Stable Discriminative Dictionary Learning via Discriminative Deviation

Nazar Khan
University of Central Florida
nazar@cs.ucf.edu

Marshall F. Tappen
University of Central Florida
mtappen@cs.ucf.edu

Abstract

Discriminative learning of sparse-code based dictionaries tends to be inherently unstable. We show that using a discriminative version of the deviation function to learn such dictionaries leads to a more stable formulation that can handle the reconstruction/discrimination trade-off in a principled manner. Results on Graz02 and UCF Sports datasets validate the proposed formulation.

1. Introduction

Sparse coding offers a generalization of vocabulary[1] based bag-of-words approaches to recognition of objects. Whereas a standard bag-of-words approach represents an input signal as an optimally sparse vector based on the closest vocabulary word, sparse coding allows representing signals using a linear combination of a few dictionary items. In order to improve upon the ultimate goal of better recognition/classification, multiple approaches attempt to compute dictionaries in a discriminative manner.

One approach for obtaining discriminative dictionaries is to compute a large overcomplete dictionary in a reconstructive manner and then to extract the more discriminative items from it using mutual information between dictionary items and class labels [3, 4, 5, 8]. But the fundamental weakness of this approach is that the initial reconstructive dictionary places a ceiling on the discriminability of the extracted dictionary.

A better alternative is to incorporate discriminability into the reconstructive dictionary learning framework [6, 7, 9]. However, these approaches suffer from the instability of the discriminative term and require careful tuning of the reconstructive and discriminative parameters in order to avoid instability.

In this work we follow this second approach and introduce a discriminative version of the deviation function that yields a more stable learning formulation by allowing the trade-off between reconstruction and discrimination to be handled in a more principled manner via constraining the search-space for the tuning parameter.

2. Preliminaries

An input signal $\mathbf{x} \in \Re^n$ can be represented using a sparse code vector $\boldsymbol{\alpha}_j \in \Re^k$ under an overcomplete $(n < k)$ dictionary $\mathbf{D}_j \in \Re^{n\times k}$ obtained as the solution to the sparse coding problem

$$\boldsymbol{\alpha}_j = \arg\min_{\boldsymbol{\alpha}\in\Re^k} ||\mathbf{x} - \mathbf{D}_j\boldsymbol{\alpha}||_F^2 \; s.t \; ||\boldsymbol{\alpha}||_0 \leq L \qquad (1)$$

where L is the sparsity factor (maximum number of non-zero coefficients in $\boldsymbol{\alpha}$)[2]. This can be thought of as a generalization of standard vocabulary based bag-of-words approaches where an input signal is represented as an optimally sparse vector consisting of only one non-zero coefficient corresponding to the closest vocabulary word. The reconstruction error $\mathcal{R}_j$ for signal $\mathbf{x}$ under dictionary $\mathbf{D}_j$ can be computed as

$$\mathcal{R}_j = ||\mathbf{x} - \mathbf{D}_j\boldsymbol{\alpha}_j||_F^2 \qquad (2)$$

For a set of M signals $\mathbf{x}_1 \ldots \mathbf{x}_M$, the optimal reconstructive dictionary $\mathbf{D}$ and sparse codes $\boldsymbol{\alpha}$ can be computed via

$$\mathbf{D}, \boldsymbol{\alpha} = \arg\min_{\mathbf{D},\boldsymbol{\alpha}} \sum_{i=1}^{M} \mathcal{R}(\mathbf{x}_i) \qquad (3)$$

which can be solved via the KSVD [1] or MOD [2] algorithms.

For N class classification, per-class dictionaries $\mathbf{D}_1 \ldots \mathbf{D}_N$ can be learned and a test signal $\mathbf{x}$ can be classified via $\arg\min_{j=1\ldots N} \mathcal{R}_j$. In order to make the dictionaries more discriminative we incorporate a discriminative deviation function into the learning framework and this is explained next.

[1] Alternative terms in literature are codebooks, dictionaries.

[2] In the rest of the paper, sparsity factor L is implied on every sparse code $\boldsymbol{\alpha}$.

