Nonnegative Matrix Factorization and Its Application to Pattern Analysis and Text Mining

Jacek M. Zurada
Electrical and Computer Engineering
University of Louisville
Louisville, USA
Spoleczna Akademia Nauk,
90-011 Lodz, Poland
jacek.zurada@louisville.edu

Tolga Ensari
Computer Engineering
Istanbul University
Istanbul, Turkey
ensari@istanbul.edu.tr

Ehsan Hosseini Asl
Electrical and Computer Engineering
University of Louisville
Louisville, USA
ehsan.hosseiniasl@gmail.com

Jan Chorowski
TTA Techtra
ul. Muchoborska 18
54-424 Wrocław
Poland
jan.chorowski@techtra.pl

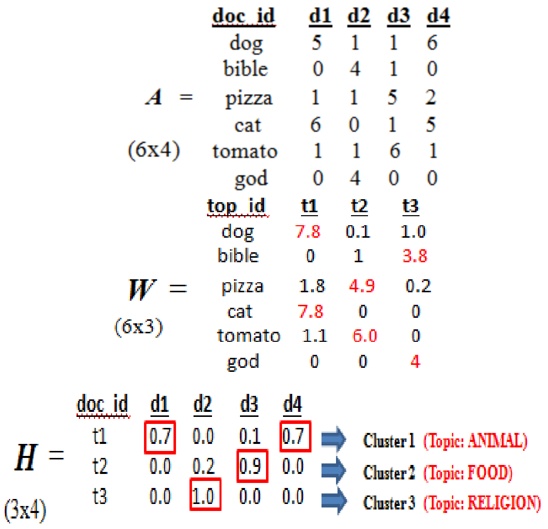
***Abstract*—Nonnegative Matrix Factorization (NMF) is one of the most promising techniques to reduce the dimensionality of the data. This presentation compares the method with other popular matrix decomposition approaches for various pattern analysis tasks. Among others, NMF has been also widely applied for clustering and latent feature extraction. Several types of the objective functions have been used for NMF in the literature. Instead of minimizing the common Euclidean Distance (EucD) error, we review an alternative method that maximizes the correntropy similarity measure to produce the factorization. Correntropy is an entropy-based criterion defined as a nonlinear similarity measure. Following the discussion of maximization of the correntropy function, we use it to cluster document data set and compare the clustering performance with the EucD-based NMF. Our approach was applied and illustrated for the clustering of documents in the 20-Newsgroups data set. The comparison is illustrated with 20-Newsgroups data set. The results show that our approach produces per average better clustering compared with other methods which use EucD as an objective function.**

***Keywords*—Nonnegative Matrix Factorization; Correntropy; Principal Component Analyis; Face recognition**

I. Introduction

The ever-increasing amount of data recorded, stored and processed worldwide necessitates the development of new representations and is becoming a major task for data analysis research [1, 2, 3, 4, 17, and 18]. Dimensionality reduction of the data is a technique that describes each multidimensional data sample with a small number of coefficients that are the sample's coordinates in a new, particular to this dataset, feature space. Often dimensionality reduction is accomplished by finding factorizations of a matrix representing the dataset. Most widely-known methods are Principal Component Analysis (PCA), Independent Component Analysis (ICA) and Singular Value Decomposition (SVD). Recently defined Nonnegative Matrix Factorization (NMF) approach also has been successfully applied in pattern recognition. It is an unsupervised learning method that also reduces the dimensionality of the data. It has also been used for several applications [5-8, 14-16, 19, 20].

Matrix factorization methods treat the data as an $m \times n$ matrix in which every column represents a data sample.



