Hessian Regularized Sparse Coding for Human Action Recognition

Weifeng Liu[1], Zhen Wang[1], Dapeng Tao[2,3], and Jun Yu[4]

[1] China University of Petroleum (East China), Qingdao, 266580, China
[2] Shenzhen Institutes of Advanced Technology, Chinese Academy of Science, Shenzhen, China
[3] The Chinese University of Hong Kong, Hong Kong, China
[4] Hangzhou Dianzi University, Hangzhou, 310018, China
1liuwf@upc.edu.cn

Abstract. With the rapid increase of online videos, recognition and search in videos becomes a new trend in multimedia computing. Action recognition in videos thus draws intensive research concerns recently. Second, sparse representation has become state-of-the-art solution in computer vision because it has several advantages for data representation including easy interpretation, quick indexing and considerable connection with biological vision. One prominent sparse representation algorithm is Laplacian regularized sparse coding (LaplacianSC). However, LaplacianSC biases the results toward a constant and thus results in poor generalization. In this paper, we propose Hessian regularized sparse coding (HessianSC) for action recognition. In contrast to LaplacianSC, HessianSC can well preserve the local geometry and steer the sparse coding varying linearly along the manifold of data distribution. We also present a fast iterative shrinkage-thresholding algorithm (FISTA) for HessianSC. Extensive experiments on human motion database (HMDB51) demonstrate that HessianSC significantly outperforms LaplacianSC and the traditional sparse coding algorithm for action recognition.

Keywords: Action recognition, sparse coding, Hessian regularization, manifold learning.

1 Introduction

Due to the development of Internet technology and smart devices, explosive growth of social videos are produced and spread on the Internet frequently. For example, every day YouTube streams more than 1 billion videos most of which are unlabeled. It is impractically expensive to manually annotate this huge volume of videos. Thus there is an emergent demand for effective methods which can help to organize this increasing visual media data including video summarizing, indexing, retrieval, classification and annotation [1]. And human action recognition is one of the most attractive research topic very recently.

Given an unknown video sequence, human action recognition aims to automatically classify ongoing actions including gestures, movement, interactions, and group

X. He et al. (Eds.): MMM 2015, Part II, LNCS 8936, pp. 502–511, 2015.

activities. Most action recognition methodologies employ spatio-temporal features to describe action in videos by concatenating video frame along time to form a 3-D space-time representation [2]. Briefly speaking, it can be divided into three categories: (1) action recognition with space-time volumes [3][4]; (2) action recognition with space-time trajectories [5][6][7] and (3) action recognition with space-time local features [8][9][10][11][12][14][15][16]. The methods with space-time volumes [3][4] recognize human actions by measure the similarity between the test video volume and template video volume. The methods with space-time trajectories [5][6][7] interpret an human action as a set of space-time trajectories which consist of a set of 2-dimensional or 3-dimensional points corresponding to human joint positions. The methods with space-time local features represent and recognize human actions using local features extracted from 3-D space-time volumes including concatenation local features at every frame [8][9][12][13] or on interest points [10][11][14][15].

Considering the redundancy of the space-time features for action representation, it is essential to employ a proper representation to reveal the underlying process of these observations. Sparse coding has received growing attentions because of its promising performance in machine learning, signal processing, neuroscience and statistics. Sparse coding aims to learn a dictionary and simultaneously the sparse coordinates w.r.t. the dictionary to represent the observations. It yields an easier interpretation because each data point is represented as a linear combination of a small set of dictionary atoms. And also sparse coding has some considerable connection with biological vision mechanism [17]. Hence a lot of variant algorithms and applications of sparse coding have been developed in recent years. In brief, sparse coding algorithms can be categorized into the following groups: (1) reconstructive sparse coding [18][19][20][21][22], (2) structured sparse coding [23][24][25] and (3) manifold regularized sparse coding [26][27][28][29][30]. Reconstructive sparse coding minimizes the data reconstruction error by different optimization algorithms including matching pursuit [19] and basis pursuit [18]. Structured sparse coding exploits the structure sparsity for a certain purpose such as group sparsity [23], hierarchical sparsity [24] and latent space [25]. Manifold regularized sparse coding exploits the local geometry of the data distribution by graph regularization including graph Laplacian [26][27] and Hessian [28][29].