3. Discriminative Deviation Function

For a set of values $x_1, \ldots, x_N$ deviation is defined as the difference between an observed value x_i and the mean $\overline{x}$. For a signal belonging to class i we define reconstruction error based discriminative deviation as

$$\mathcal{D}_i = \mathcal{R}_i - \frac{\sum_{j=1}^{N} \mathcal{R}_j}{N} \tag{4}$$

which is positive if $\mathcal{R}_i$ is above the mean $\frac{\sum_{j=1}^{N} \mathcal{R}_j}{N}$ and negative if $\mathcal{R}_i$ is below the mean. Minimizing $\mathcal{D}_i$ for a signal from class i encourages the reconstruction error $\mathcal{R}_i$ to be lowest among $\mathcal{R}_1, \ldots, \mathcal{R}_N$. This leads to more discriminability and allows us to obtain the following discriminative dictionary learning formulation

$$C(\{\mathbf{D}\}_{j=1}^{N}) = \min_{\{\mathbf{D}\}_{j=1}^{N}} \sum_{i=1}^{N} \sum_{l \in S_i} (\mathcal{D}_{li} + \gamma \mathcal{R}_{li}) \tag{5}$$

where S_i is the set of input signals belonging to class i and $\mathcal{D}_{li}$ is the discriminative deviation $\mathcal{D}_i(\mathbf{x}_l)$ of signal $\mathbf{x}_l$ for class i and $\mathcal{R}_{li}$ is the reconstruction error $\mathcal{R}_i(\mathbf{x}_l)$. The reconstructive weight $\gamma > 0$ controls the trade-off between discrimination and reconstruction.

One can show via Jensen's inequality that $\mathcal{D}_i$ is a lower-bound on the discriminative softmax function ($\mathcal{C}_i = \log \sum_{j=1}^{N} e^{(-\lambda(\mathcal{R}_j - \mathcal{R}_i)}$) used by Mairal *et al.* [6]. Therefore, objective function (5) is also a lower-bound on the discriminative cost function found in [6] with very similar behavior as demonstrated in Figure 1. It is important to note that this behavior is achieved without the discriminative parameter λ from [6].

In [6], a continuation strategy is proposed for stable iterative minimization whereby parameter values are initially set to values corresponding to stable reconstructive optimization and gradually changed to move towards the more discriminative but less stable optimization. However, the search space for the parameters remains unclear. We show in the next section how cost function (5) can be made more stable by constraining the search space of the reconstructive parameter γ and using it as a true trade-off parameter.

4. Stable Discriminative Dictionary Learning (SDDL)

By constraining γ to lie between 0 and 1, the following more balanced objective function can be obtained

$$C(\{\mathbf{D}\}_{j=1}^{N}) = \min_{\{\mathbf{D}\}_{j=1}^{N}} \sum_{i=1}^{N} \sum_{l \in S_i} (1 - f(\gamma)) \mathcal{D}_{li} + \gamma \mathcal{R}_{li} \tag{6}$$

where γ is used as a true trade-off parameter. The function $f(\cdot)$ introduces a non-linearity that allows a larger range of values of γ to be considered before running into instability issues. We choose $f(\gamma) = \sqrt{\gamma}$. As a result, the weight $1 - \sqrt{\gamma}$ of the less stable discriminative term remains small for a larger range of γ values while allowing the weight γ of the more stable reconstructive term to drop more drastically.

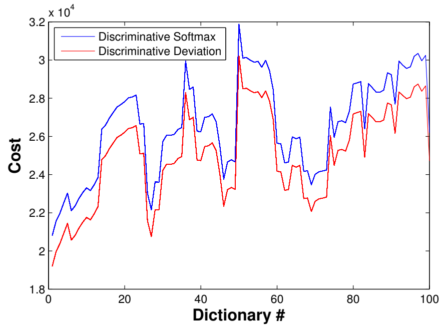

Figure 1. Comparison of the discriminative deviation based objective function (5) **with the discriminative softmax based objective function from [6] for 100 different dictionary configurations. Function** (5) **exhibits similar behavior without the need for a discriminative parameter as in [6].**

Cost function (6) can be optimized via Newton iterations, MOD [2], or KSVD [1]. We optimize by employing the MOD algorithm.

5. Experiments and Results

To validate our formulation, we perform pixel-wise classification on the Graz02 bikes dataset and on the UCF Sports action dataset.

Graz02 We select the first 300 images of the bike category from the Graz02 dataset and use odd numbered images for training and even numbered images for testing. For each training image, dense SIFT features are computed from overlapping patches of size 32×32 with a grid spacing of 12 pixels. For testing images the grid spacing is set to 4.

