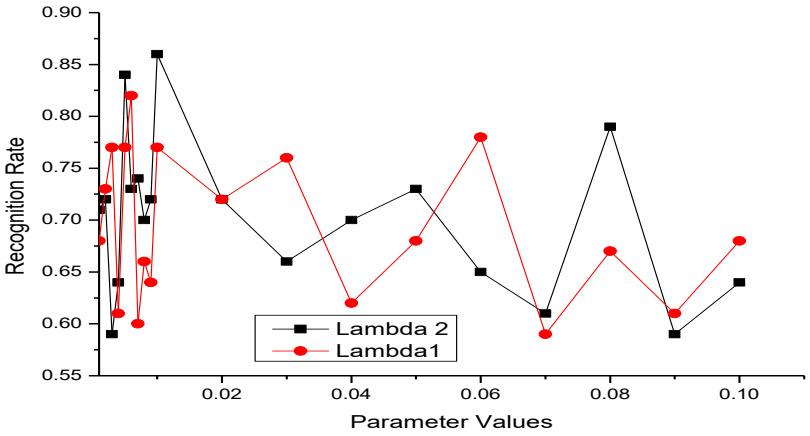



**Figure 2**: Recognition rate Vrs Variations of parameter values on TRECVID Dataset

**Table 1:** parameter values implemented on the various data sets

| Dataset | Parameter Values used | |
|---|---|---|
| | $\lambda_1$ | $\lambda_2$ |
| TRECVID 2012 | 0.007 | 0.01 |
| OV | 0.006 | 0.03 |
| YOUTUBE | 0.5 | 0.01 |

***Table 1*** *above indicates the various parameter selections with their respective datasets. Based on the aforementioned values, an experiment was conducted to validate the performance of our proposed algorithms and the findings are detailed below. Different datasets have different parameter settings because the structure of each dataset differs hence the different parameter values.*

*4.3. Results and Discussions*

4.3.1. Experimental results and analysis TRECVID 2012 video dataset

The recognition rates of video semantic detection of the proposed SLSDDL with varying video features for the initial dictionaries are as indicated in table 2 below. It could be seen that, the SLSDDL gives an optimum performance when the size of the dictionary is 122 x 400 and thus resulted in the highest accuracy of semantic analysis and hence is chosen.

**Table 2**. Classification results on different dictionary atoms of TRECVID 2012

| Number of the dictionary atoms | 10 | 20 | 30 | 40 | 50 |
|---|---|---|---|---|---|
| Accuracy Rate (%) | 62 | 72.7 | 77.1 | 84.83 | 82.6 |

The recognition rates of LLC, FDDL, LSDL, KSVD and SLSDDL are as shown in Table 3. It could be seen that the proposed SLSDDL comparatively obtained the highest recognition rates than the other recognition approaches. As indicated by Fig 3, the average recognition rates of various video semantic detection methods associated with each class on the TRECVID video dataset is also presented, which clearly suggests that the proposed method is best among all the comparative approaches on most category. The recognition rate of the proposed method on the TRECVID dataset for instance outperformed KSVD by 8.23% and 7.86% against LSDL which confirms the effectiveness

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 January 2018 doi:10.20944/preprints201801.0101.v1



of the proposed method. Hence, the proposed SLSDDL approach could effectively improve best the accuracy rates of TRECVID video semantic feature detections.

**Table 3**. The recognition rate of different video semantic analysis algorithms of TRECVID 2012

| Comparative methods | LLC | FDDL | LSDL | KSVD | SLSDDL |
|---|---|---|---|---|---|
| Accuracy(%) | 65.89 | 64.49 | 76.97 | 76.60 | 84.83 |

