SPARSE LOCALIZED FACIAL MOTION DICTIONARY LEARNING FOR FACIAL EXPRESSION RECOGNITION

Chan-Su Lee

Yeungnam University
Dept. of Electronic Engineering
Gyeongsan, Korea

Rama Chellappa

University of Maryland at College Park
Dept. of Electrical and Computer Engineering
College Park, MD, U.S.A.

ABSTRACT

This paper presents a new framework for facial motion modeling with applications to facial expression recognition. First, we design sparse localized facial motion dictionaries from dense motion flow data of facial expression image sequences. Regularization based on spatial localized support map in addition to the sparsity constraints enables spatially localized dictionary learning. Proposed localized dictionaries are effective for local facial motion description as well as global facial motion analysis. Experimental results using CK+ database shows promising results for automatic facial expression recognition from motion flow data.

Index Terms— Facial expression recognition; dictionary learning; motion analysis

1. INTRODUCTION

Facial expression is important for human communication and understanding the emotional state of a subject. Driver assistance systems analyze human facial expressions and estimate the level of drowsiness of a driver to provide a warning if it is necessary. Humanoid robots and service robots that interact with humans can be benefit by understanding human emotional state. In this paper, a localized dictionary model of facial motion flow, which provides localized sparse motion components for the analysis of facial expressions, is presented.

Many facial models have been developed based on 2D and 3D landmarks [1], facial appearances [2], geometry [3], and 2D/3D spatio-temporal representations [4]. Many questions remain unanswered. What is the best model for facial expression analysis? How do human process and categorize facial expression signals in the brain? Neurophysiological studies show that there are distinct neural substrates dedicated to facial expressions as well as faces [5]. Psychological studies also shows that the human brain effectively minimizes correlation by low overlapping local regions in facial expression recognition tasks [6].

Recently, region- or part-based approaches have been investigated in computer vision and graphics communities for object and human detection [7], action recognition [8, 9], video analysis [10, 11], and facial motion analysis [12, 13, 14], and synthesis [15, 16]. In the case of facial expression analysis, region- or part-based approaches have shown improvements in facial expression recognition [13, 14] using manual annotation of region corresponding to local facial motion activations in different expressions. In addition, region-based approaches are also supported by psycho-visual experiments that track eye-movements. For example, salient facial region-based hierarchical features have shown better performance in facial expression recognition than conventional ones [17].

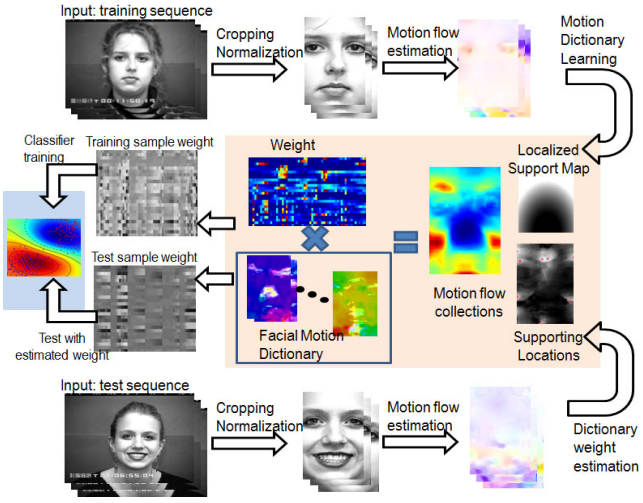


Fig. 1. A flow diagram of the proposed system

To overcome manual, predefined region-based approaches, facial patch-based learning methods have been suggested for automatic extraction of important facial local region [18]. However, these approaches are limited by the interactions between predefined regular grid patches or regular patch masks [19].

Nonnegative matrix factorization (NMF) [20] and its sparse constrained extensions [21] are well known approaches for part-based object representation. They have been applied

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

for part-based activity recognition [9], or facial action unit intensity estimation [22], and expression recognition [23]. However, it is difficult to identify localized motion components based on NMF and its extensions.

The main contribution of this paper is a new framework to learn spatially localized dictionary in image space for facial expression analysis. The localized dictionary is useful not only for facial expression analysis but also for facial action unit analysis. In addition, the proposed dictionary is based motion flow, which is robust to changes in facial skin color or illumination.