A =
(6x4)

doc_id	d1	d2	d3	d4
dog	5	1	1	6
bible	0	4	1	0
pizza	1	1	5	2
cat	6	0	1	5
tomato	1	1	6	1
god	0	4	0	0

W =
(6x3)

top_id	t1	t2	t3
dog	7.8	0.1	1.0
bible	0	1	3.8
pizza	1.8	4.9	0.2
cat	7.8	0	0
tomato	1.1	6.0	0
god	0	0	4

H =
(3x4)

doc id	d1	d2	d3	d4	
t1	0.7	0.0	0.1	0.7	➡ Cluster1 (Topic: ANIMAL)
t2	0.0	0.2	0.9	0.0	➡ Cluster2 (Topic: FOOD)
t3	0.0	1.0	0.0	0.0	➡ Cluster3 (Topic: RELIGION)

Figure 1. A small document-term matrix and its nonnegative factorization. Indicated are terms that are most important for each topic and assignments of documents into topics (clusters).

This matrix is approximated by a product of two rank k matrices, as follows:

$$A \approx WH,$$

where A is the data matrix, W is the $m \times k$ matrix of basis vectors and H is a $k \times n$ matrix that gives the coordinates of samples in the feature space. We can think of the factorization as of a decomposition of the j-th sample (j-th column of A, $A_{:j}$) into a linear combination of features given by columns of W:

$$A_{:j} = \sum_i W_{:i} H_{ij}.$$

For instance, consider a small artificial data set of documents shown in Figure 1. The documents are encoded in the bag-of-words format. Here, A is a 6×4 data matrix formed by 4 documents, 6 words of interest, and 3 topics. The topics are ANIMAL (d1 and d4), RELIGION (d2) an d FOOD (d4). Once the factorization is computed we can cluster the documents [14]. It can be accomplished by assigning each document to the topic that contributes the

978-1-4673-4471-5/$25.00 © 2013, IEEE

most to its nonnegative representation (has the largest entry in the matrix H). The matrices W and H that form a nonnegative factorization are shown in Figure 1. We have indicated the most important words for each topic and demonstrated the cluster memberships.

II. Nonnegative Matrix Factorization (NMF)

The intuitive definition of matrix factorization asks to find two matrices, W and H, whose product approximates a given matrix. Specific matrix factorization schemes are differentiated by the error function used to describe the quality of the approximation and by the constraints imposed on the elements of W and H. The family of nonnegative factorizations imposes that elements of W and H be nonnegative. A necessary prerequisite is that the data matrix A must also contain nonnegative elements only. Fortunately, this is often the case. In example documents in a bag-of-words format or images are non-negative. Application areas for NMF include face recognition, bioinformatics, text mining and audio (speech) processing [2, 3, 6-8, 14, 19]. Clustering task is also one of the main topics for NMF and it has been extensively applied and discussed in the literature [6, 8, 14-16].

Nonnegative factorizations are motivated by their enhanced interpretability. When subtraction is forbidden no cancelations occur in the topic (cluster) definitions and meanings can be deduced, as in the example shown in Figure 1. Indeed, for many applications subtraction is not meaningful. However, except for NMF, other matrix factorization methods generally allow the subtraction of values. These values can be faces, audio or gene expression levels according to application areas. But in these cases, basis values (for instance images for face recognition) are not physically intuitive.

Typically, in NMF the factorization objective is the Euclidean distance between the elements of A and the elements of WH (i.e. the Frobenius norm of the difference $A - WH$). This measure is well-known and often used in the literature. However, other distance (or similarity) measures can be used, and they will often produce different factorizations. In example, calculations derived from the Kullback-Leibler divergence have often been studied in the literature [1-4, 9, and 14]. Often the loss function is chosen to match a specific application domain. In [7], authors used the Itakura-Saito (IS) divergence as an objective function and in [5] authors used the β-divergence.

In [1-4], authors used a distance measure based on the Kullback-Leibler divergence. The measure $D_s(u_j, w_i)$ is a symmetric divergence of $u_j$with respect w_i given by [2]:

$$D_s(u_j, w_i) = D(u_j||w_i) + D(u_j||w_i),$$

where:

$$D(x||z) = \sum_l \frac{x(l)}{\|x\|_1} \log\left(\frac{x(l)\|z\|_1}{z(l)\|x\|_1}\right).$$

Another distance measure suitable for the NMF is the correntropy function, described in details in Section III.

