Coherent and non-coherent dictionary for action recognition

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

***Abstract*—In this paper, we propose sparsity based coherent and non-coherent dictionary for action recognition. First, the input data is divided into different clusters and number of clusters depends on number of action categories. Within each cluster, we seek data items of each action category. If number of data items exceeds threshold in any action category, these items are labeled as *coherent*. In a similar way, all coherent data items from different clusters form a coherent group of each action category and data which are not part of the coherent group belong to *non-coherent* group of each action category. These coherent and non-coherent groups are learned using K-SVD (K- Singular Value Decomposition) dictionary learning. Since the coherent group has more similarity among data, only few atoms need to be learned. In non-coherent group, there is a high variability among the data items. So we propose an orthogonal projection based selection to get optimal dictionary in order to retain maximum variance in the data. Finally, the obtained dictionary atoms of both groups in each action category are combined and then updated using Limited Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) optimization algorithm. The experiments are conducted on challenging datasets HMDB51 and UCF50 with action bank features and achieve comparable result using this state of art feature.**

***Index Terms*—coherency, sparse coding, dictionary learning.**

I. Introduction

IT is a challenging task to obtain compact and discriminative [1] representation for enormous amount of visual data for classification or recognition. Sparsity based approach has been extensively [2][3] used in many areas like action recognition, object tracking, video super resolution, face hallucination, face recognition etc. At the origin of this model lies a simple linear system of equations. A full row rank matrix, $D \in R^{m \times K}$, with $m < K$, having infinite number of solutions in $D\mathbf{x} = \mathbf{y}$. This is also called underdetermined system of linear equations. In our context, $D = [\mathbf{d_1 d_2} \dots \mathbf{d_K}]$ is called as *dictionary* where $\mathbf{d_i} \in R^m$ is referred as *dictionary atom*. In order to obtain unique solution, *regularization* is a familiar way where a function $J(\mathbf{x})$ determines the kind of solution we may obtain,

$$\min_{\mathbf{x}} J(\mathbf{x}) \quad \text{subject to } D\mathbf{x} = \mathbf{y}.$$

The l_2 norm is widely used best choice for $J(\mathbf{x})$, because of it's mathematical simplicity to obtain the solution. But it is realized that this is not a best solution because it is dense in nature. The quest for sparse solution ended up in exploring l_1 or l_0 norm solution. The l_0 norm, number of non-zero coefficients, is more attractive because it provides the sparsest solution. Despite proved it's uniqueness and global optimality [4], finding the solution is still NP hard problem. The orthogonal matching pursuit (OMP) algorithm is a practical approach to obtain l_0 solution. This is called *sparse coding*. The input signals can be learned into the dictionary D, which is known as *dictionary learning*. There are many dictionary learning algorithms such as MOD [5], K-SVD [6] etc. Unlike in signal reconstruction, for machine learning application, we look for discriminative dictionary atoms, not necessarily over-complete.

The sparse based approach is very powerful and successfully used in many machine learning applications. Wright et al. [7] used sparse coding for face recognition and reconstruction error for classification which yielded better result. Yang et al. [8] introduce Fisher discrimination criterion to get discriminative dictionary atoms. The extension of [8] presents a support vector based discriminative dictionary learning model [9]. In [10], 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 [11] by adding more discriminative terms. This exploits unlabelled data by sparse representation and solves specific task like classification.

In [12], the input data is divided into clusters and learned into local dictionaries. The global dictionary is trained from atoms of these local dictionaries. This helps to reduce computational time and increase performance in image processing applications. In our work, we treat coherent and non-coherent data items seperately and learned them as separate dictionaries. Daniele et al. [13] learned dictionary with low mutual coherence by sparse representation followed by dictionary update using iterative projections and rotations. The main characterestics of dictionary learning is the mutual coherene among dictionary atoms. In order to reduce this mutual coherence, Mansour Nejati et al. [14] propose a coherence regularized dictionary learning which explicitly imposes a coherence regularizer to learn the dictionary. In [15], fixed coherence dictionary is made by maximizing pairwise decorrelations of atoms in the dictionary.