2. SYSTEM OVERVIEW

The proposed system processes video sequences and recognizes facial expressions. The system has three stages: 1) localized motion flow dictionary learning, 2) dictionary-based feature extraction for training data and classifier training if necessary, 3) dictionary-based feature extraction for test data and facial expression recognition. To process localized facial motion from a video sequence, global head translation, in-plane head rotation, scaling effects are eliminated based on face normalization from two eye locations as explained in Sec. 4.1. Optical flow from the first frame of each sequence is estimated (Sec. 3.1) and used for dictionary learning, training, and testing of facial expressions.

In the motion flow dictionary learning stage, we collect motion flow from training sequence and iteratively learn motion flow dictionary elements. After learning dictionaries of several different sizes, the proper dictionary size is decided based on reconstruction error and sparsity of the learned dictionary. For training and testing, the weight values of learned dictionary components are used as feature vectors. In the testing stage, the expression labels are estimated for each frame and the majority voting is used to select the class labels. Fig. 1 shows the overall system diagram of the proposed facial expression recognition system.

3. FRAMEWORKS: SPARSE LOCALIZED FACIAL MOTION DICTIONARY LEARNING

3.1. Motion flow estimation

Let $\tilde{I}(x, y, t)$ be a original video sequence, $I(x, y, t)$ be a normalized facial motion video sequence, and $(x(t), y(t))$ be the trajectory of a point in the image plane, then the brightness constancy assumption states that $I(x(t), y(t), t)$ is constant. Thus,

$$\frac{d}{dt} I(x(t), y(t), t) = 0. \tag{1}$$

To solve the under-determined linear system, the sum of total variation $\boldsymbol{\mu}$ and L_1 regularization terms are proposed and solved by convex optimization [24]. As a result, the collection of $\boldsymbol{\mu}(x, y, t) = (\mu_1(x, y), \mu_2(x, y), t)$ represents motion flow of facial expression from video sequences. This motion flow sequence is invariant to appearance and skin color variations of subjects.

3.2. Dictionary learning by spatial regularization and sparsity constraints

When normalized flow size is $S = w \times h$, and the frame number is N, the flow data can be represented by a tensor $S \times M \times N$, where $M = 2$ is motion flow dimension in our 2D image sequence (X axis motion component, and Y axis motion component). This data set can be represented by a matrix $\boldsymbol{X}$, whose size is $(P \times N)$, using tensor unfolding [25], where $P = S \times M$ is a feature dimension.

The proposed method is aimed at decomposing captured motion flow sequence into sparse, localized motion dictionary components. Dictionary learning [26] for data matrix $\boldsymbol{X}$ or its i'th $P \times 1$ column vector $\boldsymbol{x}_i$ discovers sparse representation of the data by

$$\min_{\boldsymbol{D}, \boldsymbol{\alpha}} \sum_{i=1}^{N} \|\boldsymbol{x}_i - \boldsymbol{D}\boldsymbol{\alpha}_i\|_F^2 + \lambda \|\boldsymbol{\alpha}_i\|_1, \tag{2}$$

where $\boldsymbol{D}$ in $\mathbb{R}^{P \times K}$ is a dictionary with K elements, $\boldsymbol{\alpha}_i$ in $\mathbb{R}^K$ is a sparse representation of i'th data using dictionary $\boldsymbol{D}$, and λ is non-negative. The dictionary learning scheme can be extended by adding additional constraints [27]. The matrix factorization step can be formulated as a joint regularized minimization problem. The constraints on the weight $\boldsymbol{\alpha}_i$ are essential to prevent the weight from getting too large [16] and motion flow components getting arbitrary small.

To find an appropriate regularization for localized motion components, two factors have to be counted. First, the k'th column of $\boldsymbol{D}$ forms a two spatial coordinate:

$$\boldsymbol{d}_k^{(j)} = [\mu_j(x, y)]_k, j = 1, or\ 2 \tag{3}$$

Each j corresponds to the X axis or Y axis motion components in the dictionary component k. Therefore, it is required to consider this inherent group structure. The l_1/l_2 norm can be used for group sparsity. Second, to derive a motion flow basis where facial motion occurs locally, we enforce each motion flow dictionary to be centered around a set of local areas. The spatial locality needs to be forced during dictionary learning procedure. Inspired by [16], which proposed localized deformation components of 3D mesh data for facial animation, spatially-varying regularization parameters are enforced into the constraints. The final objective function is derived as follows:

$$\min_{\boldsymbol{D}, \boldsymbol{\alpha}} \quad \sum_{i=1}^{N} \quad \|\boldsymbol{x}_i - \boldsymbol{D}\boldsymbol{\alpha}_i\|_F^2 + \sum_{k=1}^{K} \sum_{j=1}^{M} \Lambda_k \|\boldsymbol{d}_k^{(j)}\|_2, \tag{4}$$

$$s.t. \quad \max(\boldsymbol{\alpha}_i) = 1, \boldsymbol{\alpha}_i \geq 0 \tag{5}$$

