Smooth Sparse Coding via Marginal Regression for Learning Sparse Representations- Supplementary material

1. Data set Description

1.1. CMU Multi-pie face recognition:

The face recognition experiment was conducted on the CMU Multi-PIE dataset. The dataset is challenging due to the large number of subjects and is one of the standard data sets used for face recognition experiments. The data set contains 337 subjects across simultaneous variations in pose, expression, and illumination. We ignore the 88 subjects that were considered as outliers in (Yang et al., 2010) and used the rest of the images for our face recognition experiments. We follow (Yang et al., 2010) and use the 7 frontal extreme illuminations from session one as train set and use other 20 illuminations from Sessions 2-4 as test set.

1.2. 15 Scenes Categorization:

We also conducted scene classification experiments on the 15-Scenes data set. This data set consist of 4485 images from 15 categories, with the number of images each category ranging from 200 to 400. The categories corresponds to scenes from various settings like kitchen, living room etc. Similar to the previous experiment, we extracted patches from the images and computed the SIFT features corresponding to the patches. The categorization results are reported in Table 2. The accuracy using smooth sparse codes is better than previous reported results on this data set using standard sparse coding techniques for e.g., (Yang et al., 2009).

1.3. Caltech-101 Data set:

The Caltech-101 data set consists of images from 101 classes like animals, vehicles, flowers, etc. The number of images per category varies from 30 to 800. Most images are of medium resolution (300 × 300). All images are used a gray-scale images. Following previous standard experimental settings for Caltech-101 data set, we use 30 images per category and test on the rest. Average classification accuracy normalized by class frequency is used for evaluation. Similar to the previous experiment, we extracted patches from the images and computed the SIFT features corresponding to the the patches. Table 2 shows the accuracy of sparse coding and smooth sparse coding. Note that sparse coding on SIFT achieves one of the best results on the Caltech-101 data set. The proposed smoothing approach further improves the accuracy and achieves competitive results on this benchmark data set.

1.4. Activity recognition

The KTH action dataset consists of 6 human action classes. Each action is performed several times by 25 subjects and is recorded in four different scenarios. In total, the data consists of 2391 video samples. The YouTube actions data set has 11 action categories and is more complex and challenging (Liu et al., 2009). It has 1168 video sequences of varied illumination, background, resolution etc. We randomly densely sample blocks (400 cuboids) of video from the data sample and extract HOG-3d features and constructed the video features as described above. .

1.5. Youtube person data set

Similar to the experiments using the feature smoothing kernel, in this section we report results on experiment conducted using the time smoothed kernel. Specifically, we used the YouTube person data set (Kim et al., 2008) in order to recognize people, based on time-based kernel smooth sparse coding. The dataset contains 1910 sequences of 47 subjects. The approach for this experiment is similar to (Yang et al., 2009). We extracted SIFT descriptors for every 16×16 patches sampled on a grid of step size 8. Then we use smooth sparse coding with time kernel to learn the codes and max pooling to get the final representation of a video sample. Pre-processing steps like face extraction or face tracking was not used in this experiment. Finally, linear svm was used for classification of video sequences based on person present in the video sequences.

2. Experiments using Temporal Smoothing

In this section we describe an experiment conducted using the temporal smoothing kernel on the Youtube persons dataset. We extracted SIFT descriptors for every 16×16 patches sampled on a grid of step size 8 and used smooth sparse coding with time kernel to learn the codes and max pooling to get the final video representation. We avoided pre-processing steps such as face extraction or face tracking. Note that in the previous action recognition video experiment, video blocks were densely sampled and used for extracting HoG-3d features. In this experiment, on the other hand, we extracted SIFT features from individual frames and used the time kernels to incorporate the temporal information into the sparse coding process.

Method	Fused Lasso	SC	SSC-tricube
Accuracy	68.59	65.53	71.21

Table 1. Linear SVM accuracy for person recognition task from YouTube face video dataset.