The outline of the approach is shown in figure 1. In this work, we show how coherency among data can be exploited using the sparse based approach. For non-coherent data, an orthogonal projection based selection is used to obtain discriminative dictionary atoms. Then the obtained dictionary atoms are updated to enhance the recognition performance. The section II describes coherent and non-coherent dictionary

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

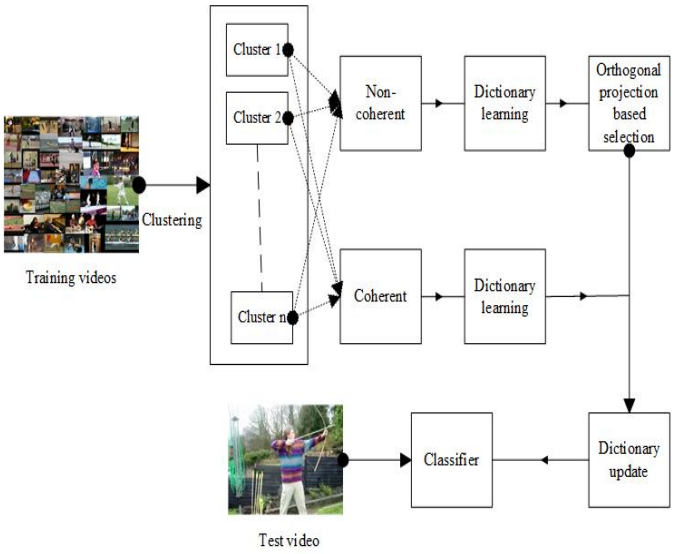


Fig. 1: Block diagram of the proposed approach for action recognition. Dotted arrow indicates that cluster may or may not have coherent or non-coherent group.

learning for each action category. The updation of obtained dictionary is explained in section III and the experimental study is discussed in section IV. Finally, section V concludes the entire work.

II. Coherent and non-coherent dictionary learning

Initially, we cluster input data $Y = [\mathbf{y_1}, \mathbf{y_2}, \ldots, \mathbf{y_N}] \in R^{m \times N}$ into n clusters using k-means and number of clusters, i.e. n, depends on number action categories in the dataset. We seek natural coherency by grouping training data into n clusters. These clusters are $Y = \{C_1, C_2, \ldots, C_n\}$, where C_i denotes i^{th} cluster. In each cluster, we look for coherent and non-coherent data items which are to be learned as separate dictionaries. Each coherent and non-coherent group are learned by K-SVD dictionary learning which alternates sparse coding and dictionary update. For sparse coding, it uses OMP which looks for minimum l_0 norm and T is the sparsity constraint for this optimization problem:

$$\underset{D,X}{\operatorname{argmin}} \|Y - DX\|_F^2 \quad \text{subject to} \quad \forall i \ \ \|\mathbf{x_i}\|_0 \leq T, \qquad (1)$$

each sparse vector $\mathbf{x_i} \in R^K$ in sparse matrix $X = [\mathbf{x_1 x_2} \ldots \mathbf{x_N}]$ represents corresponding y_i in the input data $Y = [\mathbf{y_1}, \mathbf{y_2}, \ldots, \mathbf{y_N}] \in R^{m \times N}$. The notation $\|.\|_F$ and $\|.\|_0$ denotes frobenius norm and l_0 norm, respecively. After getting X, the dictionary D will be updated using singular value decomposition (svd) unlike in MOD which uses pseudo inverse for dictionary update. Sections II-A and II-B detail how to group and learn coherent and non-coherent data items.

A. Learning coherent actions

In each cluster C_i, the data are grouped based on their action categories. For grouping, there is a constraint for minimum number of data items to group. If it satisfies the constraint, then these data items are labeled as *coherent*. Similarly, coherent data of particular action category, say c, are grouped from all clusters to form coherent group G^c_{cohe} as:

$$G^c_{cohe} = [G^c_1 G^c_2 \ldots G^c_i \ldots G^c_n], \quad 1 \leq c \leq p, i \in \{1, 2, \ldots, n\}$$

where p and n denote number of classes and clusters, respectively. The coherent group, G^c_i, may not exist in all clusters because of the minimum grouping constraint. Then each G^c_{cohe} is learned into the dictionary D^c_{cohe} using K-SVD dictionary learning. Coherent group contains similar data items, so that we can exploit sparsity by learning into few dictionary atoms. The advantage of this grouping is that only few dictionary atoms are enough to approximate the input data which leads to the reduction of overall dictionary size and computational time. If there is more coherency in the input data, we can obtain very compact dictionary while achieving good recognition performance. All other data items which are not part of the coherent group belong to *non-coherent* group which is treated in a different manner as discussed in the next section.

