Robust Feature Extraction to Utterance Fluctuation of Articulation Disorders Based on Random Projection

Toshiya Yoshioka, Tetsuya Takiguchi, Yasuo Ariki

Graduate School of System Informatics, Kobe University, Japan
yoshioka@me.cs.scitec.kobe-u.ac.jp, takigu@kobe-u.ac.jp, ariki@kobe-u.ac.jp

Abstract

We investigated the speech recognition of a person with an articulation disorder resulting from the athetoid type of cerebral palsy. The articulation of the first speech tends to become unstable due to strain on speech-related muscles, and that causes degradation of speech recognition. In this paper, we introduce a robust feature extraction method based on PCA (Principal Component Analysis) and RP (Random Projection) for dysarthric speech recognition. PCA-based feature extraction performs reducing the influence of the unstable speaking style caused by the athetoid symptoms. Moreover, we investigate the feasibility of random projection for feature transformation in order to gain more performance in dysarthric speech recognition task. Its effectiveness is confirmed by word recognition experiments.

Index Terms: articulation disorders, speech recognition, PCA, random projection, ROVER

1. Introduction

Recently, the importance of information technology in the welfare-related fields has increased. For example, sign language recognition using image recognition technology [1][2][3], text-reading systems from natural scene images [4][5][6], and the design of wearable speech synthesizers for voice disorders [7] [8] have been studied.

There are 34,000 people with speech impediments associated with articulation disorders in Japan alone, and it is hoped that speech recognition systems will one day be able to recognize their voices. One of the causes of speech impediments is cerebral palsy. Cerebral palsy results from damage to the central nervous system, and the damage causes movement disorders. Three general times are given for the onset of the disorder: before birth, at the time of delivery, and after birth. Cerebral palsy is classified as follows: 1) spastic type 2) athetoid type 3) ataxic type 4) atonic type 5) rigid type, and a mixture of types [9].

In this paper, we focused on a person with an articulation disorder resulting from the athetoid type of cerebral palsy. Athetoid symptoms develop in about 10-15% of cerebral palsy sufferers. In the case of a person with this type of articulation disorder, the first movements are sometimes more unstable than usual. That means, in the case of speaking-related movements, the first utterance is often unstable or unclear due to the athetoid symptoms, and that causes degradation of speech recognition. Therefore, we recorded speech data for a person with an articulation disorder who uttered each of the words five times, and investigated the influence of the unstable speaking style caused by the athetoid symptoms.

The goal of front-end speech processing in ASR is to obtain a projection of the speech signal to a compact parameter space where the information related to speech content can be extracted. In current speech recognition technology, MFCC (Mel-Frequency Cepstrum Coefficient) is being widely used. The feature is uniquely derived from the mel-scale filter-bank output by DCT (Discrete Cosine Transform). The low-order MFCCs account for the slowly changing spectral envelope, while the high-order ones describe the fast variations of the spectrum. Therefore, a large number of MFCCs is not used for speech recognition because we are only interested in the spectral envelope, not in the fine structure. In [10], PCA-based feature extraction has been studied. Also, [11] proposed a robust feature extraction method based on PCA instead of DCT in a dysarthric speech recognition task, where the main stable utterance element is projected onto low-order features while fluctuation elements of speech style are projected onto high-order ones. Therefore, the PCA-based filter will be able to extract stable utterance features only (Fig. 1). The proposed method improved the recognition accuracy, but the performance was not sufficient when compared to that of persons with no disability.

Random projection has been suggested as a means of space mapping, where a projection matrix is composed of the columns defined by the random values chosen from a probability distribution. In addition, the Euclidean distance of any two points is approximately preserved through the projection. Therefore, random projection has also been suggested as a means of dimensionality reduction [12]. In contrast to conventional techniques such as PCA, which find a subspace by optimizing certain criteria, random projection does not use such criteria; therefore, it is data independent. Moreover, it represents a computationally simple and efficient method that preserves the structure of the data without introducing significant distortion [13]. Goel et al [13] have reported that random projection has been applied to various types of problems, including information retrieval (e.g., [14]), image processing (e.g., [15][16]), machine learning (e.g., [17][18][19]), and so on. Although it is based on a simple idea, random projection has demonstrated good performance in a number of applications, yielding results comparable to conventional dimensionality reduction techniques, such as PCA.

