Large Margin Dimensionality Reduction for Action Similarity Labeling

Xiaojiang Peng, Yu Qiao, Qiang Peng and Qionghua Wang

***Abstract*—Action recognition in videos is receiving extensive research interest due to its wide applications. This task needs to assign a specific action class for each video. In this paper, we study the problem of action similarity labeling (ASLAN) that is to verify whether two action videos present the same type of action or not. We show that both Fisher vector (FV) and vector of locally aggregated descriptors (VLAD) with dense trajectory features can achieve state-of-the-art performance on the ASLAN benchmark. Our main contribution is to develop a large margin dimensionality reduction (LMDR) method to compress high-dimensional FV and VLAD. Specially, we leverage the hinge loss objective function and stochastic gradient descent to optimize the discriminative projection matrix of these vectors. Extensive experiments on the ASLAN dataset indicate that our LMDR method not only reduces the dimension significantly but also improves the verification performance.**

***Index Terms*—Action similarity labeling, large margin dimensionality reduction, VLAD, Fisher vector, similarity learning.**

I. Introduction

HUMAN action analysis in videos has become a highly active research area due to its wide applications in video surveillance, human-computer interface, and content based video retrieval [1], [2], [3], [4], [5], [6]. There exist several tasks in this area such as action recognition [2], [4], action detection [7] and Action Similarity LAbeliNg (ASLAN) [8]. This article addresses the ASLAN task: given a pair of videos, we wish to decide whether the videos present the same type of action or not.

Performance in the ASLAN task mainly depends on video representations and the similarity measure used to compare video pairs. For video representation, most of the methods in the ASLAN task followed those in the action recognition [8], [9], [10], [11], such as the popular bag-of-words (BoW) [12] representation with Space–Time Interest Points (STIPs) [1] and dense trajectories [4]. Orit *et al.* [8] provided the baseline performance using the STIP features with the BoW model, and improved it by introducing metric learning [9]. Orit *et al.* [10] proposed motion interchange patterns (MIP) with the BoW model for action recognition and action similarity labeling. Yair *et al.* [11] presented several variants of MIP, namely histMIP and DoGMIP, which replace the original patches employed in MIP by the Histogram of Gradient Orientations (HOG) and Difference of Gaussian (DoG), respectively. Yair *et al.* [11] also combined these variants of MIP with dense trajectory and the Motion Boundary Histograms (MBH) features, and achieved good results in both action recognition and ASLAN tasks. For similarity measure of video pairs, Orit *et al.* [8] applied 12 basic (dis-)similarities in the baseline experiments, and introduced the One-Shot-Similarity Metric Learning (OSSML) method [13] for the ASLAN task [9]. Both Orit *et al.* [9], [10] and Yair *et al.* [11] employed the Cosine Similarity Metric Learning (CSML) [14] after performing a PCA dimension reduction on the BoW representation.

Recent studies indicate that advanced feature encoding methods other than vector quantization (i.e., hard assignment [12]) can steadily improve action recognition performance [15]. Among these methods, Fisher vector (FV) [16] and vector of locally aggregated descriptors (VLAD) [17] are probably the two most effective approaches. One limitation of using FV and VLAD for ASLAN is their high dimension, which not only implies in a large computational cost but also may harm verification performance.

In this paper, we develop a large margin dimensionality reduction (LMDR) method to deal with this problem, which jointly performs dimensionality reduction and similarity learning. Our LMDR leverages the hinge loss function and stochastic gradient descent to optimize the projection matrix. Experiments show that our LMDR method not only can reduce the dimension significantly but also can improve the performance of both FV and VLAD on the ASLAN benchmark.

II. Methodology

The framework of our action similarity labeling approach is shown in Fig. 1. For a pair of videos, i) we extract the improved dense trajectories which are described by concatenating the HOG, Histogram of Optical Flow (HOF), and MBH descriptors; ii) we encode the descriptors in each video by VLAD (or FV) using a codebook pre-trained by K-means or Gaussian Mixture Models (GMM) and pool them to yield video-level representations; iii) we apply the proposed LMDR method to project these representations to a discriminative subspace; iv) we transform the two compact video representations to a single video-pair representation by performing

This work is partly supported by Natural Science Foundation of China (91320101, 61036008, 60972111), Shenzhen Basic Research Program (JC201005270350A, JCYJ20120903092050890, JCYJ20120617114614438), 100 Talents Program of CAS, Guangdong Innovative Research Team Program (201001D0104648280), and the 2013 Doctoral Innovation Funds of Southwest Jiaotong University.

