From Dictionary of Visual Words to Subspaces: Locality-constrained Affine Subspace Coding

Peihua Li, Xiaoxiao Lu, Qilong Wang
School of Information and Communication Engineering, Dalian University of Technology.

Motivation It is known that the feature data are often located on some low-dimensional manifold, leveraging the geometry of the manifold may bring benefits. The locality-constrained linear coding (LLC) [3] characterizes the geometry of feature space by a dictionary of visual words, which provides a crude, piecewise constant approximation of the manifold [1]. However, the geometric structure surrounding the words are not considered. We present a novel encoding method called Locality-constrained Affine Subspace Coding (LASC). We explicitly model the geometric structure of the immediate neighborhoods of the visual words by low-dimensional linear subspaces. The dictionary of affine subspaces thus obtained provides a piecewise linear approximation of the underlying manifold [1]. Figure 1 shows the flowchart of the LASC method and comparison with LLC.

We first define an ensemble of low-dimensional subspaces attached to some representative points:

$$\mathcal{S}_i = \left\{\mu_i + \mathbf{A}_i\mathbf{x}_i,\ \mathbf{x}_i \in \mathbb{R}^p\right\},\ i = 1,\ldots,M \tag{1}$$

where μ_i indicates a representative point and $\mathbf{A}_i$ is an $n \times p$ matrix whose columns form a basis of the linear subspace. Indeed, $\mathcal{S}_i$ defines a local coordinate system and all of these local coordinate systems put together characterize the holistic structure of the manifold.

Method Our idea is to represent a feature $\mathbf{y}$ by its top-k most neighboring affine subspaces, and meanwhile constraining the projection of $\mathbf{y}$ in each subspace by the proximity measure (PM) of the feature to this subspace. Specifically, the objective function of LASC is formulated as

$$\min_{\forall\, \mathbf{x}_i} \sum_{\mathcal{S}_i \in \mathcal{N}_k^S(\mathbf{y})} \left\{ \left\|(\mathbf{y}-\mu_i) - \mathbf{A}_i\mathbf{x}_i\right\|_2^2 + \lambda d(\mathbf{y},\mathcal{S}_i)\left\|\mathbf{x}_i\right\|_2^2 \right\}, \tag{2}$$

where $\lambda > 0$ is a regularization parameter, $\mathcal{N}_k^S(\mathbf{y})$ is the neighborhood of $\mathbf{y}$ defined by the k closest subspaces, and $d(\mathbf{y},\mathcal{S}_i)$ indicates the PM value of $\mathbf{y}$ to $\mathcal{S}_i$. Three PMs are considered based on the reconstruction error (d_r), the assumption of spherical Gaussian (d_s) and general Gaussian (d_p).

We segment the feature space by the simple k-means algorithm. Assume each cluster can be modeled by a low-dimensional linear subspace. For cluster i, we employ the PCA to preserve the p most signficant directions $\mathbf{u}_{i,j}$ with corresponding variances $\sigma_{i,j}^2$, j=1,…,p. Let $\mathbf{A}_i = \mathbf{U}_i = \left[\mathbf{u}_{i,1},\cdots,\mathbf{u}_{i,p}\right]$. Clearly (2) decouples into independent Ridge regression problems in $\mathbf{x}_i$, and the solution can be written as

$$\mathbf{x}_i = w_{\mathbf{y}}^i \mathbf{z}_i = w_{\mathbf{y}}^i \mathbf{U}_i^T(\mathbf{y}-\mu_i),\ \mathbf{z}_i \in \mathbb{R}^p \tag{3}$$

for $\mathcal{S}_i \in \mathcal{N}_k^S(\mathbf{y})$, and $w_i = (1+\lambda d(\mathbf{y},\mathcal{S}_i))^{-1}$. Thus far we can write out the *first-order* LASC vector for the feature $\mathbf{y}$ as $\mathbf{x} = \left[\mathbf{x}_1^T,\ldots,\mathbf{x}_i^T,\ldots,\mathbf{x}_M^T\right]^T$.

We propose to leverage the second-order information based on Fisher information metric (FIM) [2]. After some derivations, we obtain the second-order LASC vector

$$\mathbf{x}_i^{\cdot 2} = w_{\mathbf{y}}^i \mathbf{f}_{\lambda_i} = \frac{w_{\mathbf{y}}^i}{\sqrt{2}}\left[\frac{z_{i,1}^2}{\sigma_i^2} - 1,\ldots,\frac{z_{i,p}^2}{\sigma_p^2} - 1\right]^T. \tag{4}$$