The main contributions of this paper are the following. Firstly, we introduce a PCA-based feature extraction approach to extract stable utterance features only. Secondly, PCA-based features are projected using various random matrices. Then, we use the same number of dimensions for the projected space as that of the original space. There may be some possibility of finding a random matrix that gives better speech recognition accuracy among random matrices, since we are able to produce various RP-based features (using various random matrices). Therefore, a vote-based combination method is introduced in order to obtain an optimal result from many (infinite) random matrices, where ROVER combination [20] is applied to the results from the ASR systems created from each RP-based

SLPAT 2013, 4th Workshop on Speech and Language Processing for Assistive Technologies, pages 129–133,
Grenoble, France, 21–22 August, 2013. ©2013 Association for Computational Linguistics

feature.

The rest of this paper is organized as follows. Section 2 describes a PCA-based feature extraction method. In Section 3, the proposed feature projection method using random orthogonal matrices, and, a vote-based combination method are explained. Results and discussion for the experiments on a dysarthric speech recognition task are given in Section 4. Section 5, concludes the paper with a summary of our proposed method, contribution, and future work.

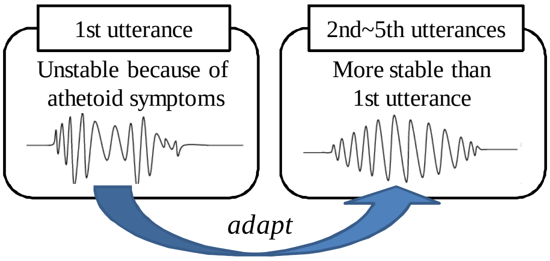
1st utterance
Unstable because of athetoid symptoms
2nd~5th utterances
More stable than 1st utterance
adapt

Figure 1: Corrective strategy for articulation disorders.

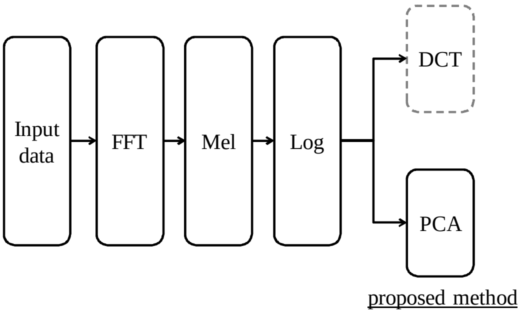

Figure 2: Feature extraction using PCA.

2. Feature extraction using PCA

Robust feature extraction was proposed based on PCA with the more stable utterance data instead of DCT (Fig. 2), where PCA is applied to the mel-scale filter bank output [11].

In this paper, we computed the filter (eigenvector matrix) using the more stable utterance. Then we applied the filtering operation to the first utterance (unstably articulated utterance) in the log-spectral domain. Given the frame of short-time analysis t and frequency ω, we represent the first utterance $\mathbf{Y}_t(\omega)$ as the multiplication of the stable speech $\mathbf{X}_t(\omega)$ and the fluctuation element of speaking style $\mathbf{H}(\omega)$ in the linear-spectral domain:

$$\mathbf{Y}_t(\omega) = \mathbf{X}_t(\omega) \cdot \mathbf{H}(\omega) \tag{1}$$

The multiplication can be converted to addition in the log-spectral domain as follows:

$$\log \mathbf{Y}_t(\omega) = \log \mathbf{X}_t(\omega) + \log \mathbf{H}(\omega) \tag{2}$$

Next, we use the following filtering based on PCA in order to extract the feature of stable speech only:

$$\hat{\mathbf{X}} = \mathbf{V}^T \mathbf{Y}_{log} \tag{3}$$

For the filter (eigenvector matrix), $\mathbf{V}$ is derived by the eigenvalue decomposition of the centered covariance matrix of a stable speech data set, in which the filter consists of the eigenvectors corresponding to the D dominant eigenvalues.

3. Proposed method

3.1. RP-based feature projection method

This section presents a feature projection method using random orthogonal matrices. The main idea of random projection arises from the Johnson-Lindenstrauss lemma [21]; namely, if original data are projected onto a randomly selected subspace using a random matrix, then the distances between the data are approximately preserved.

Random projection is a simple yet powerful technique, and it has another benefit. Dasgupta [17] has reported that even if distributions of original data are highly skewed (have ellipsoidal contours of high eccentricity); their transformed counterparts will be more spherical.