We define the range of local support from the center of each motion flow component by range $[r_{min}, r_{max}]$. The spatial support region within the pre-defined range maps into $[0, 1]$ range linearly to the regularization strength.

3.3. Optimization

It is difficult to directly optimize Eq. 4 due to non-convexity. However, an iterative solution can be achieved since if one of the components, either $\boldsymbol{D}$ or $\boldsymbol{\alpha}$ is known, the problem becomes convex.

Initialization: Initial dictionary component $\boldsymbol{d}_k$, where $k = 1, 2, \cdots, K$, can be extracted by finding components which explain the maximum variation in the data. Spatial location with the maximum variance is found and modeled after applying a local support map to explain the variation in the localized dictionary. The next component can be found by removing variance explained by current selected components. The number of component K is predefined based on reconstruction error and sparsity.

Optimization of weights: Given a dictionary $\boldsymbol{D}$, a collection of dictionary components, the weight vector $\boldsymbol{\alpha}$ can be solved by constrained linear least squares problem. The block-coordinate descent algorithm [28] can be used for optimization of weights for each component.

Optimization of local motion flow components: The optimization of $\boldsymbol{D}$ can be solved using convex optimization for a given weight matrix $\boldsymbol{A}$ in $\mathbb{R}^{K \times N}$, which is a collection of weight vectors, $\boldsymbol{\alpha}$s. To optimize the l_1/l_2 norm regularizer, the Alternative Direction Method of Multipliers (ADMM) [29] is applied. By introducing a dual variable $\boldsymbol{Z} \in \mathbb{R}^{P \times K}$, the optimization objective can be rewritten in a form compatible with ADMM as follows:

$$\begin{aligned} \arg\min_{\boldsymbol{D},\boldsymbol{Z}} \quad & \|\boldsymbol{X} - \boldsymbol{D} \cdot \boldsymbol{A}\|_F^2 + \Omega(\boldsymbol{Z}) \\ s.t. \quad & \boldsymbol{D} - \boldsymbol{Z} = 0, \Omega(\boldsymbol{Z}) = \sum_{k=1}^{K} \sum_{j=1}^{M} \Lambda_k \|\boldsymbol{z}_k^{(j)}\|_2. \end{aligned} \tag{6}$$

The ADMM algorithm initializes $\boldsymbol{U} \in \mathbb{R}^{P \times K}$ to zero and then iterates the following steps. The ADMM algorithm iteratively update $\boldsymbol{D}, \boldsymbol{Z}, \boldsymbol{U}$ sequentially [16].

$$\begin{aligned} \boldsymbol{D} &\leftarrow \arg\min_{D} \|\boldsymbol{X} - \boldsymbol{D} \cdot \boldsymbol{A}\|_F^2 + \frac{\rho}{2}\|\boldsymbol{D} - \boldsymbol{Z} + \boldsymbol{U}\|_F^2 \\ \boldsymbol{Z} &\leftarrow \arg\min_{\boldsymbol{Z}} \left(\Omega(\boldsymbol{Z}) + \frac{\rho}{2}\|\boldsymbol{D} - \boldsymbol{Z} + \boldsymbol{U}\|_F^2\right) \\ \boldsymbol{U} &\leftarrow \boldsymbol{U} + \boldsymbol{D} - \boldsymbol{Z} \end{aligned}$$

4. EXPERIMENTAL RESULTS

4.1. Face normalization and flow estimation

From the landmark points provided in the database, we compute left and right eye centers, scale and align them so that the eye centers are at a fixed location. In our experiments, from the first frame, the eye location is estimated and applied to the given sequence with the same global transformation. In real situations, a face detector at the initial frame can provide the required information for face normalization. If large head motion is present, repeated normalization based on eye location will help remove the global transformation and extract the local motion from the optical flow.