X. Peng is with the School of Information Science and Technology, Southwest Jiaotong University, Chengdu, China (e-mail: xiaojiangp@gmail.com).

Y. Qiao is with the Shenzhen Key Lab of Comp. Vis. & Pat. Rec., Shenzhen Institutes of Advanced Technology, CAS, China (e-mail:yu.qiao@siat.ac.cn).

Q. Peng is with the School of Information Science and Technology, Southwest Jiaotong University, Chengdu, China (e-mail: qpeng@swjtu.edu.cn).

Q. Wang is with the School of Electronics and Information Engineering, Sichuan University, Chengdu, China.

(*Corresponding author: Y. Qiao*).

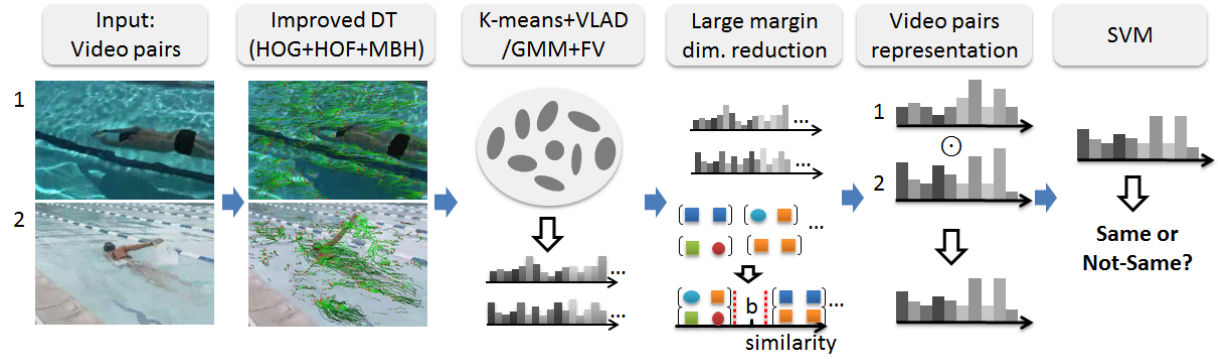

Fig. 1. The flow chart of our action similarity labeling approach.

point-wise multiplications, which is better than directly using their (dis)similarities [8] in our test. Finally, a binary SVM is learned to map the video-pair representation to a binary decision of same/not-same as [10].

A. Encoding Methods Revisited

VLAD. Jégou *et al.* proposed VLAD in [17]. Similar to standard BoW, a dictionary $\mathbf{D} = [\mathbf{d}_1, \mathbf{d}_2, ..., \mathbf{d}_K] \in \mathcal{R}^{d\times K}$ is first learned by K-means from training data. Let $\mathbf{X}_n = [\mathbf{x}_1, ...\mathbf{x}_N] \in \mathcal{R}^{d\times N}$ denote a set of local descriptors from a video. For each word $\mathbf{d}_i$, a vector $\mathbf{v}_i$ is yielded by aggregating the differences between the assigned features and $\mathbf{d}_i$:

$$\mathbf{v}_i = \sum_{\mathbf{x}_j : NN(\mathbf{x}_j)=i} (\mathbf{x}_j - \mathbf{d}_i), \tag{1}$$

where $NN(\mathbf{x}_j) = i$ denotes that the nearest neighborhood of $\mathbf{x}_j$ within $\mathbf{D}$ is $\mathbf{d}_i$. The VLAD representation is the concatenation of all the d-dimensional vectors $\mathbf{v}_i$ and is therefore a Kd dimensional vector.