First, we choose an n-dimensional random vector, $\mathbf{p}$, and let $\mathbf{P}^{(l)}$ be the l-th $n \times d$ matrix whose columns are vectors, $\mathbf{p}_1^{(l)}$, $\mathbf{p}_2^{(l)}$, ..., $\mathbf{p}_d^{(l)}$. Then, an original n-dimensional vector, $\mathbf{x}$, is projected onto a d-dimensional subspace using the l-th random matrix, $\mathbf{P}^{(l)}$, where we compute a d-dimensional vector, $\mathbf{x}'$, whose coordinates are the inner products $\mathbf{x}'_1$ = $\mathbf{p}_1^{(l)} \cdot \mathbf{x}, \ldots, \mathbf{x}'_d$ = $\mathbf{p}_d^{(l)} \cdot \mathbf{x}$.

$$\mathbf{x}' = \mathbf{P}^{(l)^T} \mathbf{x} \tag{4}$$

In this paper, we investigate the feasibility of random projection for speech feature transformation. As described above, a random projection from n dimensions to d (= n) dimensions is represented by an $n \times d$ matrix, $\mathbf{P}$. It has been shown that if the random matrix $\mathbf{P}$ is chosen from the standard normal distribution (with mean 0 and variance 1, referred to as $N(0,1)$), then the projection preserves the structure of the data [21]. In this paper, we use $N(0,1)$ for the distribution of the coordinates. The random matrix, $\mathbf{P}$, can be calculated using the following algorithm [13][17].

- Choose each entry of the matrix from an independent and identically distributed (i.i.d.) $N(0,1)$ value.
- Make the orthogonal matrix using the Gram-Schmidt algorithm, and then normalize it to unit length.

Orthogonality is effective for feature extraction because the HMMs used in speech recognition experiments consist of diagonal covariance matrices. Fig. 3 shows examples of random matrices from $N(0,1)$.

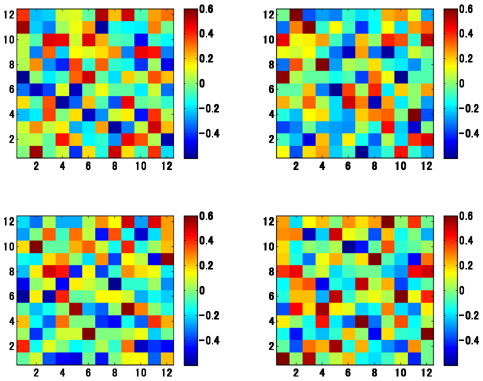

Figure 3: Examples of random matrices 12 dim (12 × 12).

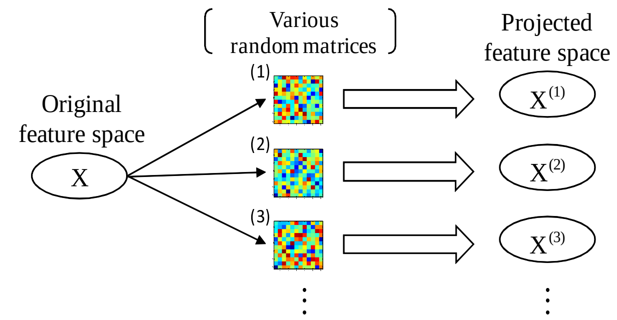

Figure 4: Random projection on the feature domain. An original feature is transformed to various features using various random matrices. (Eq. 4)

3.2. Vote-based combination

As mentioned in the previous section, we can make many (infinite) random matrices from $N(0, 1)$ (Fig. 4). Since there may be some possibility of finding a random matrix that gives better performance, we will have to select the optimal matrix or the optimal recognition result from them. To obtain the optimal result, a majority vote-based combination is introduced in this paper, where ROVER combination is applied to the results from the ASR systems created from each RP-based feature.

Fig. 5 shows an overview of the vote-based combination. First, random matrices, $\mathbf{P}^{(l)}$ (l = 1, ..., L), are chosen from the standard normal distribution, with mean 0 and variance 1. Speech features are projected using each random matrix. An acoustic model corresponding to each random matrix is also trained. For the test utterance, using each acoustic model, an ASR system outputs the best scoring word by itself. To obtain an optimal result from among all the results for random projection, voting is performed by counting the number of occurrences of the best word for each RP-based feature.

For example, in the case of L = 20, 20 kinds of new feature vectors are calculated using 20 kinds of random matrices. Then, we train the 20 kinds of acoustic models using 20 kinds of new feature vectors. In the test process, 20 kinds of recognition results are obtained using 20 kinds of acoustic models. To obtain a single hypothesis from among 20 kinds of recognition results, voting is performed.