B. Learning non-coherent actions

The non-coherent group has high variability among data items, because it is scattered in many clusters. So, we need to learn more dictionary atoms compared to coherent group discussed in the subsection II-A. The selection of minimum number of discriminative dictionary atoms effectively is a challenging task. Likewise in coherent group, non-coherent items in each action category c are grouped into G^c_{ncohe} and learned into the dictionary $D^c_{cohe} = [\mathbf{d_1 d_2} \ldots \mathbf{d_k}]$, where $\mathbf{d_i} \in R^m$ represents dictionary atom. The most variant dictionary atoms are to be selected from this dictionary. For this purpose, we propose orthogonal projection based selection to include maximum variability among the dictionary atoms. Here, one data item is to be picked randomly from D^c_{ncohe} and make it as residual vector $\mathbf{r}$. Now the current D^c_{ncohe} has only $(k-1)$ dictionary atoms. Initially, the closest dictionary atom from D^c_{ncohe} to the residual $\mathbf{r}$ to be found by projecting $\mathbf{r}$ onto the dictionary atoms. For this purpose, error $e(i)$ is computed as,

$$e(i) = min_{z_i} \|\mathbf{d_i} z_i - \mathbf{r}\|_2^2 \qquad \forall \mathbf{d_i} \in D^c_{ncohe}, \qquad (2)$$

and the optimal choice $z_i = \frac{\mathbf{d_i} \cdot \mathbf{r}}{\|\mathbf{d_i}\|_2^2}$, where $\mathbf{d_i} \cdot \mathbf{r}$ denotes dot product between $\mathbf{d_i}$ and $\mathbf{r}$. Now the closest vector $\mathbf{d_{i_1}}$ to $\mathbf{r}$ can be found by looking $e(i_1) \leq e(i)$ for all $\mathbf{d_i}$ in D^c_{ncohe}. Then this $\mathbf{d_{i_1}}$ is removed from D^c_{ncohe} and added to empty set A. After getting $\mathbf{d_{i_1}}$, the residual $\mathbf{r}$ needs to be updated as $\mathbf{r} = \mathbf{r} - \mathbf{d_{i_1}} \mathbf{z_{i_1}}$ and normalized to unit norm. The updated $\mathbf{r}$ is orthogonal to $\mathbf{d_{i_1}}$. In the next iteration, we can find $\mathbf{d_i}$ which is closest to updated residual $\mathbf{r}$ using the same procedure. In each iteration, one vector from D^c_{ncohe} is chosen and added to set A . At the t^{th} iteration, A contains t selected vectors viz. $\{\mathbf{d_{i_1}}, \mathbf{d_{i_2}}, \ldots, \mathbf{d_{i_t}}\}$ and then the updated residual becomes orthogonal to all dictionary atoms in A. So, the residual can be updated as,

$$\mathbf{r} = \mathbf{r} - A(A^T A)^{-1} A^T \mathbf{r}, \qquad (3)$$

where, with some abuse of notation, we use A to refer set of dictionary atoms as well as matrix of dictionary atoms. The set A usually contains few atoms, so it does not take much computational time to calculate inverse of the matrix while updating residual.

The non-coherent dictionary after selecting most variant dictionary atoms denoted as $D^{c^*}_{ncohe}$ which is cascaded to D^c_{cohe} to obtain final dictionary of action category c, ie., $D^c = [D^c_{cohe} D^{c^*}_{ncohe}]$. Then the dictionary D^c to be updated.

III. Update the dictionary of each action

In each action category c, two dictionaries are obtained viz. D^c_{cohe} and $D^{c^*}_{ncohe}$. These two dictionaries are cascaded to form dictionary of each action category c, ie., D^c. Here, we update the dictionary D^c using input data of the action category c. An unconstrained non-linear optimization algorithm L-BFGS (Limited memory BFGS) [16] has been used to update the dictionary. This approximates Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm using a limited amount of memory. It is based on gradient projection method. The matrix $Y^c \in R^{m \times N^c}$ be the input data of action category c and N^c denotes number of input data belong to the same action category. The sparse matrix $X^c \in R^{k \times N^c}$ can be obtained using OMP algorithm. The input data Y^c is approximated using dictionary D^c and sparse matrix X^c. Now the approximation becomes, $Y^c \approx D^c X^c$.