Extensions of the NMF methodology involve imposing other constraints on the matrices W and H , such as sparseness or orthogonality. Bayesian approaches and other conditions for factorization have also been considered [3, 8].

The typical algorithm used to compute the NMF factorization with the Euclidean distance measure begins with W and H randomly initialized. It then uses the multiplicative update rules to minimize the error function [1,4]:

$$\mathrm{H_{ij}} \leftarrow \mathrm{H_{ij}} \frac{(\mathrm{W^T A})_{\mathrm{ij}}}{(\mathrm{W^T W H})_{\mathrm{ij}}}$$

$$\mathrm{W_{ij}} \leftarrow \mathrm{W_{ij}} \frac{(\mathrm{A H^T})_{\mathrm{ij}}}{(\mathrm{W H H^T})_{\mathrm{ij}}}$$

The rules ensure that at each iteration the error function does not increase, while the matrices W and H stay non-negative. The rules are applied iteratively until convergence.

Faster converging alternatives to the multiplicative updates, that have been proposed for the NMF include the projected gradient descent (PGD) and the alternating least squares (ALS) algorithm [2, 16].

III. Correntropy Similarity Measure

We have recently proposed to use the correntropy similarity measure as an objective function for nonnegative matrix factorization [26, 27]. The correntropy is a localized similarity measure between two random variables that was proposed in [9-12, 14]. It can be used as a cost function for NMF. We use it to calculate the element-wise similarity between the matrix A and its factorization:

$$Corr(A, WH) = \sum_{\mathrm{i,j}} \exp\left(\frac{-\left(A_{ij} - (WH)_{ij}\right)^2}{2\sigma^2}\right) \quad (1)$$

where σ is a parameter of the correntropy similarity measure. We note that for NMF we need to minimize the negative of correntropy since it is a similarity and not a distance measure [14].

It can easily be seen from eq. 1 that $Corr(A, WH)$ is always bounded and nonnegative. Moreover, the correntropy saturates when the disagreement between elements of A and its factorization WH is large. This property is important. It makes correntropy insensitive to outliers, because errors for badly approximated elements have less influence on the factorization. We illustrate correntropy as the error surface in Figure 2. It shows the errors for a single element of $1 + Loss(A, WH)$. We can change the shape of the function and control the level of saturation by adjusting the parameter σ. When σ is large little saturation occurs. Lowering σ causes that more and more elements of the difference $A - WH$ saturate and are treated as outliers.

IV. Exemplary Applications of NMF

A. Document Clustering with NMF

For the first real life example we report the result of a comparison between quality of NMF factorizations based on the Euclidean distance and based on correntropy [14, 26]. The evaluation analyses the quality of clusters computed from factorizations. We have used the 20-newsgroups data set, which is one of the popular benchmarks used for clustering and classification of the text data. It has approximately 11,000 documents taken from 20 different newsgroups pertaining to various subjects.

After the factorization process, we obtain W and H. H can be used to group the data (A) into r clusters by choosing the largest value of each column in H.

The 20 newsgroups data contains ground-truth document labels which can be used to evaluate the quality of the clustering. We evaluate the clustering performance with the entropy measure. Total entropy for a set of clusters is calculated as the weighted mean of the entropies of each cluster weighted by the size of each cluster. Firstly, we calculate the distribution of the data for each cluster. For class j we compute p_{ij}, the probability that a member of cluster i belongs to class j as $p_{ij} = m_{ij}/m_i$, where m_i is the number of objects in cluster i and m_{ij} is the number of objects of class j in cluster i. Entropy of each cluster i is defined as:

$$e_i = -\sum_{j=1}^{L} p_{ij} log_2(p_{ij}),$$

where L is the number of classes. Entropy of the full data set as the sum of the entropies of each cluster i weighted by the size of each cluster:

$$e = \sum_{i=1}^{K} \frac{m_i}{m} e_i$$

where K is the number of clusters and m is the total number of data points [24].