In this paper, we propose Hessian regularized sparse coding (HessianSC) for action recognition. In contrast to Laplacian, Hessian has richer null space and favors the solution varying linearly w.r.t. geodesic distance [31]. Then Hessian regularization can better preserve the local geometry and lead to better extrapolation capability. Hence, HessianSC can achieve smoother sparse coding that preserves local similarity and result more excellent performance than traditional sparse coding algorithms. We also present the fast iterative shrinkage-thresholding algorithm (FISTA) [32] for the optimization of HessianSC. In the sense of Nemirovsky and Yudin [33], FISTA is one "optimal" first order method for sparse coding [32] with an $O(1/k^2)$ complexity. Finally, we carefully implement HessianSC for action recognition and conduct experiments on the HMDB51 database [34]. To evaluate the performance of HessianSC, we also compare HessianSC with some baseline algorithms including traditional sparse coding and Laplacian regularized sparse

coding (LaplacianSC). The experimental results verify the effectiveness of HessianSC by comparison with the baseline algorithms.

The rest of this paper is assigned as follows. Section 2 provide a brief description of the proposed Hessian regularized sparse coding (HessianSC) algorithm. Section 3 introduces the optimization scheme of HessianSC using FISTA. Section 4 reports some experimental results followed with conclusions in section 5.

2 Hessian Regularized Sparse Coding

Suppose we are given N examples $S = \{x_i\}_{i=1}^N$, sparse coding aims to learn the sparse representation w_i of each example x_i simultaneously with a dictionary D. In the following section of this paper, we use $X = [x_1, \cdots, x_N] \in R^{m \times N}$ to denote the data matrix of examples and $W = [w_1, \cdots, w_N] \in R^{d \times N}$ to denote the sparse codes matrix w.r.t. to the dictionary $D = [D_1, \cdots, D_d] \in R^{m \times d}$. Then sparse coding can be formulated as follows:

$$\min_{D,W} \frac{1}{2N} \|X - DW\|_F^2 + \lambda_1 \sum_{i=1}^N \|w_i\|_1 , \text{s.t.} \left\|D_j\right\|_2 \leq 1, 1 \leq j \leq d. \tag{1}$$

Under manifold assumption [35], it is crucial to explore the local geometry because the sparse codes w_i, w_j of two examples x_i and x_j respectively are close to each other if the two examples are close in the intrinsic geometry of the data distribution. Hence in this paper, we integrate Hessian regularization into the objective function of sparse coding and reformulate HessianSC as below:

$$\min_{D,W} \frac{1}{2N} \|X - DW\|_F^2 + \lambda_1 \sum_{i=1}^N \|w_i\|_1 + \lambda_2 Tr(WHW^T) , \text{s.t.} \left\|D_j\right\|_2 \leq 1, 1 \leq j \leq d. \tag{2}$$

Here H is the Hessian computed from the data matrix.

The objective function in (2) is convex w.r.t. D or W separately, but it is not convex w.r.t. both variables together. In this paper, we employ alternating optimization to solve the problem by optimizing one variable while keeping the other one fixed. Thus the solution of (2) can be generally divided into two parts: sparse coding and dictionary updating. In the following section, we detail the optimization algorithm of (2).

3 Optimization of HessianSC

The optimization of HessianSC contains two steps: (1) learning sparse codes W given fixed dictionary D and (2) updating dictionary D given fixed sparse codes W. In particular, given fixed dictionary D, the problem (2) can be written as the follow subproblem:

$$\min_W \frac{1}{2N}\|X - DW\|_F^2 + \lambda_1 \sum_{i=1}^{N}\|w_i\|_1 + \lambda_2 Tr(WHW^T). \quad (3)$$

Given fixed sparse codes W, the problem (2) can be written as the follow sub-problem:

$$\min_D \frac{1}{2N}\|X - DW\|_F^2, \text{s.t.} \left\|D_j\right\|_2 \le 1, 1 \le j \le d. \quad (4)$$

In the following, we present the optimization of subproblem (3) and (4) in detail.