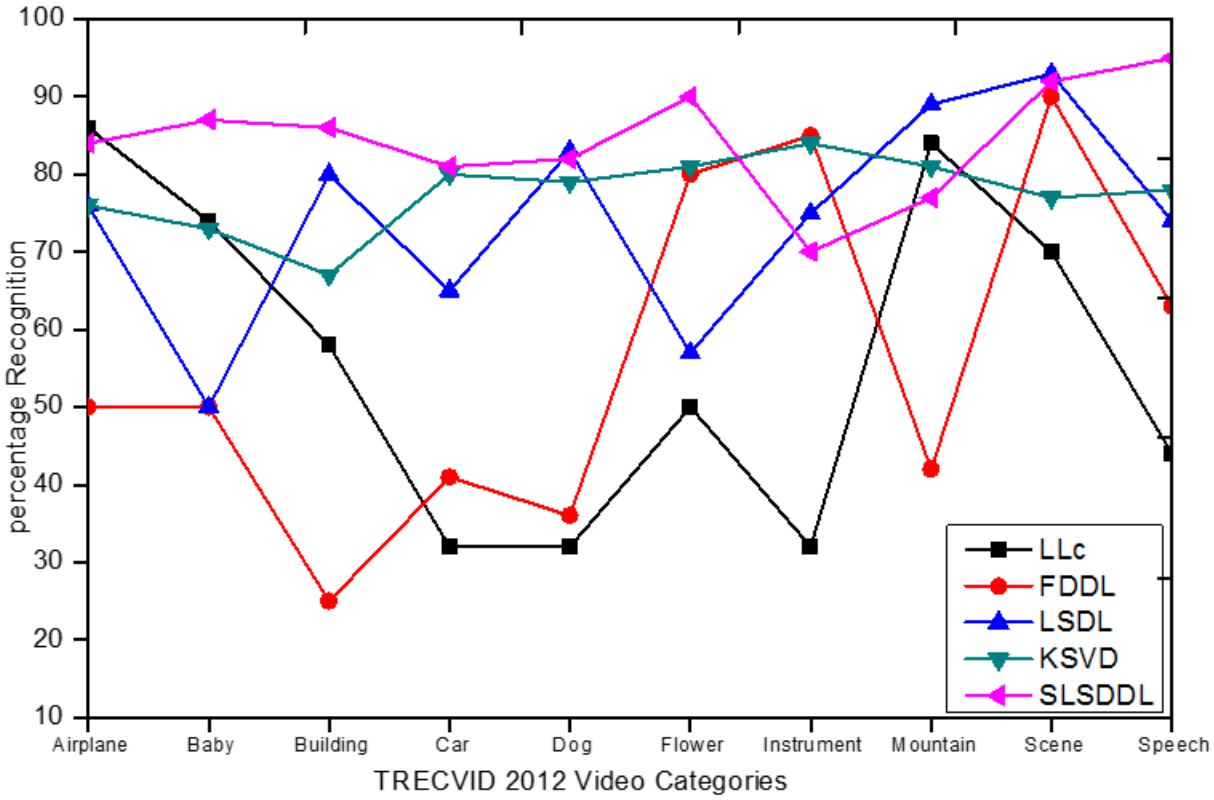



**Fig 3**: Recognition rate of TRECVID 2012 videos with different algorithms

4.3.2. Experimental results and analysis OV video dataset

The recognition rates of video semantic detection of the proposed SLSDDL with varying video features for the initial dictionaries are as indicated in table 4 below. It could be seen that, the SLSDDL gives an optimum performance when the size of the dictionary is 122 x 500 and thus resulted in the highest accuracy of semantic analysis and hence is chosen.

**Table 4**. Classification results on different dictionary atoms of OV dataset

| Number of the dictionary atoms | 10 | 20 | 30 | 40 | 50 | 60 |
|---|---|---|---|---|---|---|
| Accuracy (%) | 63.7 | 74.6 | 77.1 | 83 | 84.97 | 82.8 |

The recognition rates of LLC, FDDL, LSDL, KSVD and SLSDDL are as shown in Table 5. It could be seen that the proposed SLSDDL comparatively obtained the highest recognition rates than the other recognition approaches. As indicated by Figure 4, the average recognition rates of various video semantic detection methods associated with each class on the OV video dataset is also presented,

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 January 2018 doi:10.20944/preprints201801.0101.v1



which clearly suggests that the proposed method is best among all the comparative approaches on most categories. The recognition rate of the proposed method on the OV dataset for instance outperformed KSVD by 10.86% and 10.41% against LSDL, which goes to confirm the effectiveness of the proposed method. Hence, the proposed SLSDDL approach could effectively improve best the accuracy rates of OV video semantic feature detections.