Table 1 shows the entropy values of *NMF-PGD (EucD)* and *NMF-Corr* approaches for 20-Newsgroups data set. We graph these values (*NMF-PGD (EucD)* and *NMF-Corr (for* $\sigma = 1$, $\sigma = 0.5$ and $\sigma = 0.01$)) in Figure 3. Here, "k" denotes the assumed number of clusters and equals to the ranks of W, H. We change it from 2 to 20 to track the clustering performance. We show all entropy values in Figure 3, but for brevity we only illustrate 10 data points in Table 1. Since lower entropy values indicate better clustering performance, it can be seen from Table 1 and Figure 3, that *NMF-Corr* ($\sigma = 0.5$) demonstrates superior clustering performance than *NMF-PGD (EucD)* for every evaluated number of clusters.

Experiments and comparative results between NMF-PGD (EucD) and NMF-Corr show that NMF-Corr ($\sigma = 0.5$) has better clustering performance than NMF-PGD (EucD). Therefore, we can conclude that correntropy-based NMF ($for\ \sigma = 0.5$) has comparatively better clustering performance vs. EucD-based NMF for the evaluated data set. However, NMF-Corr does not show improved performance for $\sigma = 1$ and specifically worst performance for $\sigma = 0.01$. This can be seen from Figure 3 and Table 1. Also, the deterioration of clustering results for σ values below 0.5 requires further studies. One additional question

Table 1. Entropy of 20-Newsgroups data set with NMF-PGD (EucD) and NMF-Corr.

Number of Clusters (k)	NMF-PGD (EucD)	NMF-Corr *(σ = 1)*	**NMF-Corr *(σ = 0.5)***	NMF-Corr *(σ = 0.01)*
r = 2	3.84	3.86	**3.85**	4.30
r = 3	3.86	3.79	**3.58**	4.27
r = 4	3.78	3.49	**3.50**	4.27
r = 5	3.74	3.60	**3.38**	4.24
r = 6	3.49	3.36	**3.30**	4.23
r = 7	3.44	3.28	**3.26**	4.20
r = 8	3.30	3.26	**2.94**	4.19
r = 9	3.30	3.34	**3.13**	4.18
r = 10	3.16	3.23	**2.93**	4.20

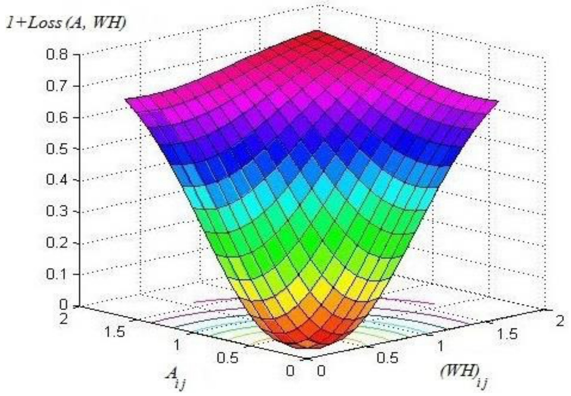


Figure 2. Correntropy objective function 1+Loss(A,WH) with σ=0.5,m=n=1.

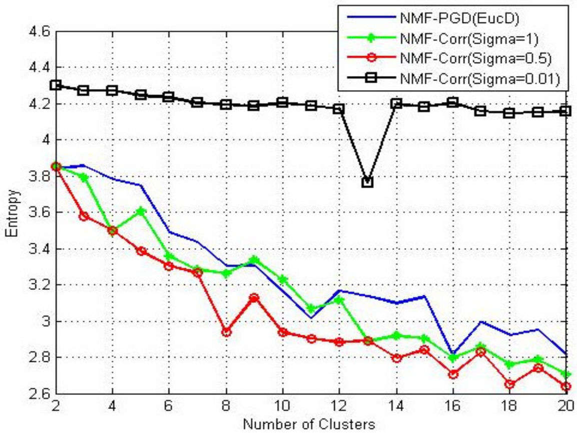


Figure 3. Entropy comparison for NMF-PGD (EucD) and NMF-Corr.

is whether this dependence on σ value is a property of the method or else whether it lies in the properties of the data for which experiments have been conducted. This will warrant further studies.