3.1 Learning Sparse Codes W with Fixed D

In this section, we describe the optimization of subproblem (3) using FISTA [32]. Subproblem (3) can be expressed as the general form:

$$\min_W \{F(W) \equiv f(W) + g(W)\}, \quad (5)$$

where $f(W) = \frac{1}{2N}\|X - DW\|_F^2 + \lambda_2 Tr(WHW^T)$, $g(W) = \lambda_1 \sum_{i=1}^{N}\|w_i\|_1$. $f(W)$ and $g(W)$ are both convex functions.

Adopting gradient algorithm, subproblem (5) leads to the iterative scheme:

$$W_k = \text{argmin}_W \left\{Q_L(W, W_{k-1}) \equiv f(W_{k-1}) + \langle W - W_{k-1}, \nabla f(W_{k-1})\rangle + \frac{L}{2}\|W - W_{k-1}\|^2 + \lambda_1 \sum_{i=1}^{N}\|w_i\|_1\right\} \quad (6)$$

where $\langle A, B\rangle = Tr(A^T B)$, and L is the Lipschitz constant of ∇f. Ignoring constant terms, (6) can be rewritten as:

$$W_k = p_L(W_{k-1}) \equiv \text{argmin}_W \left\{\frac{L}{2}\left\|W - \left(W_{k-1} - \frac{1}{L}\nabla f(W_{k-1})\right)\right\|^2 + \lambda_1 \sum_{i=1}^{N}\|w_i\|_1\right\}. \quad (7)$$

Since l_1 norm is separable, subproblem (7) can then be solved using the shrinkage operator as follows:

$$w_i = \mathcal{T}_{\frac{\lambda_1}{L}}\left(w_{k-1} - \frac{1}{L}\nabla f(w_{k-1})\right),$$

where $\mathcal{T}_\alpha(x)_j = \left(|x_j| - \alpha\right)_+ sgn(x_j)$.

Then we can state the optimization of subproblem (3) using FISTA with back-tracking stepsize in Table 1.

3.2 Update Dictionary D with Fixed W

Subproblem (4) is a l_2-constrained least squares problem and can be equally rewritten as:

$$\min_D \|X - DW\|_F^2, \text{s.t.} \left\|D_j\right\|_2 \le 1, 1 \le j \le d \quad (8)$$

In this section, we describe the optimization of (8) using Largrange dual [36].

Table 1. FISTA optimization for subproblem (3)

— **Input**: $X, D, H, \lambda_1, \lambda_2$
— **Output**: W
— **Step 0**: chose $W_0, Z_1 = W_0, L_0, \eta > 1, t_1 = 1$
— **Step *k***:
— 1. set $\bar{L} = L_{k-1}$
— 2. repeat $\bar{L} = \eta\bar{L}$, until $F\left(p_{\bar{L}}(Z_k)\right) \le Q_{\bar{L}}(p_{\bar{L}}(Z_k), Z_k)$
— 3. set $L_k = \bar{L}$
— 4. update
— $W_k = p_{\frac{1}{L_k}}(Z_k)$
— $t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}$
— $Z_{k+1} = W_k + \left(\frac{t_k - 1}{t_{k+1}}\right)(W_k - W_{k-1})$

Consider $\beta = [\beta_1, \cdots, \beta_d]$ as the Lagrange multiplier, the Lagrange dual function of subproblem (8) can be written as follows:

$$g(\beta) = \min_D \mathcal{L}(D, \beta) = \min_D \|X - DW\|_F^2 + \sum_{j=1}^{d} \beta_j \left(D_j^T D_j - 1\right)$$

$$= \min_D Tr\left((X - DW)^T (X - DW)\right) + Tr(D^T DB) - Tr(B)$$

$$= \min_D Tr(X^T X - 2D^T XW^T + W^T D^T DW + D^T DB - B) \tag{9}$$

where $B = diag(\beta)$ is $d \times d$ diagonal matrix with diagonal entry $B_{jj} = \beta_j$ for all j.

