Facial Expression Recognition Using Non-negative Matrix Factorization

Symeon Nikitidis, Anastasios Tefas and Ioannis Pitas
Artificial Intelligence & Information Analysis Lab
Department of Informatics
Aristotle University of Thessaloniki, Greece

E-mail: pitas@aiia.csd.auth.gr

www.aiia.csd.auth.gr

Presentation Outline

- Why is important to recognize facial expressions?
- Facial Expression From the Image Processing Perspective
  - Subspace Methods
  - NMF Basics
- Discriminant NMF Methods
  - Discriminant NMF (DNMF)
  - Projected Gradient Discriminant NMF (PGDNMF)
  - Subclass Discriminant NMF (SDNMF)
- Experimental results
- Conclusions

Informative Content of Facial Expressions

- Human communication by nonverbal means (gestures and essentially facial actions).
- Facial actions important source for understanding humans emotional state and intension.
- Key importance to various fields e.g. human behavior analysis, psychiatry, HCI, entertainment etc.

Universal Facial Expressions

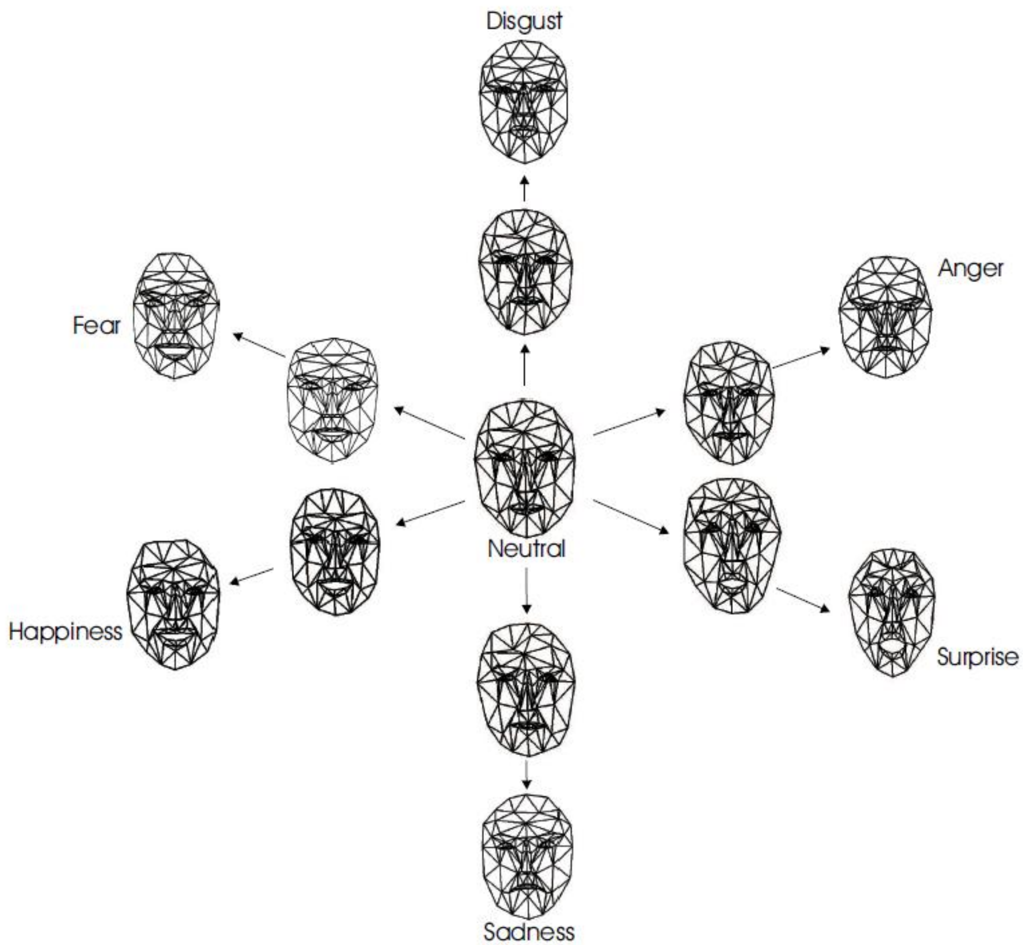
- Anger
- Fear
- Disgust
- Happiness
- Sadness
- Surprise
- Neutral

Dimensionality Reduction

- Facial image space dimensionality much higher than that required.
- Necessitates to perform dimensionality reduction to extract the appropriate facial features.
- Reduce computational complexity and boost performance of succeeding algorithms.
- Two popular approaches:
  - Grid-based Methods
  - Subspace Methods

Grid-Based Methods

- Grid is a parameterised face mask specifically developed for model-based coding of human faces .
- A popular facial wireframe model is the Candide grid.
- Facial expression information extraction is performed by facial feature point tracking.



Subspace Methods

- Among the most popular dimensionality reduction methods are the subspace based algorithms.
- Aim to discover latent facial features by projecting the facial image to a linear/nonlinear low dimensional subspace where a certain criterion is optimized.