B. Occluded Face Recognition Using NMF

In the second example we report the results of an application of NMF to the problem of occluded face recognition [26].

Face recognition is one of the well-studied real life problems. Several methods have been defined and applied for this task. Above mentioned methods and Neural Networks (NN) have been studied to recognize face images [1, 19-26]. In fact, faces are not clear for daily life, because some obstacles can be in front of the face. These obstacles can be scarf, glasses, hats or some occlusion on the face. Therefore, occluded face recognition is important area in pattern analysis. There are many studies in the literature for occluded face recognition task, especially using PCA and NMF [19-26].

In this section, we evaluate the recognition performance of occluded face images on ORL face data set. We have compared PCA, NMF and correntropy based NMF (NMF-Corr) formulations by evaluating quality of recognition rates computed from factorizations. The ORL data consists of 40 persons, each photographed in 10 different poses. The data set was partitioned into two equal parts for training and testing. We have resized face images from original 112x92 pixels to 56x46 pixels for efficient computation.

Face recognition in the NMF and NMF-Corr linear subspace is performed by first computing the pseudo-inverse of the W matrix as $W^{+} = W^{T}(WW^{T})^{-1}$. Then, all samples were encoded using this pseudo inverse. Finally, we have used 1 nearest-neighbor (1-NN) classifier for the recognition process.

In order to generate occluded faces, we have used randomly located black patches for both training and testing face images. In this way we test the robustness of the compared dimensionality reduction methods to noise on both training and testing data. Each patch covers from 10% to %50 of the face image at a random location. Sample patched face images can be seen from Figure 4.

Recognition results have been obtained by running each method (PCA, NMF and NMF-Corr) 10 times, and then average recognition rate has been calculated.

NMF-Corr and NMF algorithms were run with the random initial matrices W and H. For NMF-Corr, we set stopping criteria at most 1000 iterations and relative tolerance 10^{-4}. PCA, NMF and NMF-Corr basis images has been shown in Figure 5, respectively.

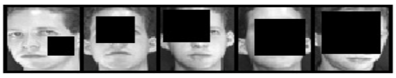
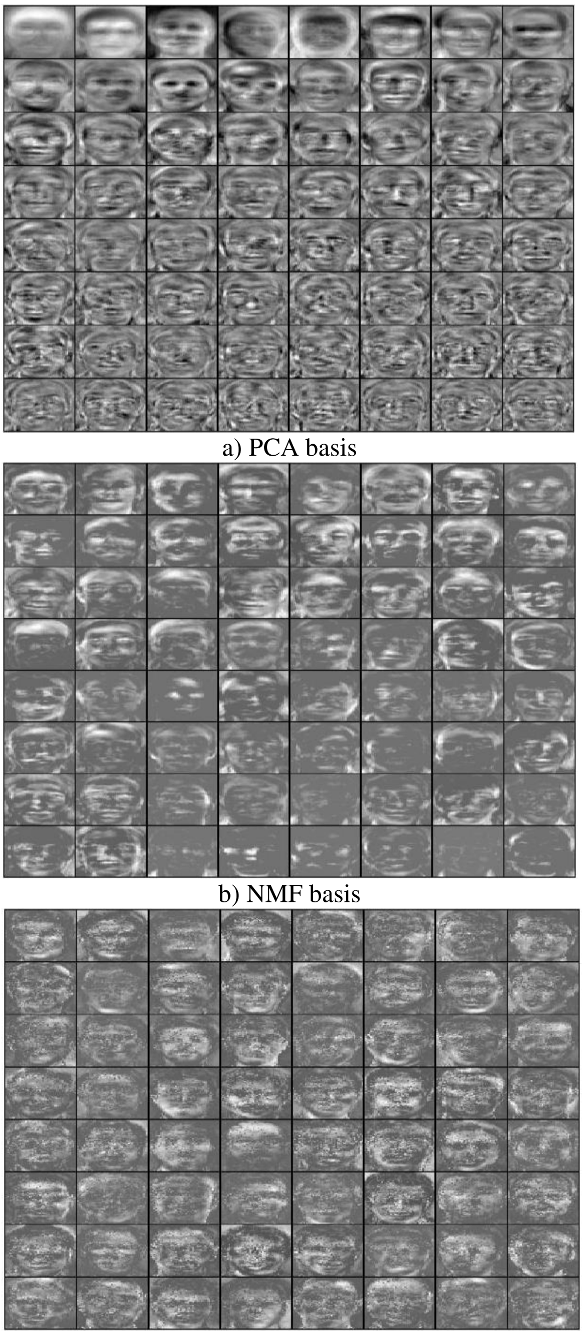