Set the first order derivative of $\mathcal{L}(D, \beta)$ w.r.t. D to zero, we have

$$D^* W W^T - X W^T + D^* B = 0.$$

Then, we have

$$D^* = XW^T (WW^T + B)^{-1}. \tag{10}$$

Substituting (10) into (9), the Lagrange dual of subproblem (4) can be written as:

$$g(\beta) = Tr(X^T X - XW^T (WW^T + B)^{-1} WX^T - B). \tag{11}$$

After solving the maximization of (11) w.r.t β by using Newton's method, we obtain the optimal dictionary D^* as $D^* = XW^T (WW^T + B^*)^{-1}$.

4 Experiments

To evaluate the effectiveness of the proposed HessianSC, we apply support vector machines as classifier to the sparse codes obtained by HessianSC for action recognition. We conduct the experiments on the HMBD51 database [1]. HMDB51 contains 6849

video clips of 51 distinct action categories, each containing at least 101 clips. Each clip was validated by at least two human observes to ensure consistency. The 51 actions categories can be grouped in five types: (1) general facial actions, (2) facial actions with object manipulation, (3) general body movements, (4) body movements with object interaction and (5) body movements for human interaction (see Table 2). Figure 1 shows some sample frames of different categories from the HDMB51 database.

Fig. 1. Some sample frames from HDMB51 database

We implement bag-of-words on a concatenation of the HOG and HOF features to obtain action descriptors. The HOG and HOF features have been shown to be state-of-the-art descriptors for action representation. In this paper, we use the HOG and HOF features around 3D Harris corners which are provided by Kuehne et al. [1]. In particular, we cluster a subset of 100000 features sampled from the training videos with the *k*-means algorithm and form a set of 2000 visual words. By matching every local point descriptors to the nearest visual words, each action clip can be represented with a 2000-dimensional feature vector which is a histogram over the index of the matched code book entries. Sequently, these action descriptors are used to obtain sparse codes by HessianSC.

Table 2. Actions categories in five types

Action groups	Action labels
General facial actions	smile, laugh, chew, talk
Facial actions with object manipulation	smoke, eat, drink
General body movements	cartwheel, clap hands, climb, climb stairs, dive, fall on the floor, backhand flip, hand-stand, jump, pull up, push up, run, sit down, sit up, somersault, stand up, turn, walk, wave
Body movements with object interaction	brush hair, catch, draw sword, dribble, golf, hit something, kick ball, pick, pour, push something, ride bike, ride horse, shoot ball, shoot bow, shoot gun, swing baseball bat, sword exercise, throw
Body movements for human interaction	fencing, hug, kick someone, kiss, punch, shake hands, swordfight

According to [1], we select 70 training and 30 testing clips from each action class to form our experiment dataset including a training set that contains 3570 clips and a test set that contains 1530 clips. We compare the proposed HessianSC with Laplacian regularized sparse coding (LaplacianSC) and the traditional sparse coding algorithms. In our experiments, The number of dictionary atoms is set to 200, the parameters λ_1 and λ_2 are tuned from the candidate set $\{1 \times 10^e | -30, \cdots, 10\}$, and the number of the neighbors for computing Hessian and Laplacian is set to 100 empirically. Considering multiple action category recognition, we adopt one vs. one method that selects two action categories each time for classification.

Figure 2 shows the confusion matrix on selected 20 action categories. Although errors look like to be spread across category labels randomly, HessianSC performs significantly better than LaplacianSC and the traditional sparse coding methods.

Figure 3 illustrates the accuracy performance on the selected 20 single action categories. From Figure 3, we can see that HessianSC outperforms than the baseline algorithms including LaplacianSC and traditional sparse coding in most cases.