Non-negative Matrix Factorization (NMF)

- Unsupervised matrix decomposition method.
- Requires both the decomposed data and the yielding factors to contain non-negative elements.
- Original data are reconstructed using only additive combinations of the resulting basic elements.
- Distinguishes NMF from PCA, ICA, SVD

Non-negative Matrix Factorization (NMF)

NMF considers factorizations of the form:

$$X \approx ZH$$

where $X \in R_{+}^{F*L}$ is the decomposed data matrix (1 column contains 1 image), $Z \in R_{+}^{F*M}$ contains the basis images and $H \in R_{+}^{M*L}$ the coefficients of the linear combination.

Non-negative Matrix Factorization (NMF)

- NMF training aims to learn different facial parts and approximate the appropriate weights to reconstruct the original facial images.
- Consistent with the psychological intuition of combining parts to form the whole regarding the objects representation in the human brain.

Non-negative Matrix Factorization (NMF)

- Approximation error metrics :
  - Kullback-Leibler (KL) divergence

$$\mathcal{O}(\mathbf{X}||\mathbf{ZH}) \triangleq \sum_{j=1}^{L} KL(\mathbf{x}_j||\mathbf{Zh}_j) = \sum_{j=1}^{L}\sum_{i=1}^{F}\left(x_{i,j}\ln(\frac{x_{i,j}}{\sum_k z_{i,k}h_{k,j}}) + \sum_k z_{i,k}h_{k,j} - x_{i,j}\right)$$

  - Frobenius norm

$$\mathcal{O}(\mathbf{X}||\mathbf{ZH}) \triangleq ||\mathbf{X} - \mathbf{ZH}||_F^2 = \sum_{j=1}^{L}\sum_{i=1}^{F}(x_{i,j} - [\mathbf{ZH}]_{i,j})^2$$

Non-negative Matrix Factorization (NMF)

- NMF optimization problem:

$$\min_{\mathbf{Z},\mathbf{H}} \mathcal{O}(\mathbf{X}||\mathbf{ZH})$$

$$\text{subject to:} \quad z_{i,k} \geq 0 \quad , h_{k,j} \geq 0, \quad \forall i, j, k.$$

- Using an appropriately designed auxiliary function and the EM algorithm a set of multiplicative update rules is derived.

$$h_{k,j}^{(t)} = h_{k,j}^{(t-1)} \frac{\sum_i z_{i,k}^{(t-1)} \frac{x_{i,j}}{\sum_l z_{i,l}^{(t-1)} h_{l,j}^{(t-1)}}}{\sum_i z_{i,k}^{(t-1)}}, \quad \acute{z}_{i,k}^{(t)} = z_{i,k}^{(t-1)} \frac{\sum_j h_{k,j}^{(t)} \frac{x_{i,j}}{\sum_l z_{i,l}^{(t-1)} h_{l,j}^{(t)}}}{\sum_j h_{k,j}^{(t)}}$$

Non-negative Matrix Factorization (NMF)

- NMF optimization problem is convex for either variable Z,H but non convex for both.
- Local minimum is reached.
- Update rules guarantee a non increasing behavior of the cost function.

Non-negative Matrix Factorization (NMF)

- Reached local minimum depends on the randomly selected initialization point.
- Sparseness achieved is rather a side effect than a goal, caused by the non negativity constraints.
- Tends to produce holistic basis images.

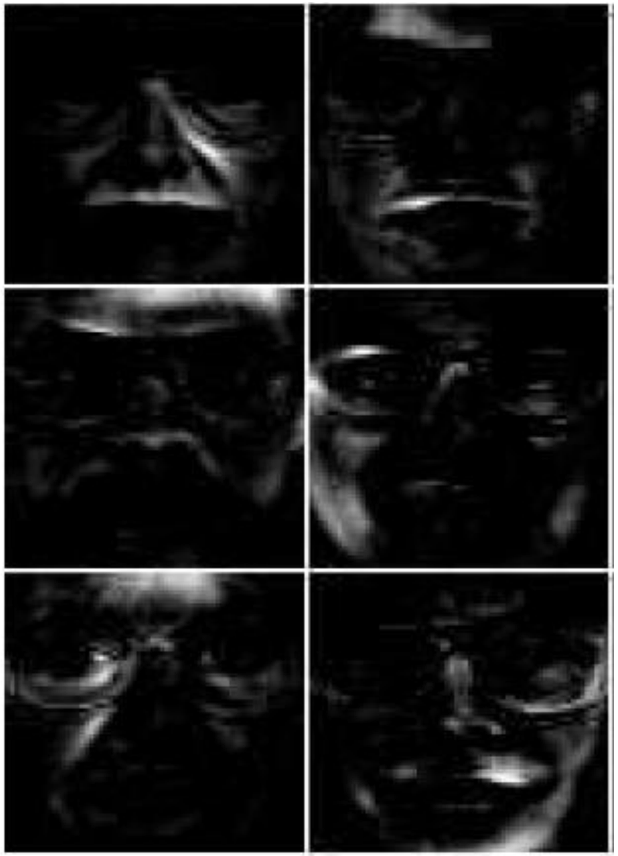
Notable NMF Variants

- Local NMF (LNMF)
- Discriminant NMF (DNMF)
- Projected Gradients DNMF (PGDNMF)
- Subclass Discriminant NMF (SDNMF)