Figure 4 Randomly located occluded face samples from ORL face dataset with 10%, 20%, 30%, 40% and 50% patch sizes (From left to right).



a) PCA basis

b) NMF basis

c) NMF-Corr basis

Figure 5. Basis images of PCA, NMF and NMF-Corr for 64 grids.

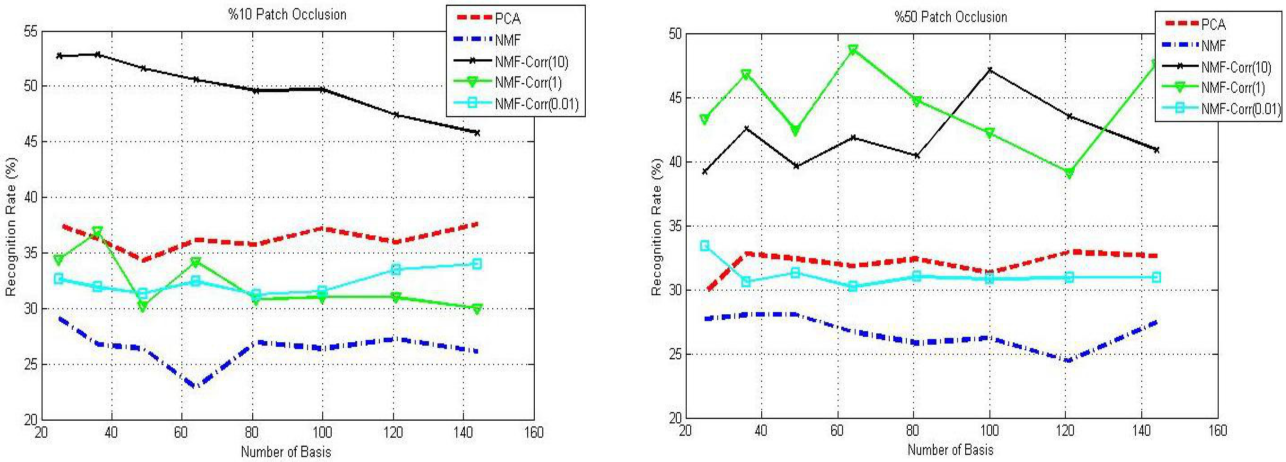


Figure 6. Recognition rates (%) versus number of basis images for 10% and 50% patch occlusions (On the legend, values in paranthesis indicate the corresponding σ parameter for NMF-Corr).