Fisher vector. For Fisher vectors, we assume the generation process of local descriptors $\mathbf{X}$ can be modeled by a probability density function $p(\cdot;\theta)$ with parameter θ. The gradient of the log-likelihood w.r.t a parameter can describe how that parameter contributes to the generation process of $\mathbf{X}$ [18]. The probability density function is usually modeled by Gaussian Mixture Model (GMM). An improved version of Fisher vectors proposed by Perronnin *et al.* [16] is as follows,

$$\mathcal{G}^{\mathbf{X}}_{\mu_k} = \frac{1}{N\sqrt{\pi_k}} \sum_{n=1}^{N} \gamma_n(k) \left(\frac{\mathbf{x}_n - \mu_k}{\sigma_k} \right), \tag{2}$$

$$\mathcal{G}^{\mathbf{X}}_{\sigma_k} = \frac{1}{N\sqrt{2\pi_k}} \sum_{n=1}^{N} \gamma_n(k) \left[\frac{(\mathbf{x}_n - \mu_k)^2}{\sigma_k^2} - 1 \right], \tag{3}$$

where $\gamma_n(k)$ is the weight of local feature $\mathbf{x}_n$ for the i-th Gaussian [16]. The final Fisher vectors representation is the concatenation of all the $\mathcal{G}^{\mathbf{X}}_{\mu_k}$ and $\mathcal{G}^{\mathbf{X}}_{\sigma_k}$ which is a $2Kd$ dimensional super vector.

B. Large Margin Dimension Reduction

Both VLAD and FV representations are high-dimensional, which can be storage-consuming and time-consuming for further processing. To obtain compact and discriminative video representations, we propose the larger margin dimensionality reduction method to compress these high-dimensional vectors.

Without loss of generality, we consider the VLAD representation here, and it can be generalized to FV. Suppose the VLAD for video pairs are ϕ_i and ϕ_j, respectively. We aim to learn a discriminative projection matrix $\mathbf{U} \in \mathbb{R}^{p\times Kd}, p \ll Kd$, which projects $\phi \in \mathbb{R}^{Kd}$ to low-dimensional one $\mathbf{U}\phi \in \mathbb{R}^p$, such that the inner product $< \mathbf{U}\phi_i, \mathbf{U}\phi_j >$ between the representations of video pairs is larger than a learnt threshold $b \in \mathbb{R}$ if the video pairs present the same type of action, and smaller otherwise. We further impose that these conditions are satisfied with a margin of at least one w.r.t b (see the 4th column in Fig. 1), resulting in the constraints:

$$y_{ij}(\phi_i^{\mathrm{T}}\mathbf{U}^{\mathrm{T}}\mathbf{U}\phi_j - b) > 1, \quad y_{ij} \in \{+1, -1\} \tag{4}$$

where $y_{ij} = 1$ denotes the video pairs contain the same type of action, and $y_{ij} = -1$ otherwise.

To learn the projection matrix $\mathbf{U}$, we leverage the hinge-loss function and optimize the following objective function:

$$\arg\min_{U,b} \sum_{(i,j)} \max\{1 - y_{ij}(\phi_i^{\mathrm{T}}\mathbf{U}^{\mathrm{T}}\mathbf{U}\phi_j - b), 0\} \tag{5}$$

Note that Eq.(5) is different from standard SVM, as it is a quadratic optimization about matrix $\mathbf{U}$. The optimal $\{U, b\}$ can be found by stochastic sub-gadient descent method. At each iteration t, we randomly sample a batch of video pairs $\mathcal{V}$ and perform the following update for the projection matrix:

$$\mathbf{U}_{t+1} = \begin{cases} \mathbf{U}_t, \text{if } y_{ij}(\phi_i^T\mathbf{U}^T\mathbf{U}\phi_j - b) > 1, \forall (i,j) \in \mathcal{V} \\ \mathbf{U}_t + \lambda \sum_{(i,j)} (-2y_{ij}\phi_i^T\phi_j\mathbf{U}), \text{ otherwise} \end{cases} \tag{6}$$

where $(-2y_{ij}\phi_i^T\phi_j\mathbf{U})$ is the sub-gradient from video pairs (i,j) in the batch $\mathcal{V}$, and λ is a given learning rate. $\mathbf{U}$ is initialized by the p largest PCA-Whitened dimensions U_0 [14], [10], [19], and b by the mean of $< \phi_i\mathbf{U}_0, \mathbf{U}_0\phi_j >$ empirically. The final b is discarded, and we keep the projection $\mathbf{U}$. Generally, $\mathbf{U}$ can be regularized by $\|\mathbf{U}_i\| = 1$ to prevent scaling $\mathbf{U}$. We call it as *Constrained LMDR* (C-LMDR). An evaluation is given in Section III-B.