**Table 5.** The recognition rate of different video semantic analysis algorithms of OV dataset

| Comparative methods | LLC | FDDL | LSDL | KSVD | SLSDDL |
|---|---|---|---|---|---|
| Accuracy(%) | 69.44 | 56.85 | 74.56 | 74.11 | 84.97 |

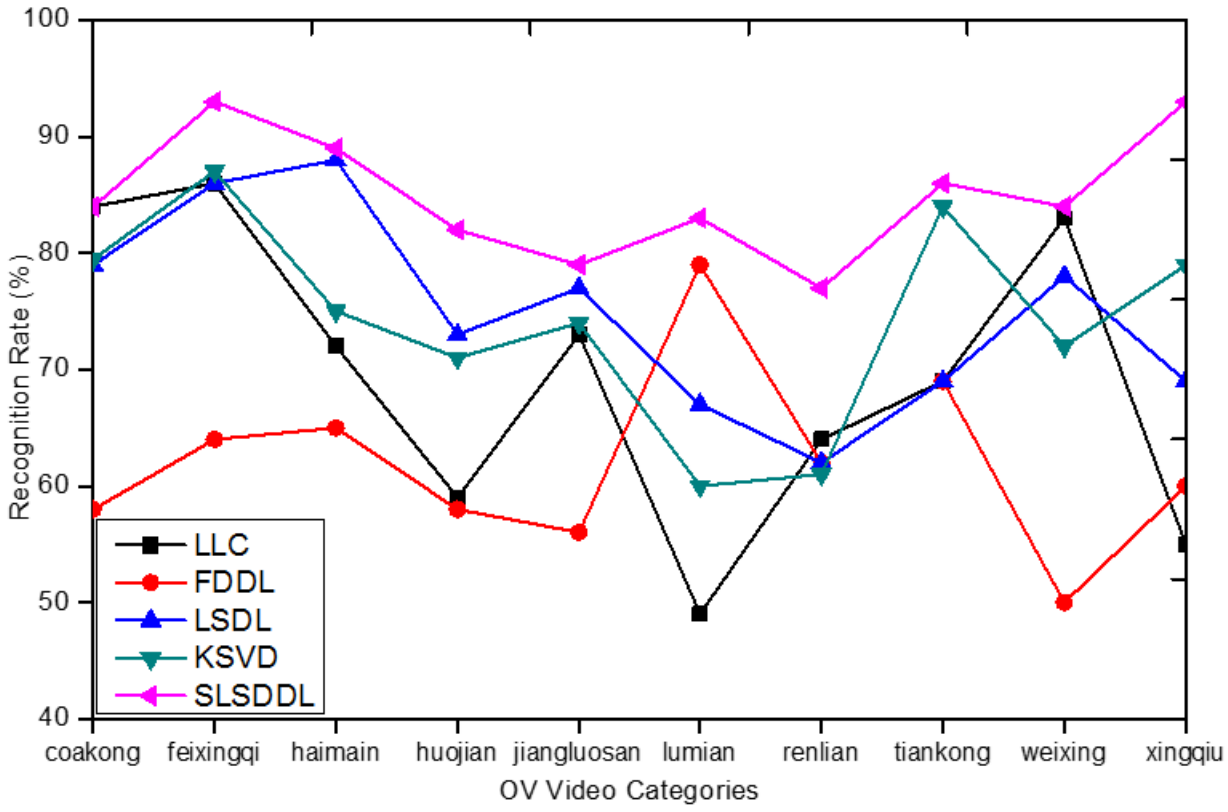



**Fig 4:** Recognition rate of OV videos with different algorithms

4.3.3 Experimental results and analysis YouTube video dataset

The recognition rates of video semantic detection of the proposed SLSDDL with varying video features for the initial dictionaries are as indicated in table 6 below. It could be seen that, the SLSDDL gives an optimum performance when the size of the dictionary is 122 x 550 and thus resulted in the highest accuracy of semantic analysis and hence is chosen.