For this case, we also compared to the more standard fused-lasso based approach (Tibshirani et al., 2005). Note that in fused Lasso based approach, in addition to the standard L_1 penalty, an additional L_1 penalty on the difference between the neighboring frames for each dimensions is used. This tries to enforce the assumption that in a video sequence, neighboring frames are more related to one another as compared to frames that are farther apart.

Table 1 shows that smooth sparse coding achieved higher accuracy than fused lasso and standard sparse coding. Smooth sparse coding has comparable accuracy on person recognition tasks to other methods that use face-tracking, for example (Kim et al., 2008). Another advantage of smooth sparse coding is that it is significantly faster than sparse coding and the used lasso.

3. Generalization bounds for learning problems

In this section, we provide two generalization bounds for learning problems, corresponding to slow-rates and fast rates, based on covering numbers. We first state the following general lemma regarding generalization error bounds with slow rates for a learning problem with given covering number bounds.

Lemma 1 ((Vainsencher et al., 2011) **).** *Let $\mathcal{Q}$ be a function class of $[0, B]$ functions with covering number $(\frac{C}{\epsilon})^d > \frac{e}{B^2}$ under $|\cdot|_\infty$ norm. Then for every $t > 0$ with probability at least $1-e^{-t}$, for all $f \in \mathcal{Q}$, we have:*

$$\mathsf{E}\, f \leq E_n f + B\left(\sqrt{\frac{d\ln(C\sqrt{n})}{2n}} + \sqrt{\frac{t}{2n}}\right) + \sqrt{\frac{4}{n}}.$$

Next, we state general lemma regarding generalization error bounds with fast rates

Lemma 2 ((Vainsencher et al., 2011) **).** *Let $\mathcal{Q}$ be a function class of $[0, 1]$ functions that can be covered for any $\epsilon > 0$ by at most $(C/\epsilon)^d$ balls of radius ϵ in the $|\cdot|_\infty$ metric, where $C \geq e$ and $\beta > 0$. Then with probability at least $1 - \exp{(-t)}$ we have for all functions $f \in \mathcal{Q}$,*

$$\mathsf{E}\, f \leq (1+\beta)E_n f + K(d, m, \beta)\frac{d\ln(Cm) + t}{n},$$

where $K(d, m, \beta) = \sqrt{2\left(\frac{9}{\sqrt{n}} + 2\right)\left(\frac{d+3}{3d}\right) + 1} + \left(\frac{9}{\sqrt{n}} + 2\right) + \left(\frac{d+3}{3d}\right) + 1 + \frac{1}{2\beta}$.

Note that $K(d, m, \beta)$ is non-increasing in d, m as a consequence of which we immediately have the following corollary, which we use in the statement of our main theorem for fast rates.

Corollary 1. *Let $\mathcal{Q}$ be as above. For $d \geq 20$, $m \geq 5000$ and $\beta = 0.1$, we have with probability at least $1 - \exp{(-t)}$ for all functions $f \in \mathcal{Q}$,*

$$Ef \leq (1.1)E_n f + 9\frac{d\ln(Cm) + t}{n}.$$

The proofs of Lemma 1 and Lemma 2 could be found in (Vainsencher et al., 2011). Obtaining generalization bounds for the problem under consideration follows directly, given the above two general statements and our theorem on covering numbers (Theorem 1).

References

Kim, M., Kumar, S., Pavlovic, V., and Rowley, H. Face tracking and recognition with visual constraints in real-world videos. In *CVPR*, 2008.

Liu, J., Luo, J., and Shah, M. Recognizing realistic actions from videos in the wild. In *CVPR*, 2009.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. Sparsity and smoothness via the fused lasso. *JRSS:B*, 67, 2005.

Vainsencher, D., Mannor, S., and Bruckstein, A.M. The sample complexity of dictionary learning. *JMLR*, 2011.

Yang, J., Yu, K., Gong, Y., and Huang, T. Linear spatial pyramid matching using sparse coding for image classification. In *CVPR*, 2009.

Yang, J., Yu, K., and Huang, T. Supervised translation-invariant sparse coding. In *CVPR*, 2010.