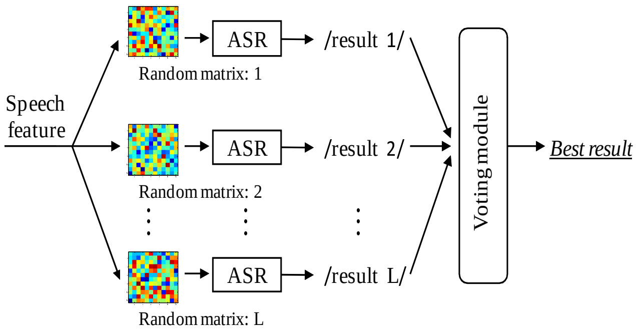

Figure 5: Overview of the vote-based combination.

4. Evaluation

4.1. Experimental conditions

The proposed method was evaluated on a word recognition task for one male with an articulation disorder. For the conducted experiments, we recorded 210 words included in the ATR Japanese speech database. Each of the 210 words was repeated five times (Fig. 6). The speech signal was sampled at 16 kHz and windowed with a 25-msec Hamming window every 10 msec.

It was difficult to recognize an utterance of an articulation disorder using an acoustic model trained by utterances of physically unimpaired persons. Therefore, in this paper, we trained the acoustic model using the utterances of a person with an articulation disorder. When we recognized the 1st utterance, the 2nd through 5th utterances were used for training. We iterated this process for each utterance. The acoustic models consist of a HMM set with 54 context-independent phonemes and 8 mixture components for each state. Each HMM has three states and three self-loops.

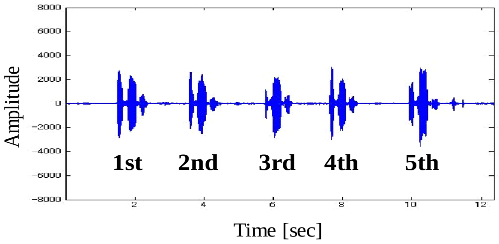

Figure 6: Example of recorded speech data.

4.2. Experiment 1

In Experiment1, recognition results were obtained for each utterance of a person with an articulation disorder using speaker-dependent model.

The system was trained using 24-dimensional feature vectors consisting of 12-dimensional MFCC parameters, along with their delta parameters.

Table 1: Recognition results [%] for each utterance in Experiment 1

1st	2nd	3rd	4th	5th
75.7	86.7	92.9	90.5	88.6

Table 1 shows the results obtained in Experiment 1. In a person with an articulation disorder, the recognition rate for the 1st utterance was 75.7%. As can be seen in Table 1, it was significantly lower than other utterances. It is considered that the speaker experiences a more strained state during the first utterance compared to subsequent utterances because the first utterance is the first intentional movement. Therefore, athetoid symptoms occur and articulation becomes difficult. It is believed that this difficulty causes fluctuations in speaking style and degradation of the recognition rates.

4.3. Experiment 2

The aim of Experiment 2 is to evaluate the improvement introduced by the use of a PCA-based feature extraction method. For Experiment 2, PCA was applied to 24 mel-scale filter bank output. Then, we computed the filter $\mathbf{V}$ using the 2nd through 5th utterances (the more stable utterances). We experimented on the number of principal components, using 11, 13, 15, 17, and 19 dimensions. Then, the delta coefficients were also computed. Comparison results between the baseline method (MFCC) and the PCA-based feature extraction method for the 1st utterance were shown in Fig. 7.

As can be seen in Fig. 7, the use of PCA instead of DCT improved the recognition rate for the 1st utterance from 76.7% (15-dimensional MFCC and their delta) to 80.5% (17-principal components and their delta). This results gives the evidence of the improvement introduced by the use of PCA instead of DCT when dealing with the 1st utterance. In addition, the recognition rates of the other utterances were equal to those of MFCC.

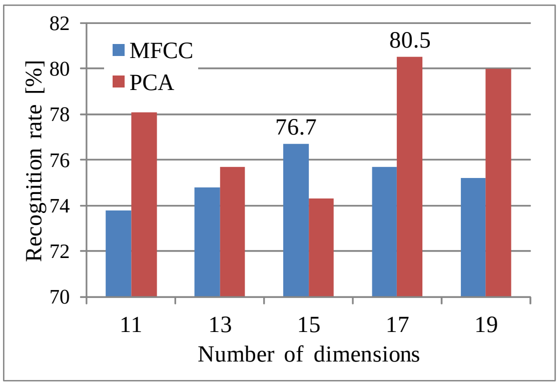

Figure 7: Comparison of DCT and PCA for the 1st utterance in Experiment 2.