4.2. Sparse motion dictionary learning and its interpretation

From the collection of motion flow data, we learn the sparse motion dictionary as presented in Sec. 3. During the dictionary learning stage, we have to decide dictionary size K. The dictionary size is determined by looking at the reconstruction error and sparsity of the given dictionary size. Fig. 2 (a), and (c) show the reconstruction error and sparsity convergence as a function of iteration. However, if we compare the final reconstruction error, it is not always the case that large dictionaries yield lower reconstruction errors due to sparsity constraints. In the case of Fig. 2 (b), we see that the reconstruction error is minimized when the dictionary size K=40, which dictionary size is used for the evaluation of the training and test dataset.

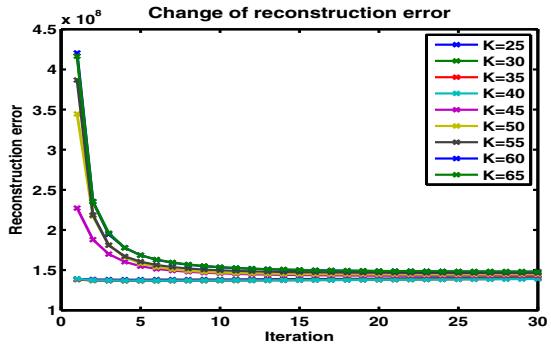
The learned dictionary shows sparse and localized facial motion primitives. Fig. 3 shows flow intensities of learned dictionary components. Regions with bright intensity correspond to area of strong motion flow. The figure shows localized brightness spot in each dictionary element.

Table 1. Facial expression recognition accuracy (CK+).

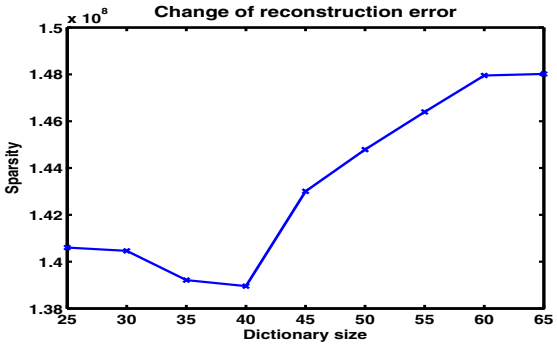
NNLS [30] (texture)	NNLS [30] (motion)	Proposed sparse dictionary (motion), 1-NN
59.02	79.39	**86.7**

4.3. Facial expression recognition

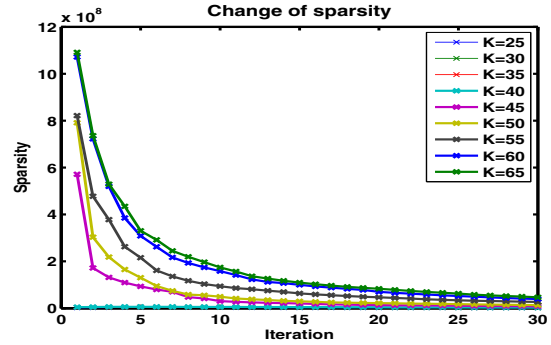
We evaluated the performance of facial expression recognition using the CK+ facial expression database [31], which is one of the well-known facial expression database. The database provides not only expression type but also the action units and facial landmark tracking result. To focus on the proposed sparse local facial motion dictionary performance, we used a very primitive 1-NN(nearest neighborhood) classifier and the majority voting scheme for expression recognition. For the comparison of the performance we created fourfold sets. We evaluated the performance of nonnegative least square classifier(NNLS) [30], which is based on approximation by sparse non-negative linear combinations optimized for classification. By using motion features, the performance is improved by 20% in (NNLS) [30] classifier. The proposed



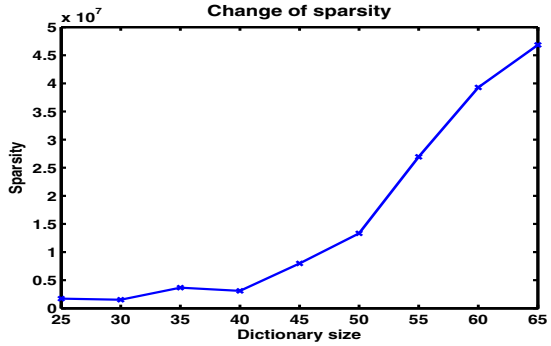
(a) Reconstruction error (by iteration)



(b) Reconstruction error (by dictionary size)



(c) Sparsity (by iteration)