The cost function and gradient matrix are to be computed for the update. So the cost function J can be written as,

$$J = \frac{1}{2N^c} \|D^c X^c - Y^c\|_F^2 + \frac{\lambda}{2N^c} \sum_i \sum_j d_{ij}^2, \qquad (4)$$

where d_{ij} is the element in i^{th} row and j^{th} column in the matrix D^c and the regularization parameter λ is determined by empirically. The vectorized form of gradient matrix is given by,

$$\frac{\partial J}{\partial D^c} = \frac{1}{N^c}(D^c X^c - Y^c) X^{cT} + \frac{\lambda}{N^c} D^c. \qquad (5)$$

All updated dictionaries of each action categories are cascaded to form final dictionary $D = [D^1 D^2 \dots D^n]$. The dictionary D is used for the experiment which is discussed in the next section.

1) Reconstruction error: In this experiment, reconstruction error is used to evaluate the efficacy of the learned dictionaries and OMP algorithm to obtain sparse vector of the test vector $\mathbf{y}$. We use two ways to get reconstruction error. One way is to calculate reconstruction error for each dictionary D^c separately. In this, sparse vector $\mathbf{x^c}$ is obtained using corresponding D^c and $\mathbf{y}$. For the second way, dictionaries from each action category, D^c, are cascaded to form dictionary D and obtain sparse vector $\mathbf{x}$ using D and $\mathbf{y}$. Finally, equal weightages are given to both result by late fusion. The action category of minimum reconstruction error will be assigned to test input $\mathbf{y}$, ie.,$\min_c \|\mathbf{y_i} - D^c \mathbf{x^c}\|^2$, where the sparse vector $\mathbf{x^c}$ contains coefficients corresponding to the atoms in the dictionary D^c.

IV. Experiment

We demonstrate our proposed approach on two challenging datasets viz. UCF50 [17] and HMDB51 [18]. The state of art feature Action bank [19] has been used to represent each action videos. Action bank comprises many individual action detectors which constitutes mid-level representation of action data. The non-coherent groups are learned into dictionary of larger size as compared to coherent groups to maintain high variability in non-coherent group. However, in this experiment, coherent and non-coherent group are learned into dictionary size of 10% and 20% of input data, respectively. The sparsity constraint T is 10 and value of λ for dictionary update is 1. Moreover, the grouping constraint, ie., minimum number of coherent data items, is taken as 10 in this experiment.