Relation to previous methods. Our LMDR learns a discriminative projection matrix to increase the similarities of video pairs with "same" annotations and meanwhile separate

TABLE I
THE PERFORMANCE (ACC.±SE(AUC)) OF COMPACT VLAD BY PCA AND THE PROPOSED LMDR WITH CHANGING COMPRESSION RATIO.

Compression Ratio	Dim. (Energy)	PCA	LMDR
4,655	11 (10%)	58.72±3.94 (62.69)	63.60±2.39 (69.65)
826	62 (20%)	60.28±3.58 (64.80)	65.77±2.62 (72.37)
275	186 (30%)	60.25±3.47 (64.82)	65.72±2.16 (72.58)
139	369 (40%)	60.27±3.48 (64.82)	66.63±2.36 (72.93)
85	600 (50%)	60.25±3.47 (64.82)	66.83±2.65 (73.26)
59	874 (60%)	60.25±3.47 (64.82)	**67.13±2.84 (73.53)**
43	1,194 (70%)	60.25±3.47 (64.82)	66.42±2.55 (73.46)
1	51,200 (100%)	Baseline (Full Dim.): 61.38±3.25 (66.39)	

TABLE II
THE PERFORMANCE OF COMPACT FISHER VECTOR BY PCA AND THE PROPOSED LMDR WITH CHANGING COMPRESSION RATIO.

Compression Ratio	Dim. (Energy)	PCA	LMDR
5,389	19 (10%)	63.10±2.50 (68.98)	66.20±2.63 (72.70)
883	116 (20%)	63.40±2.83 (69.30)	67.52±3.30 (74.49)
354	289 (30%)	63.43±2.81 (69.30)	67.82±2.90 (74.82)
200	513 (40%)	63.43±2.82 (69.30)	68.37±3.07 (**75.40**)
132	775 (50%)	63.43±2.82 (69.30)	**68.72**±2.95 (75.26)
96	1,067 (60%)	63.43±2.82 (69.30)	68.08±2.95 (74.94)
1	102,400 (100%)	Baseline (Full Dim.): 63.75±2.03 (69.28)	

those video pairs with "not-same" annotations by a large margin. This is closely related to CSML [14] and distance metric learning of *large margin nearest neighbor* (LMNN) [20], [21], [22]. The CSML encourages the margin between positive and negative samples to be as large as possible, which tries to maximize the following objective:

$$g(U) = \sum_{\forall y_{ij}=1} CS(\mathbf{U}\phi_i, \mathbf{U}\phi_j) - \alpha \sum_{\forall y_{ij}=-1} CS(\mathbf{U}\phi_i, \mathbf{U}\phi_j) \tag{7}$$

where $CS(x, y) = x^T y / \|x\| \|y\|$ denotes cosine similarity. We also use similarity measure here, but unlike CSML that ensures the margin between the sum of positive and negative samples, we try to guarantee the margin between each positive and negative sample to be larger than 2 (see Eq.(4) and Fig.1).

LMNN learns a full-rank Mahalanobis matrix based on a distance measure [20]. This is impractical in our case since the dimension of FV is more than 100,000. We note there exists another method also named as LMDR in [22]. It jointly learned the Mahalanobis matrix and L1-norm SVM for image classification. The L1-norm SVM served as feature selection which can lead to low dimension. This method suffers the same problem as LMNN.

III. EXPERIMENTS

A. Experimental Protocol and Setup

To validate the proposed method, we conduct extensive experiments on the widely-used ASLAN benchmark [8]. The ASLAN dataset contains 3,631 action videos collected from the web, to a total of 432 action classes. The protocol used for the experiments is the 10-fold leave-one-out cross-validation scheme. The dataset is divided into ten splits, each of which includes 300 pairs of same-type actions and 300 pairs of different-type actions. In each experiment, nine splits are used for training, with the remaining split used for testing. Results are reported by calculating a ROC curve and measuring both the area under curve (AUC) and the averaged accuracy ± standard errors for the ten splits.

We extract improved dense trajectories (IDT) using the code from Wang [4]. Each trajectory is described by concatenating HOG, HOF, and MBH descriptors, which is a 396-dimensional vector. We reduce the dimensionality of these descriptors to 200 by performing PCA and Whitening, and fix the codebook size to 256 for both VLAD and FV as usual in the literature [4]. Therefore, the original VLAD and FV are 51,200 and 102,400 dimensional, respectively. The batch size of stochastic sub-gradient descent is fixed to 10 and λ to 0.01.