(d) Sparsity (by dictionary size)

Fig. 2. Reconstruction errors and sparsity during iteration in different dictionary size

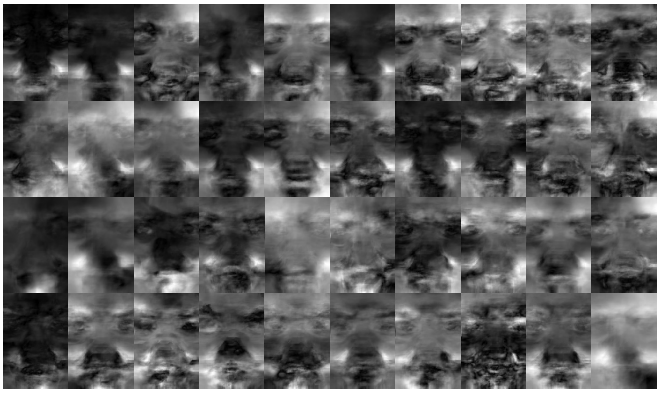


Fig. 3. Learned motion flow dictionary elements(K=40).

motion dictionary based approaches outperform classification optimized nonnegative matrix factorization by 7%. Table 1 shows the recognition accuracies. The confusion matrix corresponding to four-fold cross validation test is given in Table 2.

Table 2. Confusion Matrix: proposed (4 fold) (CK+).

	An	Co	Di	Fe	Ha	Sa	Su
An	**84.8**	2.2	6.5	2.2	6.5	4.35	0.0
Co	4.5	**81.8**	0.0	0.0	9.1	0.0	4.6
Di	8.8	0.0	**89.5**	0.0	1.75	0.0	0.0
Fe	14.8	0.0	7.4	**63.0**	3.7	7.4	3.7
Ha	1.5	0.0	4.5	1.5	**91.0**	0.0	1.5
Sa	3.3	3.3	3.3	3.3	3.3	**80.0**	3.3
Su	1.2	1.2	0.0	1.2	2.5	0.0	**93.8**

5. CONCLUSIONS

This paper presented a new framework to extract sparse, localized facial motion components for facial expression recognition. The proposed spatially localized facial motion components will be useful not only for facial expression recognition but also for other local facial motion analysis like facial action unit recognition. The proposed approach is robust to appearance variations in different subject since it relies on facial motion flow. In the future, we will extend the model for nonlinearity using kernelization. In addition, instead of the simple 1-NN classifier and the majority voting scheme for the sequence data, we plan to experiment with classifiers like SVM and temporal sequence analysis using latent SVM.

Acknowledgements: This research of the first author was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2012R1A1B4003830).

6. REFERENCES

[1] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, "Active shape models: Their training and applications," *CVIU*, vol. 61, no. 1, pp. 38–59, 1995.

[2] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," in *ECCV*, 1998, vol. 2, pp. 484 – 498.

[3] Kolja Kähler, Jörg Haber, and Hans-Peter Seidel, "Geometry-based muscle modeling for facial animation," in *No description on Graphics interface 2001*, 2001, pp. 37–46.

[4] Li Zhang, Noah Snavely, Brian Curless, and Steven M. Seitz, "Spacetime faces: high resolution capture for modeling and animation," *ACM Trans. Graph.*, vol. 23, no. 3, pp. 548–558, 2004.

[5] Mette T. Posamntier and Hervé Abdi, "Processing faces and facial expressions," *Neuropsychology Review*, vol. 13, no. 3, pp. 113–143, 2003.

[6] Marie L. Smith, Garrison W. Cottrell, Frédéric Gosselin, and Philippe G. Schyns, "Transmitting and decoding facial expressions," *Psychological Science*, vol. 16, no. 3, pp. 184–189, 2005.

[7] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *PAMI*, vol. 32, no. 9, pp. 1627–1645, 2010.

[8] K. N. Tran, I. A. Kakadiaris, and S. K. Shah, "Part-based motion descriptor image from action recognition," *Pattern Recognition*, vol. 45, no. 7, pp. 2562–2572, 2012.

[9] Christian Thurau and Vàclav Hlavác, "Pose primitive based human action recognition in videos or still images," in *CVPR*, 2008, pp. 1–8.

[10] Yang Yang, Imran Saleemi, and Mubarak Shah, "Discovering motion primitives for unsupervised grouping and one-shot learning of human actions, gestures, and expressions," *PAMI*, vol. 35, no. 7, pp. 1635–1648, 2013.