4.4. Experiment 3

In order to test the effectiveness of a RP-based feature projection method, in Experiment 3, two RP-based features were evaluated. Each feature description was found below:

1. PCA[17]→RP[17] + ΔRP[17]:
   Random projection is applied to PCA-based features at the t-th frame, $\mathbf{x}(t) \in \mathrm{R}^{17}$, and the new feature, $\mathbf{y}(t) \in \mathrm{R}^{17}$, is obtained.

   $$\mathbf{y}(t) = \mathbf{P}^{(l)^T}\mathbf{x}(t) \qquad (5)$$

   Then, the new feature also has the delta parameter of projected feature, $\mathbf{y}(t)$. The final system feature dimensionality is 34.

2. PCA[17]→RP[17] + ΔPCA[17]:
   Random projection is applied to PCA-based features, $\mathbf{x}(t) \in \mathrm{R}^{17}$, and the new feature, $\mathbf{y}(t) \in \mathrm{R}^{17}$, is obtained. Then, the new feature also has the delta coefficient of original feature, $\mathbf{x}(t)$. The final system feature dimensionality is 34.

We investigated the performance of random projections for various random matrices (l = 20, 40, 60, 80, and 100) sampled from $N(0, 1)$. Tables 2 and 3 show the recognition rate versus the number of random matrices for each feature. The results of "RP w/o combination" show the maximums, means, and minimums obtained from each random projection without ROVER-based combination.

Table 2: Word recognition rate (%) for the 1st utterances using feature 1 in various random matrices. (The recognition rate of PCA-based features is 80.5%)

Number of random matrices	RP combination based on ROVER	RP w/o combination		
		Max.	Mean	Min.
20	79.5%	80.5%	76.5 %	72.9%
40	80.0%	81.0%	76.8%	72.9%
60	80.5%	**83.3%**	76.8%	72.9%
80	80.5%	**83.3%**	76.8%	72.4%
100	80.5%	**83.3%**	76.8%	72.4%

Table 3: Word recognition rate (%) for the 1st utterances using feature 2 in various random matrices. (The recognition rate of PCA-based features is 80.5%)

Number of random matrices	RP combination based on ROVER	RP w/o combination		
		Max.	Mean	Min.
20	83.3%	81.9%	79.5%	76.7%
40	**85.2%**	83.8%	79.6%	71.9%
60	**85.2%**	83.8%	79.5%	71.9%
80	84.8%	83.8%	79.5%	71.9%
100	84.8%	83.8%	79.5%	71.9%

Table 2 shows the performance results obtained using feature 1 in Experiment 3. As can be seen in Table 2, the maximums of random projections without ROVER-based combination for 60, 80, and 100 random matrices were higher than the recognition rate of PCA-based features. However, even if ROVER-based combination is applied, we could not show further performance increases in our experiments using feature 1.

The recognition results obtained using feature 2 are shown in Table 3. As can be seen in Table 3, the results for feature 2 indicated that the vote-based random-projection combination improved the recognition rate from 80.5% (17-dimentional PCA and their delta) to 85.2% using the combination of 40 or 60 random matrices, although the means of random projections without combination for some random matrices was lower than the recognition rate of the original features.

We can see that the combination of random projection and ROVER outperforms both the baseline method (MFCCs) and the PCA-based feature extraction method. This result gives the evidence of the improvement introduced by the feature transformation based on random projection and the use of ROVER to obtain an optimal result. One of the possible reasons the random projection improves the recognition rates may be that if distributions of original data are skewed (have ellipsoidal contours of high eccentricity), their transformed counterparts will become more spherical [17]. However, there were 'bad' projections that cause degradation of speech recognition accuracy compared with the recognition of original features. Therefore, more research will be needed to investigate the effectiveness of the random projection methid for speech features.

5. Conclusions

As a result of this work, a method for recognizing dysarthric speech using a robust PCA-based feature extraction and transformation based on random projection has been developed. In the feature extraction, PCA is applied to the mel-scale filter bank output. It can be expected that PCA will project the main stable utterance elements onto low-order features, while elements associated with fluctuations in speaking style will be projected onto high-order features. Moreover, the proposed method transforms the PCA-based features using various random matrices. It also introduces a vote-based combination method to obtain an optimal result from the ASR systems created from each RP-based feature. Word recognition experiments were conducted to evaluate the proposed method for one male with an articulation disorder. The results of the experments showed that a method based on random projection outperformed both a baseline method (using MFCC) and a PCA-based feature extraction method.