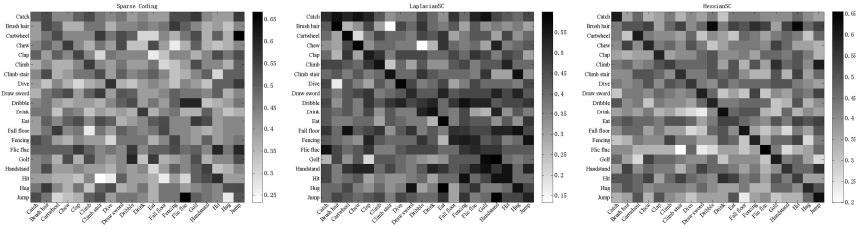

Fig. 2. Confusion matrix on selected 20 action categories

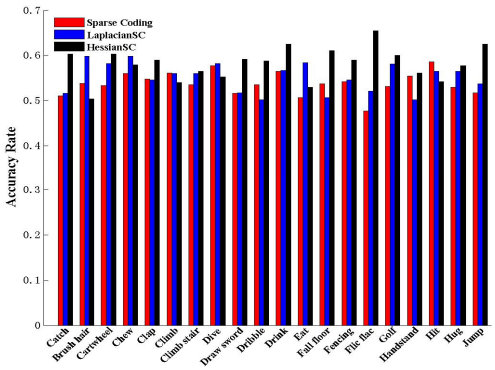

Fig. 3. Accuracy on single action category

5 Conclusion

Human action recognition have received intensive research attentions with the explosively growing of online videos. Although there are a lot of action representation methods, sparse coding has achieved stat-of-the-art performance in many computer vision applications. In this paper, we employed Hessian regularized sparse coding for human action recognition. The proposed HessianSC can well preserve local similarity benefitting from Hessian regularization. We also present a fast iterative shrinkage thresholding algorithm (FISTA) for efficient solving HessianSC. We apply HessianSC to support vector machines for action recognition. Extensive experiments on the HMBD51 database demonstrate that the proposed HessianSC significantly outperforms LaplacianSC and the traditional sparse coding algorithm for action recognition.

Acknowledgement. This paper is supported partly by National Natural Science Foundation of China (61301242, 61271407), Natural Science Foundation of Shandong Province (ZR2011FQ016), Fundamental Research for the Central Universities (13CX2096A).

References

1. Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., Serre, T.: HMDB: A large video database for human motion recognition. In: IEEE International Conference on Computer Vision (ICCV), pp. 2556–2563 (2011)
2. Aggarwal, J.K., Ryoo, M.S.: Human activity analysis: A review. ACM Computing Surveys (CSUR) 43(3), 16 (2011)
3. Ke, Y., Sukthankar, R., Hebert, M.: Spatio-temporal shape and flow correlation for action recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1–8 (2007)
4. Rodriguez, M., Ahmed, J., Shah, M.: Action MACH: A spatio-temporal maximum average correlation height filter for action recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2008)
5. Campbell, L.W., Bobick, A.F.: Recognition of human body motion using phase space constraints. In: IEEE International Conference Computer Vision, pp. 624–630 (1995)
6. Rao, C., Shah, M.: View-invariance in action recognition. In: IEEE Conferences on Computer Vision and Pattern Recognition (CVPR), vol. 2, p. II-316 (2001)
7. Sheikh, Y., Sheikh, M., Shah, M.: Exploring the space of a human action. In: IEEE International Conference on Computer Vision, vol. 1, pp. 144–149 (2005)
8. Chomat, O., Crowley, J.L.: Probabilistic recognition of activity using local appearance. In: IEEE Conference on Computer Vision and Pattern Recognition, vol. 2 (1999)
9. Zelnik-Manor, L., Irani, M.: Event-based analysis of video. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 2, p. II-123 (2001)
10. Laptev, I.: On space-time interest points. International Journal of Computer Vision 64(2-3), 107–123 (2005)
11. Yilmaz, A., Shah, M.: Actions sketch: A novel action representation. In: IEEE Conference on Computer Vision and Pattern Recognition, vol. 1, pp. 984–989 (2005)
12. Blank, M., Gorelick, L., Shechtman, E., Irani, M., Basri, R.: Actions as space-time shapes. In: IEEE International Conference on Computer Vision (ICCV), vol. 2, pp. 1395–1402 (2005)