[11] Muhammad Muneeb Ullah and Ivan Laptev, "Actlets: A novel local representation for human action recognition in video," in *ICIP*, 2012, pp. 777–780.

[12] Tianhong Fang, Xi ZHao, Omar Ocegueda, Shishir K. Shah, and Ioannis A. Kakadiaris, "3d/4d facial expression analysis: An davanced annoated face model approach," *Image and vision Computing*, vol. 30, pp. 738–749, 2012.

[13] Deeppak Ghimire and Joonwhoan Lee, "Geometric feature-based facial expression recognition in image squences using multi-class adaboost and support vector machines," *Sensors*, vol. 13, pp. 7714–7734, 2013.

[14] Pierre Lemaire, Boulbaba Ben Amor, Mohsen Ardabilian, Liming Chen, and Mohamed Daoudi, "Fully automatic 3d facial expression recognition using a region-based approach," in *Proceedings of the 2011 joint ACM workshop on Human gesture and behavior understanding*, 2011, pp. 53–58.

[15] J. Rafael Tena, Fernando De la Torre, and Iain Matthews, "Interative region-based linear 3d face models," *ACM Transactions on Graphics*, vol. 30, no. 4, 2011.

[16] Thomas Neumann, Kiran Varanasi, Stephan Wenger, Markus Wacker, Marcus Magnor, and Chirstian Theobalt, "Sparse localized deformation components," *ACM Transactions on Graphics*, vol. 32, no. 6, 2013.

[17] Rizwan Ahmed Khan, Alexandre Meyer, Hubert Konik, and Saida Bouakaz, "Human vision inspired framework for facial expression recognition," in *ICIP*, 2012, pp. 2593–2596.

[18] Ling Zhong, Qingshan Liu, Peng Yang, Bo Liu, Junzhou Huang, and Dimitris N. Metaxas, "Learning active facial patches for expression analysis," in *CVPR*, 2012, pp. 2562–2569.

[19] Rodophe Jenatton, Guillaume Obozinski, and Francis Bach, "Structured sparse principal component analysis," in *Proceedings of International Conference on Artificial Intelligence and Statistics*, 2010.

[20] Daniel D. Lee and Sebastian Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, pp. 788–791, 1999.

[21] Patrik O. Hoyer, "Non-negative matrix factorization with sparseness constraints," *Journal of Machine Learning Research*, vol. 5, pp. 1457–1469, 2004.

[22] László A. Jeni, Jeffrey M. Girard, Jeffrey F. Cohn, and Fernando De La Torre, "Continuous au intensity estimation using localized, sparse facial feature space," in *IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*, 2013, pp. 1–7.

[23] Symeon Nikitidis, A. Tefas, N. Nikolaidis, and I. Pitas, "Facial expresson recognition using clustering discriminant non-negative matrix factorization," in *ICIP*, 2011, pp. 3001–3004.

[24] Javier Sánchez Pérez, Enric Meinhardt-Llopis, and Gabriele Facciolo, "Tv-l1 optical flow estimation," *Image Processing On Line*.

[25] Lieven De Lathauwer, Bart De Moor, and Joos Vandewalle, "A multilinear singular value decomposition," *SIAM J. Matrix Anal. Appl.*, vol. 21, no. 4, pp. 1253–1278, 2000.

[26] Ivana Tošić and Pascal Frossard, "Dictionary learning: What is the right representation for my signal?," *IEEE Signal Processing Magazine*, pp. 27–38, 2011.

[27] Francis Bach, Rodolphe Jenatton, Julien Mairal, and Guillaume Obozinski, "Optimization with sparisty-inducing penalities," *Found. Trends Mach. Learn.*, vol. 4, no. 1, pp. 1–106, 2012.

[28] Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro, "Online dictionary learning for sparse coding," in *ICML*, 2009, pp. 689–696.

[29] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Mach. Learn.*, vol. 3, no. 1, pp. 1–122, 2011.

[30] Yifeng Li and Alioune Ngom, "Classification approach based on non-negative least squares," *Neurocomputing*, vol. 118, pp. 41–57, 2013.

[31] P. Lucey, J. F. Cohn, T. Kanade, and J. Saragih, "The extended cohn-kanade dataset(ck+): A complete dataset for action unit and emotion-specific expression," in *CVPR Workshop*, 2010, pp. 94–101.