**Table 6.** Classification results on different dictionary atoms of YouTube

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 January 2018 doi:10.20944/preprints201801.0101.v1



| The number of the dictionary atoms | 10 | 20 | 30 | 40 | 50 | 60 |
|---|---|---|---|---|---|---|
| Accuracy (%) | 60.51 | 72.64 | 78.73 | 81.82 | 83.45 | 80.91 |

The recognition rates of LLC, FDDL, LSDL, KSVD and SLSDDL are as shown in Table 7. It could be seen that the proposed SLSDDL comparatively obtained the highest recognition rates than the other recognition approaches. As indicated by Figure 5, the average recognition rates of various video semantic detection methods associated with each class on the YouTube video dataset is also presented, which clearly suggests that the proposed method is best among all the comparative approaches on most categories. The recognition rate of the proposed method on the YouTube dataset also outperformed KSVD by 8.9% and 9.36% against LSDL which confirms the effectiveness of the proposed method. Hence, the proposed SLSDDL approach could effectively improve best the accuracy rates of YouTube video semantic feature detections.

**Table 7**. The recognition rate of different video semantic analysis algorithms of YouTube

| Comparative methods | LLC | FDDL | LSDL | KSVD | SLSDDL |
|---|---|---|---|---|---|
| Accuracy(%) | 66.31 | 64 | 74.09 | 74.55 | 83.45 |