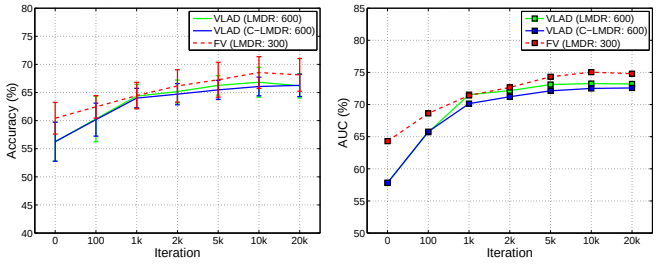

Fig. 2. The performance of LMDR for FV and VLAD with changing iterations. $p = 300$ for FV and $p = 600$ for VLAD. C-LMDR denotes the constrained LMDR. The proposed LMDR achieves best performance at iteration 10k for both FV and VLAD.

B. Results and Discussion

We first study the effect of changing dimensionality (based on the main energy from initialized PCA-Whitened matrix) for LMDR on VLAD and FV. We fix the iterations to 10k.

Dimensionality. For fair comparison, we take the original VLAD and FV as baselines for Table I and Table II, respectively. As shown in both Tables, for both VLAD and FV, our LMDR improves the baseline performance significantly with large compression ratio. Specially, it improves AUC by 7.14% with compression rate 59 for VLAD, and 6.12% with compression ratio 200 for FV. The performance changes slightly when the PCA preserving energy reaches 30%. Our LMDR method for VLAD obtains the best performance with compression rate 59, and for FV with compression ratio about 200. This indicates that VLAD and FV representations are highly redundant on the ASLAN dataset. Compared to PCA, our LMDR not only reduces the dimensionality significantly but also boosts the performance even with thousands of reduction rates.

Iterations. We also investigate the effect of iterations for the LMDR method in Fig. 2. We fix the p of our LMDR approach to 600 and 300 for VLAD and FV, respectively. The iteration 0 means that the PCA-Whitened matrix is used which is the initialization for $\mathbf{U}$. The optimal iteration for both VLAD and FV is 10k as can be observed from Fig. 2. Too many iterations may lead to overfitting which decreases the performance.

Constraint on U. To evaluate the impact of constraints on $\mathbf{U}$, we conduct an experiment using C-LMDR with VLAD. We apply augmented Lagrangian method to solve the constrained optimization [23]. The results with varying iterations are shown in Fig. 2. We observe that the results of VLAD using both C-LMDR and LMDR are very similar. Compared with LMDR, C-LMDR needs more computational cost.

To gain further insight into our results, we illustrate the most confident predictions made by our LMDR method with

Fig. 3. The most confident results using LMDR (dim.: 513) with Fisher vector. The Same/Not-Same labels are the ground truth, and the Correct/Incorrect labels show whether our method predicted correctly.

TABLE III
COMPARISON TO THE CSML METHOD.

Method	No LMDR/CSML	PCA+CSML	LMDR
VLAD (%)	61.38±3.25 (66.39)	64.18±2.05 (70.28)	**67.13**±2.84 (**73.53**)
FV (%)	63.75±2.03 (69.28)	66.17±1.83 (72.58)	**68.37**±3.07 (**75.40**)

TABLE IV
COMPARISON TO THE STATE-OF-THE-ART RESULTS.

Method	[8]	[9]	[10]	[11]	VLAD+LMDR	FV+LMDR
Acc. (%)	60.88	64.25	65.45	66.13	67.17	**68.72**
AUC (%)	65.30	69.10	71.92	73.23	73.53	**75.40**

FV. Fig. 3 presents the most confident correct/incorrect same and not-same predictions. These mistakes mainly result from changing viewpoints, complex backgrounds, ill-defined action categories, and pose ambiguity of inter-class.

C. Comparison

We compare our LMDR method to the popular PCA+CSML approach [14] in Table III. We implement the PCA+CSML method following [10]. We learn a full rank projection matrix for CSML, and therefore the projection matrix from CSML is a square matrix with the same size as PCA dimension. We sweep the same PCA dimensions as in Table I and Table II, and find that the best results from both VLAD and FV are obtained at the same dimensions as for our LMDR. From Table III, we observe that our method outperforms the CSML significantly. Note that, unlike CSML, the proposed LMDR method jointly performs dimensionality reduction and similarity learning which preserves the discriminative information effectively.