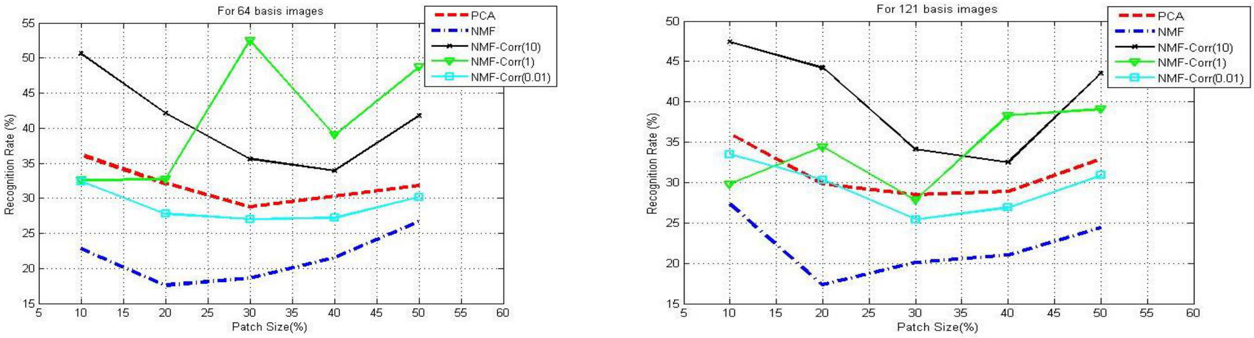


Figure 7. Recognition rates (%) versus patch sizes (%) of face images for 64 and 121 basis images.

For brevity we only illustrate for 10% and 50% occlusions in Figure 6, for different number of basis images. It can be easily seen that NMF-Corr with $\sigma = 10$ demonstrates superior recognition performance than NMF and PCA. Therefore, NMF-Corr with $\sigma = 10$ has the best accuracy. (In the case of 50% patch occlusion, generally $\sigma = 1$ has better accuracy than $\sigma = 10$). Recognition rate plots versus patch occlusion sizes have been also calculated for 25, 36, 49, 64, 81, 100, 121 and 144. Again, we only demonstrate for 64 and 121 basis in Figure 7 for brevity. Here, NMF-Corr has the best recognition rate for all patch sizes. Additionally, it can be seen from Figure 7, the graphic lines are u-shaped, because training and testing parts have been done with occluded face images.

V. Conclusion

In this contribution we have first introduced the topic of nonnegative matrix factorization and reviewed its major applications and implementations. The NMF factorizes a given data matrix into a product of two matrices that contain nonnegative elements only. Subtraction is forbidden which enhances sparsity of the patterns that are found in the data. This leads to a better interpretability of the factorization.

The usefulness of nonnegative factorizations was demonstrated using two real-life tasks: document clustering and occluded face recognition. Moreover the demonstrations used correntropy, a novel similarity measure that enhances the robustness to outliers. Experiments on both datasets have shown that using the correntropy criterion has led to better cluster purity and recognition rates than NMF and PCA.

Acknowledgments

This paper is an extended version of [14] and [26].

References

[1] Lee D., Seung H. S. "Learning the Parts of Objects with Nonnegative Matrix Factorization", Nature, Vol. 401, pp. 788-791, 1999.

[2] Berry, M. W., Browne M., Langville A. N., Pauca V. P., Plemmons R. J. , "Algorithms and Applications for Approximate Nonnegative Matrix Factorization", Computational Statistics and Data Analysis, Vol. 52, No. 1, pp. 155–173, 2007.

[3] Hoyer P. O., "Non-negative Matrix Factorization with Sparseness Constraints", Journal of Machine Learning Research 5, pp. 1457-1469, 2004.

[4] Lee D., Seung H. S., "Algorithms for Non-negative Matrix Factorization", Advances in Neural Information Processing, Vol. 13, pp. 556-562, 2001.

[5] Fevotte C. and Idier J. "Algorithms for Nonnegative Matrix Factorization with the β- Divergence", Neural Computation, Vol. 13, Issue 3, pp. 1-24, 2010.

[6] Zhao W., Ma H., Li N., "A Nonnegative Matrix Factorization Algorithm with Sparseness Constraints", Int. Conf. on Machine Learning and Cybernetics, Guilin, China, July 10-13, 2011.

[7] Fevotte C., Bertin N., Durrieu J. L., "Nonnegative Matrix Factorization with the Itakura-Saito Divergence, Neural Computation, Vol. 21, pp. 793-830, 2009.