As future work, we will continue to investigate how to select the optimal basis vector from a random matrix.

6. References

[1] T. Starner, J. Weaver, and A. Pentland, “Real-Time American Sign Language Recognition Using Desk and Wearable Computer Based Video, ” IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(12), pp. 1371–1375, 1998.

[2] J. Lin, W. Ying, and T.S. Huang, “Capturing human hand motion in image sequences,” IEEE Workshop on Motion and Video Computing, pp. 99–104, 2002.

[3] G. Fang, W. Gao, and D. Zhao, “Large vocabulary sign language recognition based on hierarchical decision trees, ” International Conference on Multimodal Interaction, pp. 125–131, 2003.

[4] V. Wu, R. Manmatha, and E. M. Riseman, “Textfinder: an automatic system to detect and recognize text images, ” IEEE Transactions on Pattern Analysis and Machine Inteligence, 21(11), pp. 1224–1229, 1999.

[5] M.K. Bashar, T. Matsumoto, Y. Takeuchi, H. Kudo, and N. Ohnishi, “Unsupervised Texture Segmentation via Wavelet-based Locally Orderless Images (WLOIs) and SOM,” International Conference on Computer Graphics and Imaging, pp. 279–284, 2003.

[6] N. Ezaki, M. Bulacu, and L. Schomaker, “Text Detection from Natural Scene Images: Towards a System for Visually Impaired Persons, ” International Conference on Pattern Recognition, pp. 683–686, 2004.

[7] T. Ohsuga, Y. Horiuchi, and A. Ichikawa, “Estimating Syntactic Structure from Prosody in Japanese Speech,” IEICE Transactions on Information and Systems, 86(3), pp. 558–564, 2003.

[8] K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, “Speaking Aid System for Total Laryngectomees Using Voice Conversion of Body Transmitted Artificial Speech,” Annual Conference of the International Speech Communication Association, pp. 1395–1398, 2006.

[9] S.T. Canale, and W.C. Campbell, “Campbell’s Operative Orthopaedics,” Mosby-Year Book, 2002.

[10] S-M. Lee, S-H. Fang, J-W. Hung, and L-S. Lee, “Improved MFCC Feature Extraction by PCA-Optimized Filter Bank for Speech Recognition,” IEEE Workshop on Automatic Speech Recognition and Understanding, pp. 49–52, 2001.

[11] H. Matsumasa, T. Takiguchi, Y. Ariki, I. LI, and T. Nakabayashi, “PCA-Based Feature Extraction for Fluctuation in Speaking Style of Articulation Disorders,” Annual Conference of the International Speech Communication Association, pp. 1150–1153, 2007.

[12] Ella Bingham, and Heikki Mannila, “Random projection in dimensionality reduction: applications to image and text data, ” Knowledge Discovery and Data Mining, pp. 245–250, 2001.

[13] N. Goel, G. Bebis, and A. Nefian, “Face recognition experiments with random projection,” Storage and Retrieval for Image and Video Databases, pp. 426–437, 2005.

[14] P. Thaper, S. Guha, and N. Koudas, “Dynamic multidimensional histograms,” International Conference on Management of Data, pp. 428–439, 2002.

[15] L. Liu, P. Fieguth, G. Kuang, and H. Zha, “Sorted Random Projections for robust texture classification, ” IEEE International Conference on Computer Vision, pp. 391–398, 2011.

[16] H. T. Ho, and R. Chellappa, “Automatic head pose estimation using randomly projected dense SIFT descriptors, ” IEEE International Conference on Image Processing, pp. 153–156, 2012.

[17] S. Dasgupta, “Experiments with random projection,” Uncertainty in Artificial Intelligence, pp. 143–151, 2000.

[18] X.Z. Fern, and C.E. Brodley, “Random projection for high dimensional data clustering: A cluster ensemble approach, ” International Conference on Machine Learning, pp. 186–193, 2003.

[19] S. Lee, and A. Nedic, “Distributed Random Projection Algo-rithm for Convex Optimization, ” IEEE Journal of Selected Topics in Signal Processing, 7(2), pp. 221–229, 2013.

[20] J.G. Fiscus, “A post-processing system to yield reduced word error rates: Recogniser output voting error reduction (rover),” IEEE Workshop on Automatic Speech Recognition and Understanding, pp. 347–352, 1997.

[21] R.I. Arriaga, and S. Vempala, “An algorithmic theory of learning: robust concepts and random projection,” IEEE Symposium on Foundations of Computer Science, pp. 616–623, 1999.