We run 30 iterations of KSVD[3] to train 2 separate reconstructive dictionaries $\mathbf{D}_f$ and $\mathbf{D}_b$ for foreground and background respectively using the training images and the provided ground-truth shape masks. Each dictionary has 256 items and the sparsity factor L is set to 8. To demonstrate the improvement of our discriminative approach over reconstructive approaches, these

[3] http://www.cs.technion.ac.il/∼ronrubin/software.html

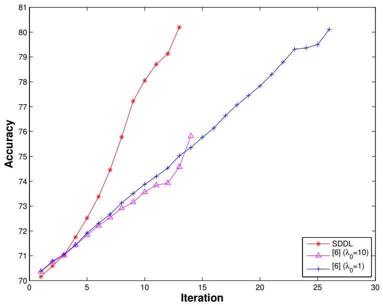

Figure 2. Stability comparison of SDDL with the formulation of [6] with high ($\lambda_0 = 10$) and low ($\lambda_0 = 1$) initializations of their discriminative parameter λ and reconstructive parameter γ initialized to 100. λ and γ were gradually updated as proposed in [6]. All three optimizations were continued until instability. For [6], learning with high discriminability leads to instability quickly while not achieving high accuracy while learning with low discriminability takes longer to achieve high accuracy. In contrast, SDDL achieves faster learning and only requires a single tuning parameter constrained between 0 and 1.

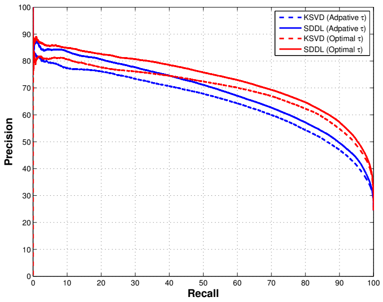

Figure 3. Comparison of precision-recall curves on the testing set of Graz02 bikes dataset using reconstructively learned dictionaries via KSVD (dashed curves) and discriminatively trained dictionaries via SDDL (solid curves). See text for details.

dictionaries are used as initial solution for the iterative optimization of (6). For each SIFT feature in a test image I, we compute the reconstruction errors $\mathcal{R}_f$ and $\mathcal{R}_b$ under both dictionaries and classify as foreground if $\mathcal{R}_f < \tau\mathcal{R}_b$ where the optimal value of $0 < \tau \leq 1$ is learned from the training data via cross-validation. Alternatively, τ can be set adaptively for each test image based on the first and second moments of the reconstruction errors. Interpolation is carried out for missing pixel values and the result is smoothed to obtain the final pixel-wise classification confidence that is used in all subsequent precision-recall curve calculations.

Figure 2 demonstrates that, compared to [6], our stable formulation (6) offers more control over the optimization due to one less parameter to search over and also due to constraining its only parameter to lie between 0 and 1. On the other hand, in [6], there is a lack of clarity as regards to what range of values to consider for the discriminative parameter λ as well as the reconstructive parameter γ.

Figure 3 compares precision-recall curves on the Graz02 bikes dataset using reconstructively learned dictionaries via KSVD (dashed curves) and discriminatively trained dictionaries via SDDL (solid curves). Blue curves represent adaptive setting of the classification parameter τ for each test image. Red curves represent τ optimally learned from the training set. It can be observed that discriminative dictionaries yield better classification performance. The benefit of learning an optimal τ from the training set can also be observed. The best achieved EER (Equal Error Rate where precision=recall) is 69.5% which is better than that achieved by [9].

UCF Sports Similar to our setup for the Graz02 dataset, we learn foreground and background dictionaries on dense STIP descriptors[4] [11] for the Diving and Gym (beam) categories from the UCF Sports actions dataset [10]. We replicate the evaluation setup of Yao *et al.* [12] who consider these two classes to be difficult. We compare against their action localization performance in Table 1. Considering that we neither do tracking nor ground-truth based initialization for test videos as in [12], our pixel classification based localization is comparable. Figure 5 demonstrates localization results on two selected frames.

6. Conclusion

We have introduced a new discriminative deviation based formulation for dictionary learning that is more stable than previous work while requiring only one tuning parameter and handling the discrimination/reconstruction trade-off in a more principled manner. Its applicability has been shown via state-of-the-art results on two real-world datasets. Ongoing efforts aim at incorporating context into the learning framework.