[8] Shahnaz F., Berry M. W., Pauca V. P., Plemmons R. J. "Document Clustering Using Nonnegative Matrix Factorization", Int. Journal of Information Processing and Management, Vol. 42, Issue 2, pp. 373-386, 2006.

[9] He R., Zheng W. S., Hu B. G., "Maximum Correntropy Criterion for Robust Face Recognition", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 33, No. 8, pp. 1561-1576, 2011.

[10] Liu W., Pokharel P. P., Principe J. C.: Correntropy, "Properties and Applications in Non-Gaussian Signal Processing", IEEE Transactions on Signal Processing, Vol. 55, No. 11, pp. 5286-5298, 2007.

[11] He R., Hu B. G., Zheng W. S., Kong X. W., "Robust Principal Component Analysis Based on Maximum Correntropy Criterion", IEEE Transactions on Image Processing, Vol. 20, No. 6, 2011.

[12] Chalasani R., Principe J. H., "Self Organizing Maps with Correntropy Induced Metric", Int. Joint Conf. on Neural Networks, Spain, pp. 1-6, 2010.

[13] Matlab Software, www.di.ens.fr/~mschmidt/Software/minConf.html

[14] Ensari T, Chorowski J, Zurada J. M., "Correntropy-based Document Clustering via Nonnegative Matrix Factorization", Int. Conf. on Artificial Neural Networks (ICANN), Lausanne, Switzerland, September 11-14, 2012.

[15] Zhao W., Ma H., Li N., "A New Non-negative Matrix Factorization Algorithm with Sparseness Constraints, Proc. of the 2011 Int. Conf. on Machine Learning and Cybernetics, Guilin, 10-13 July, 2011.

[16] Lin C. J., "Projected Gradient methods for Non-Negative Matrix Factorization", Neural Computation, 19:2756-2779, 2007.

[17] Tan P., Steinbach M., Kumar V., "Introduction to Data Mining", Pearson Addison Wesley, 2006.

[18] P. Paatero, "Least Squares Formulation of Robust Non-negative Factor Analysis", Chemometrics and Intelligent Laboratory Systems 37, 23-35, 1997.

[19] Wang Y., Jia Y., "Non-Negative Matrix Factorization Frame for Face Recognition", Int. Journal of Pattern Recognition and Artificial Intelligence, Vol. 19. No.4, pp. 495-511, 2005.

[20] Byeon W., Jeon M., "Face Recognition Using Region-based Nonnegative Matrix Factorization", Communications in Computer and Information Science, Vol. 56, pp. 621-628, 2009.

[21] Feng T., Li S. Z., Shum H. Y., Zhang H. J., "Local Non-negative Matrix Factorization as a Visual Perpeption", Int. conf. on Development and Learning, June 12-15, 2002.

[22] Shastri B. J. and Levine M. D., "Face Recognition Using Localized Features Based on Non-Negative Sparse Coding", Machine Vision and Applications, Vol. 18, No. 2, pp. 107-122, 2007.

[23] Oh H. J., Lee K. M., Lee S. U., "Occlusion Invariant Face Recognition Using Selective Local Non-negative Matrix Factorization Basis Images", Image and Vision Computing, Vol. 26, Issue 11, pp. 1515-1523, November 2008.

[24] Pan J. Y., Zhang J. S., "Large Margin Based Nonnegative Matrix Factorization and Partial Least Squares Regression for Face Recognition", Pattern Recognition Letters, Vol. 32, pp. 1822-1835, 2011.

[25] Liu H., Wu Z., Li X., Cai D., Huang T. S., "Constrained Nonnegative Matrix Factorization for Image Representation", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 34, No. 7, July 2012.

[26] Ensari T., Chorowski J., Zurada J. M., "Occluded Face Recognition Using Correntropy-based Nonnegative Matrix Factorization", International Conference on Machine Learning and Applications (ICMLA), Boca Raton, Florida, USA, December 12-15, 2012.