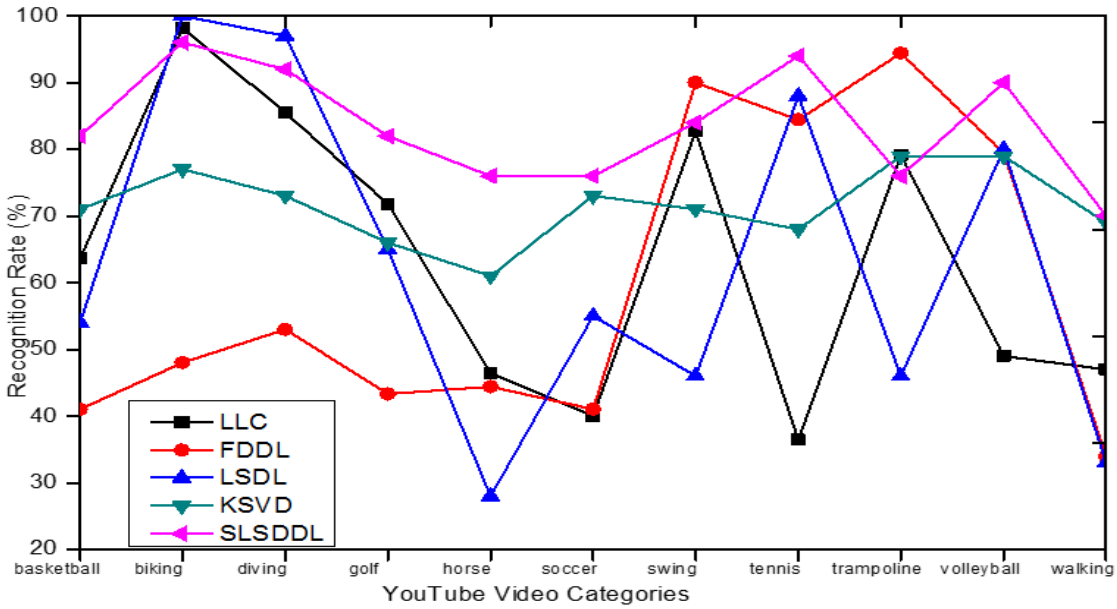



**Fig 5:** Recognition rate of YouTube videos with different algorithms

## 5. **Conclusion and Recommendations**

In this paper, a Locality Sensitive Discriminative Dictionary Learning based on sparsity is proposed. The proposed approach learns that video samples from the same category should be encoded as similar sparse coefficients in the process of video semantic detection based on sparse representation, so as to enhance the power of discrimination of sparse representation features. The paper also introduced into the structure of the locality sensitive algorithm, a lost function centered on sparse coefficients to optimize the dictionary. The SLSDDL enhances the power of discrimination

of sparse representation features based on the principles of Fishers for better dictionary optimization as a result of the constraints and also to improve video semantic classification.

Based on the experimental results, it could be seen that, the proposed SLSDDL algorithm for video semantic analysis is more effective and outperforms the other state-of-the-art approaches such as LLC, FDDL, LSDL and KSVD. Despite a superior results demonstrated by the proposed SLSDDL on the TRECVID, YouTube and OV video datasets, there is still the need to improve on the execution time and further improve the power of discrimination hence we plan use deep learning discussed in [36] for the extraction of the video features and also introduce kernel into the structure in our future work.

**Acknowledgments:** This work was supported in part by National Natural Science Foundation of China (Grant Nos.~61170126, Grant Nos.~61502208), the Natural Science Foundation of the Jiangsu Higher Education Institutions of China (Grant No. 14KJB520007), China Postdoctoral Science Foundation (Grant No. 2015M570411), Natural Science Foundation of Jiangsu Province of China (Grant No. BK20150522) and Research Foundation for Talented Scholars of JiangSu University (Grant No. 14JDG037).

**Authors Contributions:** All Authors had equation contributions in the paper and have read and approved of the final manuscript.

**Conflict of Interest:** The Authors declare no conflict of interest whatsoever.

## References

1. Zhang, Z., et al., *A survey of sparse representation: algorithms and applications.* IEEE access, 2015. **3**: p. 490-530.
2. Zhan, Y., et al., *A video semantic detection method based on locality-sensitive discriminant sparse representation and weighted KNN.* Journal of Visual Communication and Image Representation, 2016. **41**: p. 65-73.
3. Wei, C.-P., et al., *Locality-sensitive dictionary learning for sparse representation based classification.* Pattern Recognition, 2013. **46**(5): p. 1277-1287.
4. Meng, F., et al., *A Sparse Dictionary Learning-Based Adaptive Patch Inpainting Method for Thick Clouds Removal from High-Spatial Resolution Remote Sensing Imagery.* Sensors, 2017. **17**(9): p. 2130.
5. Wang, J., et al. *Locality-constrained linear coding for image classification*. in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. 2010. IEEE.
6. Zhang, Q. and B. Li. *Discriminative K-SVD for dictionary learning in face recognition*. in *Computer Vision and Pattern Recognition*. 2010.
7. Zheng, H. and D. Tao, *Discriminative dictionary learning via Fisher discrimination K-SVD algorithm*. 2015: Elsevier Science Publishers B. V. 9-15.
8. Zha, Z., et al., *Analyzing the group sparsity based on the rank minimization methods.* arXiv preprint arXiv:1611.08983, 2016.
9. Wright, J., et al., *Robust face recognition via sparse representation.* Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2009. **31**(2): p. 210-227.
10. Engan, K., 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*. 1999. IEEE.