[4] http://www.irisa.fr/vista/Equipe/People/Laptev/download/stip-2.0-linux.zip

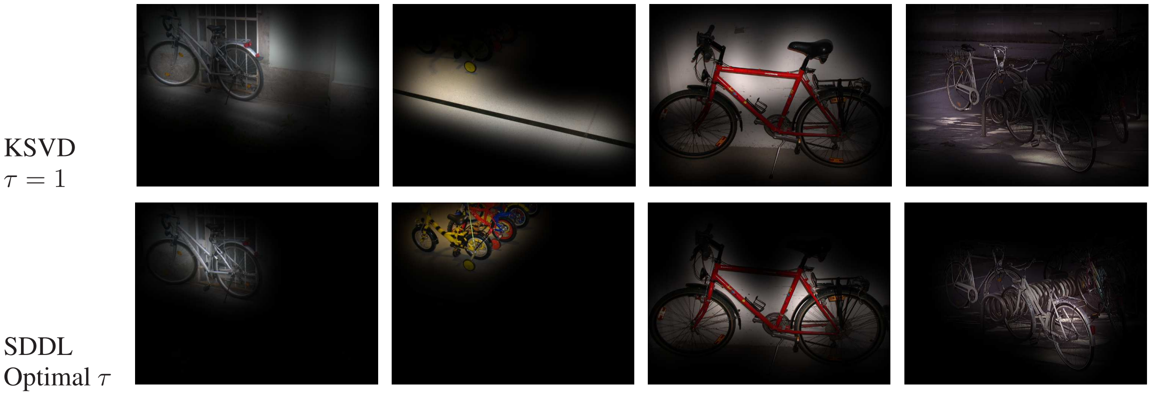

Figure 4. Row 1: Reconstructive dictionaries (KSVD) with $\mathcal{R}_f < \mathcal{R}_b$ based pixel-wise classification shows a greater tendency to classify background as foreground while Row 2: Our discriminatively learned dictionaries (SDDL) with $\mathcal{R}_f < \tau\mathcal{R}_b$ and optimal τ are able to achieve much better pixel-wise classification.

Table 1. Localization on UCF Sports. Percentage of frames with localized bounding boxes having intersection over union with ground-truth $> \frac{1}{2}$. Consider ing that we neither do tracking nor ground-truth based initialization for test videos as in [12], our pixel classification based localization is comparable.

	Gym (beam)	Diving
[12]	62%	68%
SDDL	52%	55%

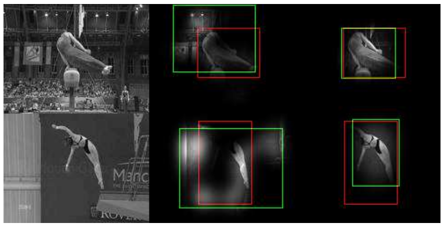

Figure 5. Dictionary based (green) and ground-truth (red) localization on UCF Sports. Left: Original frame. Middle: Untrained. Right: SDDL Trained.

7. Acknowledgements

M.F.T. was supported by NSF grants IIS-0905387 and IIS-0916868

References

[1] M. Aharon, M. Elad, and A. Bruckstein. K-svd: Design of dictionaries for sparse representation. In *SPARS*, 2005.

[2] K. Engan, S. O. Aase, and J. H. Husøy. Frame based signal compression using method of optimal directions (mod). In *ISCAS (4)*, 1999.

[3] B. Fulkerson, A. Vedaldi, and S. Soatto. Localizing objects with smart dictionaries. In *ECCV*, 2008.

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

[5] J. Liu, Y. Yang, and M. Shah. Learning semantic visual vocabularies using diffusion distance. In *CVPR*, 2009.

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

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

[8] Q. Qiu, J. Zhoulin, and R. Chellappa. Sparse dictionary-based representation and recognition of action attributes. In *ICCV*, 2011.

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

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

[11] H. Wang, M. M. Ullah, A. Kläser, I. Laptev, and C. Schmid. Evaluation of local spatio-temporal features for action recognition. In *BMVC*, 2009.

[12] A. Yao, D. Uebersax, J. Gall, and L. J. V. Gool. Tracking people in broadcast sports. In *DAGM*, 2010.