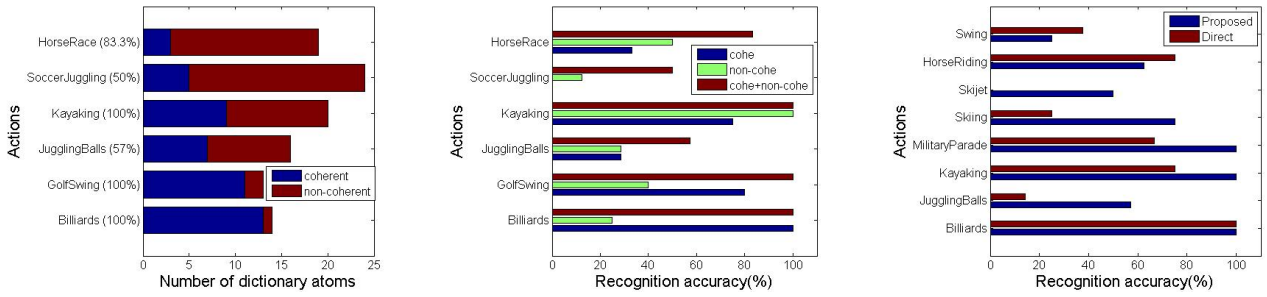
A. UCF 50 dataset

This is one of the challenging data set for action recognition. There are 50 action categories and 6950 action videos in all categories. There are 25 persons performing actions in each category. Here the input data is grouped into 50 clusters and each cluster is analysed for coherent and non-coherent data items. The obtained coherent and non-coherent dictionary are cascaded and updated as discussed in previous sections. The experimental results are taken based on Leave-One-Person-Out strategy. In figure 2(a), coherent dictionary atoms are dominating non-coherent in Golf swing and Billiards. In this case, it provides good recognition accuracy with small number of dictionary atoms which shows if coherency is more in any action category, we can have better recognition while reducing overall dictionary size. Figure 2(b) shows recognition accuracies of coherent and non-coherent dictionary separately and both. It can be observed that both coherent and non-coherent are contributing for the overall recognition accuracy. Our proposed approach is compared with direct dictionary learning in the figure 2(c) which clearly indicates splitting the data into coherent and non-coherent is worth for enhancing the recognition performance. The same number of atoms are used for both proposed and direct dictionary learning. Figure 4 gives the performance of action recognition before and after the dictionary update. It can be observed that the dictionary update clearly enhances the overall recognition performance.

B. HMDB 51 dataset

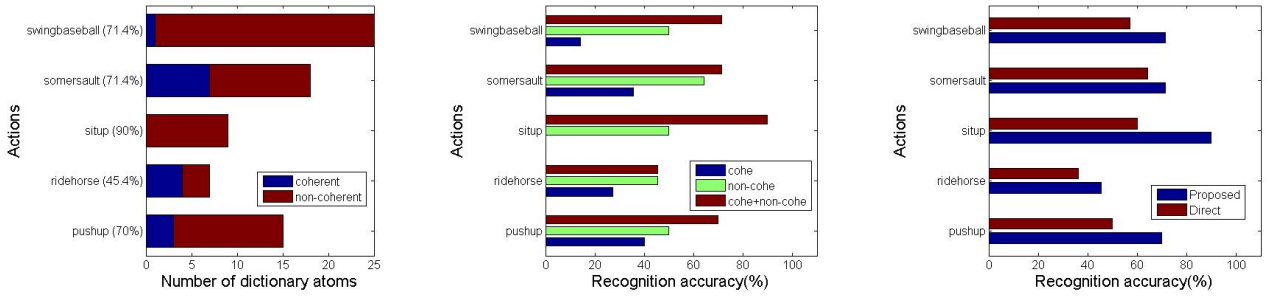
This is another challenging dataset. It has 51 action categories and 6766 action videos. The input data are clustered into 51 clusters and results are obtained based on 10-fold cross validation. In this, most of the data items are grouped in non-coherent group as shown in figure 3(a), this indicates the high variability in the dataset. As compared to coherent atoms, the non-coherent atoms are contributing more to the overall recognition performance as seen in figure 3(b). So the selection of non-coherent dictionary atoms is vital to this kind of challenging dataset. Figure 3(c) compares our proposed method with direct dictionary learning, which shows advantage of the proposed method.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/LSP.2017.2690461, IEEE Signal Processing Letters



(a) (b) (c)

Fig. 2: UCF 50: comparing recognition performances of (a) no. of coherent and non-coherent dictionary atoms (b) coherent, non-coherent and combining both dictionary (c) proposed method and direct dictionary learning



(a) (b) (c)

Fig. 3: HMDB 51: comparing recognition performances of (a) no. of coherent and non-coherent dictionary atoms (b) coherent, non-coherent and combining both dictionary (c) proposed method and direct dictionary learning

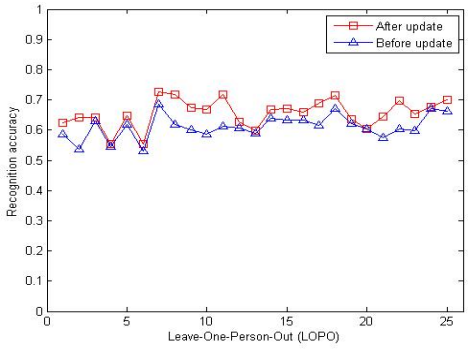


Fig. 4: Performance comparison: before and after dictionary update in UCF50. The x-axis denotes one of the 25 person taken as test data in LOPO evaluation.

C. Comparison to the state of the art

Table I compares our method with other state of art results in datasets UCF50 and HMDB51. Sadanand et al. [19] and shyju et al. [20] used same action bank features as ours and achieved performance of 57.9% and 59.3%, respectively. We improved this benchmark results using actionbank around 7%. Solmaz et al. [21] and Kliper el al. [22] achieved better results than ours, but they used different features like GIST3D, MIP etc.