13. Yu, J., Tao, D., Wang, M., Rui, Y.: Learning to Rank Using User Clicks and Visual Features for Image Retrieval. IEEE Transactions on Cybernetics (2014), 10.1109/TCYB.2014.2336697
14. Niebles, J.C., Wang, H., Fei-Fei, L.: Unsupervised learning of human action categories using spatial-temporal words. International Journal of Computer Vision 79(3), 299–318 (2008)
15. Ryoo, M.S., Aggarwal, J.K.: Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities. In: IEEE International Conference on Computer Vision (ICCV), pp. 1593–1600 (2009)
16. Hong, C., Yu, J., Chen, X.: Image-Based 3D Human Pose Recovery with Locality Sensitive Sparse Retrieval. In: 2013 IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp. 2103–2108 (2013)
17. Olshausen, B.A.: Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature 381(6583), 607–609 (1996)
18. Chen, S.S., Donoho, D.L., Saunders, M.A.: Atomic decomposition by basis pursuit. SIAM Journal on Scientific Computing 20(1), 33–61 (1998)
19. Mallat, S.G., Zhang, Z.: Matching pursuits with time-frequency dictionaries. IEEE Transactions on Signal Processing 41(12), 3397–3415 (1993)
20. Yu, J., Rui, Y., Tao, D.: Click Prediction for Web Image Reranking using Multimodal Sparse Coding. IEEE Transactions on Image Processing 23(5), 2019–2032 (2014)
21. Liu, B.-D., Wang, Y.-X., Zhang, Y.-J., Shen, B.: Learning dictionary on manifolds for image classification. Pattern Recognition 46(7), 1879–1890 (2013)
22. Liu, B.-D., Wang, Y.-X., Shen, B., Zhang, Y.-J., Hebert, M.: Self-explanatory sparse representation for image classification. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014, Part II. LNCS, vol. 8690, pp. 600–616. Springer, Heidelberg (2014)
23. Yuan, M., Lin, Y.: Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 68(1), 49–67 (2006)
24. Jenatton, R., Mairal, J., Bach, F.R., Obozinski, G.R.: Proximal methods for sparse hierarchical dictionary learning. In: The 27th International Conference on Machine Learning (ICML), pp. 487–494 (2010)
25. Jia, Y., Salzmann, M., Darrell, T.: Factorized latent spaces with structured sparsity. In: Advances in Neural Information Processing Systems, pp. 982–990 (2010)
26. Zheng, M., Bu, J., Chen, C., Wang, C., Zhang, L., Qiu, G., Cai, D.: Graph regularized sparse coding for image representation. IEEE Transactions on Image Processing 20(5), 1327–1336 (2011)
27. Gao, S., Tsang, I.W.-H., Chia, L.-T.: Laplacian sparse coding, hypergraph laplacian sparse coding, and applications. IEEE Transactions on Pattern Analysis and Machine Intelligence 35(1), 92–104 (2013)
28. Zheng, M., Bu, J., Chen, C.: Hessian sparse coding. Neurocomputing 123, 247–254 (2014)
29. Liu, W., Tao, D., Cheng, J., Tang, Y.: Multiview hessian discriminative sparse coding for image annotation. Computer Vision and Image Understanding 118, 50–60 (2014)
30. Yu, J., Wang, M., Tao, D.: Semisupervised multiview distance metric learning for cartoon synthesis. IEEE Transactions on Image Processing 21(11), 4636–4648 (2012)
31. Kim, K.I., Steinke, F., Hein, M.: Semi-supervised regression using hessian energy with an application to semi-supervised dimensionality reduction. In: Advances in Neural Information Processing Systems, pp. 979–987 (2009)
32. Beck, A., Teboulle, M.: A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM Journal on Imaging Sciences 2(1), 183–202 (2009)
33. Nemirovsky, A.S., Yudin, D.B.: Problem complexity and method efficiency in optimization (1983)

34. Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., Serre, T.: HMDB: A large video database for human motion recognition. In: IEEE International Conference on Computer Vision (ICCV), pp. 2556–2563 (2011)
35. Belkin, M., Niyogi, P., Sindhwani, V.: Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. The Journal of Machine Learning Research 7, 2399–2434 (2006)
36. Lee, H., Battle, A., Raina, R., Ng, A.Y.: Efficient sparse coding algorithms. In: Advances in Neural Information Processing Systems, pp. 801–808 (2006)