The final LASC vector, containing both the first- and second-order information, has the following form:

$$\mathbf{x} = \begin{bmatrix}\vdots\\ \mathbf{x}_i\\ \mathbf{x}_i^{\cdot 2}\\ \vdots\end{bmatrix}, \overbrace{\underbrace{\mathbf{x}_i = w_{\mathbf{y}}^i\begin{bmatrix}z_{i,1}\\ \vdots\\ z_{i,p}\end{bmatrix}}_{1^{\mathbf{st}}\text{-order vector}}, \underbrace{\mathbf{x}_i^{\cdot 2} = \frac{w_{\mathbf{y}}^i}{\sqrt{2}}\begin{bmatrix}\left(\frac{z_{i,1}}{\sigma_{i,1}}\right)^2 - 1\\ \vdots\\ \left(\frac{z_{i,p}}{\sigma_{i,p}}\right)^2 - 1\end{bmatrix}}_{2^{\mathbf{nd}}\text{-order vector}}}^{\text{LASC vector if } \mathcal{S}_i\in\mathcal{N}_k^S(\mathbf{y})\text{; otherwise } \mathbf{x}_i=\mathbf{x}_i^{\cdot 2}=\mathbf{0}} \tag{5}$$

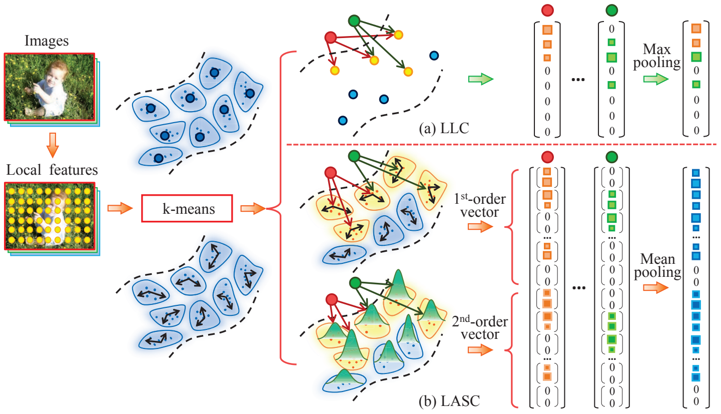

Figure 1: The dictionary of LLC (a) is a set of visual words while that of LASC (b) is an ensemble of low-dimensional linear subspaces attached to some representative points (i.e. affine subspaces). For an input feature, we find its top-k nearest subspaces and perform linear decomposition of the feature in these subspaces weighted by the proximity measures. Beyond the linear coding, we propose to leverage the second-order information.

Discussion Our method is similar to LTC [4] which aims at learning a nonlinear function by introducing local tangent directions computed by PCA. However, we intend to obtain highly distinct representation by encoding on an ensemble of affine subspaces, as opposed to the encoding of LTC on individual visual words. Moreover, we use the PRs to assign features to their k most neighboring affine subspaces and weight the coding vector, while LTC computes the weight computed using the LCC [5] coefficients by solving the LASSO problem. Last, we present the second-order encoding in each subspace based on FIM [2], which amounts to explore the geometry of the Riemannian manifold from the statistic perspective.

The FV also exploits FIM and performs local coding with respect to 5∼10 Gaussians with significant posterior probabilities [2, Appendix 2]. The FV uses a global PCA basis for dimensionality reduction and models the universal GMM also in that global system. In contrast, the LASC leverages an ensemble of local coordinate systems of varying origins and the corresponding local bases. The dimensionality reduction and coding are both relative to the local bases, which distinguishes the LASC from most of the existing coding methods.

Results Comparisons with state-of-the-arts are shown in Table 1. Note that Super vector (SV) coding [6] is a special case of LTC.

Method	VOC2007	Caltech256 (30 train)	MIT Indoor	SUN397 (50 train)
LLC [3]	57.6	41.2(-)	-	32.4(-)
SV [6]	58.2	42.4(-)	56.2	36.6(-)
FV [2]	61.8	47.4(0.1)	61.3	43.3(0.2)
LASC	**63.6**	**52.1(0.1)**	**63.4**	**45.3(0.4)**

Table 1: Comparisons on image classification benchmarks

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage. The work was supported by the NSFC (61471082).

[1] Guillermo D. Canas, Tomaso Poggio, and Lorenzo Rosasco. Learning manifolds with k-means and k-flats. *NIPS*, 2012.

[2] J. Sánchez, F. Perronnin, T. Mensink, and J. Verbeek. Image Classification with the Fisher Vector: Theory and Practice. *IJCV*, 2013.

[3] J. Wang, J. Yang, K. Yu, F. Lv, T. S. Huang, and Y. Gong. Locality-constrained linear coding for image classification. In *CVPR*, 2010.

[4] Kai Yu and Tong Zhang. Improved local coordinate coding using local tangents. In *ICML*, 2010.

[5] Kai Yu, Tong Zhang, and Yihong Gong. Nonlinear learning using local coordinate coding. In *NIPS*, 2009.

[6] X. Zhou, K. Yu, T. Zhang, and T. S. Huang. Image classification using super-vector coding of local image descriptors. In *ECCV*, 2010.