11. Aharon, M., M. Elad, and A. Bruckstein, *< img src="/images/tex/484. gif" alt="\ rm K">-SVD: An Algorithm for Designing Overcomplete Dictionaries for Sparse Representation.* Signal Processing, IEEE Transactions on, 2006. **54**(11): p. 4311-4322.
12. Pham, D.-S. and S. Venkatesh. *Joint learning and dictionary construction for pattern recognition*. in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. 2008. IEEE.
13. Mairal, J., et al. *Discriminative learned dictionaries for local image analysis*. in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. 2008. IEEE.
14. Mairal, J., et al. *Supervised dictionary learning*. in *Advances in neural information processing systems*. 2009.
15. Yang, M., et al. *Fisher discrimination dictionary learning for sparse representation*. in *Computer Vision (ICCV), 2011 IEEE International Conference on*. 2011. IEEE.
16. Yang, M., et al., *Sparse representation based fisher discrimination dictionary learning for image classification.* International Journal of Computer Vision, 2014. **109**(3): p. 209-232.
17. Xu, Y., et al., *Discriminative structured dictionary learning with hierarchical group sparsity.* Computer Vision and Image Understanding, 2015. **136**: p. 59-68.
18. Wang, P., et al., *Discriminative structured dictionary learning for image classification.* Transactions of Tianjin University, 2016. **22**: p. 158-163.
19. Chang, H., M. Yang, and J. Yang, *Learning a structure adaptive dictionary for sparse representation based classification.* Neurocomputing, 2016. **190**: p. 124-131.
20. Sun, Y., et al., *Learning discriminative dictionary for group sparse representation.* IEEE Transactions on Image Processing, 2014. **23**(9): p. 3816-3828.
21. Li, L., S. Li, and Y. Fu. *Discriminative dictionary learning with low-rank regularization for face recognition*. in *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*. 2013. IEEE.
22. Yang, J., et al. *Linear spatial pyramid matching using sparse coding for image classification*. in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. 2009. IEEE.
23. Benuwa, B.B., et al. *A Comprehensive Review of Particle Swarm Optimization*. in *International Journal of Engineering Research in Africa*. 2016. Trans Tech Publ.
24. Donoho, D.L. and M. Elad, *Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ1 minimization.* Proceedings of the National Academy of Sciences, 2003. **100**(5): p. 2197-2202.
25. Yu, K., T. Zhang, and Y. Gong. *Nonlinear learning using local coordinate coding*. in *Advances in neural information processing systems*. 2009.
26. Zou, H. and T. Hastie, *Regularization and variable selection via the elastic net.* Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2005. **67**(2): p. 301-320.
27. Jiang, Z., Z. Lin, and L.S. Davis. *Learning a discriminative dictionary for sparse coding via label consistent K-SVD*. in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. 2011. IEEE.
28. Cai, S., et al. *Support vector guided dictionary learning*. in *European Conference on Computer Vision*. 2014. Springer.
29. Feng, Z., et al., *Joint discriminative dimensionality reduction and dictionary learning for face recognition.* Pattern Recognition, 2013. **46**(8): p. 2134-2143.
30. Harandi, M. and M. Salzmann. *Riemannian coding and dictionary learning: Kernels to the rescue*. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015.
31. Yongzhao, Z., W. Manrong, and K. Jia, *Video keyframe extraction using ordered samples clustering based on artificial immune.* Journal of Jiangsu University (Natural Science Edition), 2012. **2**: p. 017.

32. Mukundan, R. *Radial Tchebichef invariants for pattern recognition*. in *TENCON 2005 2005 IEEE Region 10*. 2005. IEEE.
33. Haralick, R.M. and K. Shanmugam, *Textural features for image classification.* IEEE Transactions on systems, man, and cybernetics, 1973(6): p. 610-621.
34. Kim, M., *Efficient histogram dictionary learning for text/image modeling and classification.* Data Mining & Knowledge Discovery, 2016: p. 1-30.
35. Ojala, T., M. Pietikainen, and T. Maenpaa, *Multiresolution gray-scale and rotation invariant texture classification with local binary patterns.* IEEE Transactions on pattern analysis and machine intelligence, 2002. **24**(7): p. 971-987.
36. Benuwa, B.B., et al. *A Review of Deep Machine Learning*. in *International Journal of Engineering Research in Africa*. 2016. Trans Tech Publ.