Table IV compares our best results with the state-of-the-art performance. Most of these methods make use of STIP [8], [9], MIP [10], and dense trajectory [11] features or the combination of all the features [11] with traditional BoW model and CSML. Both VLAD and FV encoding methods with our LMDR obtain state-of-the-art results on the ASLAN benchmark.

IV. CONCLUSION

This paper proposes a large margin dimensionality reduction method to compress the Fisher vector and VLAD representation for action similarity labeling. Our method can not only reduce the dimension significantly but also improve the performance for both FV and VLAD. Experimental results show that the proposed method achieves superior performance than state-of-the-art methods on ASLAN benchmark. Our code is available at http://mmlab.siat.ac.cn/personal/pxj/.

REFERENCES

[1] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *CVPR*, 2008, pp. 1–8.
[2] J. Aggarwal and M. S. Ryoo, "Human activity analysis: A review," *ACM Computing Surveys*, vol. 43, no. 3, p. 16, 2011.
[3] M. Jain, H. Jégou, P. Bouthemy *et al.*, "Better exploiting motion for better action recognition," in *CVPR*, 2013.
[4] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *ICCV*, 2013.
[5] X. Peng, Y. Qiao, Q. Peng, and X. Qi, "Exploring motion boundary based sampling and spatial-temporal context descriptors for action recognition," in *BMVC*, 2013, pp. 1–11.
[6] L. Wang, Y. Qiao, and X. Tang, "Latent hierarchical model of temporal structure for complex activity classification," *TIP*, vol. 23, no. 2, pp. 810–822, 2014.
[7] Y. Tian, R. Sukthankar, and M. Shah, "Spatiotemporal deformable part models for action detection," in *CVPR*, 2013.
[8] O. Kliper-Gross, T. Hassner, and L. Wolf, "The action similarity labeling challenge," *PAMI*, vol. 34, no. 3, pp. 615–621, 2012.
[9] O. Kliper-Gross, T. Hassner, and L. Wolf, "One shot similarity metric learning for action recognition," in *Similarity-Based Pattern Recognition*, 2011, pp. 31–45.
[10] O. Kliper-Gross, Y. Gurovich, T. Hassner, and L. Wolf, "Motion interchange patterns for action recognition in unconstrained videos," in *ECCV*, 2012, pp. 256–269.
[11] Y. Hanani, N. Levy, and L. Wolf, "Evaluating new variants of motion interchange patterns," in *CVPRW*, 2013, pp. 263–268.
[12] J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos," in *ICCV*, 2003, pp. 1470–1477.
[13] L. Wolf, T. Hassner, and Y. Taigman, "The one-shot similarity kernel," in *ICCV*, 2009, pp. 897–902.
[14] H. V. Nguyen and L. Bai, "Cosine similarity metric learning for face verification," in *ACCV*, 2011, pp. 709–720.
[15] X. Wang, L. Wang, and Y. Qiao, "A comparative study of encoding, pooling and normalization methods for action recognition," in *ACCV*, 2012.
[16] F. Perronnin, J. Sánchez, and T. Mensink, "Improving the fisher kernel for large-scale image classification," in *ECCV*, 2010, pp. 143–156.
[17] H. Jégou, M. Douze, C. Schmid, and P. Pérez, "Aggregating local descriptors into a compact image representation," in *CVPR*, 2010, pp. 3304–3311.
[18] T. Jaakkola, D. Haussler *et al.*, "Exploiting generative models in discriminative classifiers," *NIPS*, pp. 487–493, 1999.
[19] K. Simonyan, O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Fisher vector faces in the wild," in *BMVC*, 2013.
[20] K. Q. Weinberger and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," *JMLR*, vol. 10, pp. 207–244, 2009.
[21] L. Torresani and K.-c. Lee, "Large margin component analysis," *NIPS*, vol. 19, p. 1385, 2007.
[22] K. Huang and S. Aviyente, "Large margin dimension reduction for sparse image classification," in *Statistical Signal Processing, 2007. IEEE/SP 14th Workshop on*, 2007, pp. 773–777.
[23] J. Nocedal and S. J. Wright, *Numerical Optimization*, 2nd ed. Springer, 2006.