For HMDB51, our proposed method achieved better results than all other state of art results. Sadanand et al. [19] got 26.9%, but we achieved remarkably good result 35.8% using action bank feature. Solmaz et al. [21] and Kliper el al. [22] achieved 29.2% and 29.17%, respectively. We improve it further by around 6%.

TABLE I: Comparison of our results to the state of art

Method	*Features*	*UCF50 (%)*	*HMDB51 (%)*
Sadanand et al.[19]	Action bank	57.90	26.90
Shyju et al. [20]	Action bank	59.30	23.62
Solmaz et al. [21]	GIST3D	73.70	29.20
Kliper-Gross et al. [22]	MIP	72.68	29.17
Proposed Method	Action bank	**66.30**	**35.8**

V. Conclusion

Here our aim is to deal coherent data and non-coherent data separately. The experiments prove the efficacy of the proposed approach by giving good recognition performances. To exploit sparsity, coherent group can be learned into few dictionary atoms. If the input data has more coherent data, it can drastically reduce the overall dictionary size and computational time. In this way, the dictionary can be optimized effectively while keeping discriminant information for generalization. For non-coherent group, there is high variability among the data, we use orthogonal projection based selection to get optimum discriminative dictionary atoms which is an efficient way to sustain high variability in the non-coherent data. This is a challenging task and we look more robust method in future work.

REFERENCES

[1] 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.

[2] T. Peleg, Y. C. Eldar, and M. Elad, "Exploiting statistical dependencies in sparse representations for signal recovery," *IEEE Transactions on Signal Processing*, vol. 60, no. 5, pp. 2286–2303, May 2012.

[3] R. Rubinstein, A. M. Bruckstein, and M. Elad, "Dictionaries for sparse representation modeling," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 1045–1057, June 2010.

[4] M. Elad, *Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing*, 1st ed. Springer Publishing Company, Incorporated, 2010.

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

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

[7] 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.

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

[9] 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.

[10] 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

[11] 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.

[12] S. Mukherjee and C. S. Seelamantula, "A divide-and-conquer dictionary learning algorithm and its performance analysis," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2016, pp. 4712–4716.

[13] D. Barchiesi and M. D. Plumbley, "Learning incoherent dictionaries for sparse approximation using iterative projections and rotations," *IEEE Transactions on Signal Processing*, vol. 61, no. 8, pp. 2055–2065, April 2013.

[14] M. Nejati, S. Samavi, S. M. R. Soroushmehr, and K. Najaran, "Coherence regularized dictionary learning," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2016, pp. 4717–4721.

[15] B. Mailh, D. Barchiesi, and M. D. Plumbley, "Ink-svd: Learning incoherent dictionaries for sparse representations," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2012, pp. 3573–3576.

[16] R. H. Byrd, P. Lu, J. Nocedal, and C. Zhu, "A limited memory algorithm for bound constrained optimization," *SIAM Journal on Scientific Computing*, vol. 16, no. 5, pp. 1190–1208, 1995.

[17] K. K. Reddy and M. Shah, "Recognizing 50 human action categories of web videos," *Mach. Vision Appl.*, vol. 24, no. 5, pp. 971–981, Jul. 2013. [Online]. Available: http://dx.doi.org/10.1007/s00138-012-0450-4

[18] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "Hmdb: a large video database for human motion recognition," in *International Conference on Computer Vision (ICCV)*, 2011.

[19] S. Sadanand and J. Corso, "Action bank: A high-level representation of activity in video," in *Computer Vision and Pattern Recognition (CVPR), IEEE Conference on*, June 2012, pp. 1234–1241.

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

[21] B. Solmaz, S. M. Assari, and M. Shah, "Classifying web videos using a global video descriptor," *Mach. Vision Appl.*, vol. 24, no. 7, pp. 1473–1485, Oct. 2013. [Online]. Available: http://dx.doi.org/10.1007/s00138-012-0449-x

[22] O. Kliper-Gross, Y. Gurovich, T. Hassner, and L. Wolf, "Motion interchange patterns for action recognition in unconstrained videos," in *Proceedings of the 12th European Conference on Computer Vision - Volume Part VI*, ser. ECCV'12. Berlin, Heidelberg: Springer-Verlag, 2012, pp. 256–269. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-33783-3_19