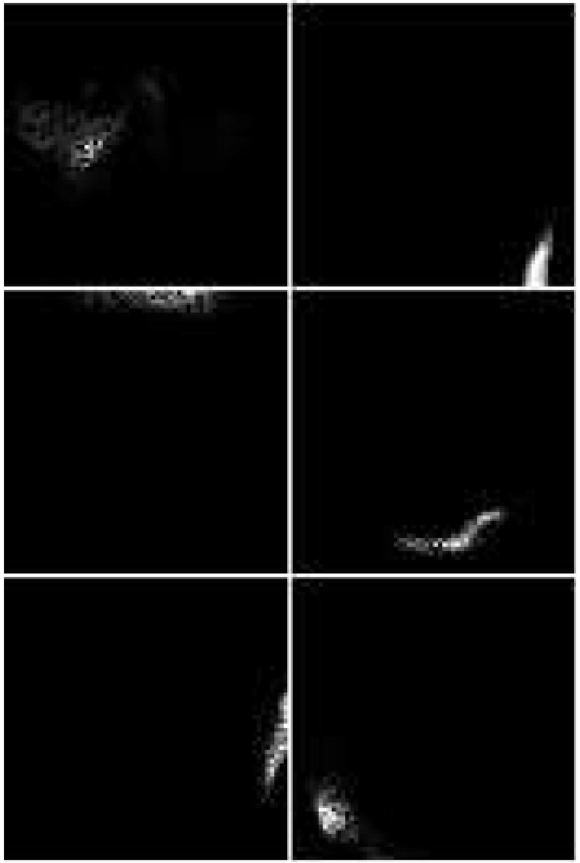
Local NMF

- To enhance basis images sparsity additional constraints imposed in the NMF decomposition cost function that:
  - enforce spatial locality of the basis images.
  - control sparsity.
  - minimize redundant information across different bases (orthogonal bases).

Local NMF



NMF Basis



LNMF Basis

Discriminant Non-negative Matrix Factorization

- DNMF is an attempt to introduce LDA-inspired discriminant constraints in the NMF decomposition cost function.
- DNMF aims to perform the projection to the low dimensional subspace in a discriminant manner.
- DNMF in contrary to NMF is a supervised learning algorithm.

Discriminant Non-negative Matrix Factorization

- DNMF uses the traces of the within and between scatter matrices also employed in Fisher discriminant criterion:

$$J(\mathbf{\Psi}) = \frac{\mathrm{tr}[\mathbf{\Psi}^T \mathbf{S}_b \mathbf{\Psi}]}{\mathrm{tr}[\mathbf{\Psi}^T \mathbf{S}_w \mathbf{\Psi}]}$$

- Seeks a projection matrix that enhances class separability.

Discriminant Non-negative Matrix Factorization

- Scatter matrices are defined considering the projected feature vectors.
- Class dispersion:

$$\mathbf{S}_b = \sum_{r=1}^{K} N_r (\boldsymbol{\mu}^{(r)} - \boldsymbol{\mu})(\boldsymbol{\mu}^{(r)} - \boldsymbol{\mu})^T$$

- Samples dispersion within the same class:

$$\mathbf{S}_w = \sum_{r=1}^{K} \sum_{\rho=1}^{N_r} (\boldsymbol{\eta}_\rho^{(r)} - \boldsymbol{\mu}^{(r)})(\boldsymbol{\eta}_\rho^{(r)} - \boldsymbol{\mu}^{(r)})^T$$

Discriminant Non-negative Matrix Factorization

- DNMF cost function:

$$D_{DNMF}(\mathbf{X}||\mathbf{ZH}) = \sum_{j=1}^{L} KL(\mathbf{x}_j||\mathbf{Zh}_j) + \alpha \mathrm{tr}[\acute{\mathbf{S}}_w] - \beta \mathrm{tr}[\acute{\mathbf{S}}_b]$$

- Goal of optimization is twofold:
  - Minimize decomposition error.
  - Find that projection matrix that maximizes the Fisher criterion.

Discriminant Non-negative Matrix Factorization

- DNMF enhances class separability by:
  - Achieving more compact classes formation in the projection subspace.
  - Classes are well discriminated in the projection subspace.
  - Optimization based on a properly designed auxiliary function.
  - The iterative optimization algorithm reaches a local minimum.

Discriminant Non-negative Matrix Factorization

- Optimization leads to the following multiplicative update rule for H:

$$h_{k,j}^{(t)} = \frac{T_1 + \sqrt{T_1^2 + 4(2\gamma - (2\gamma + 2\delta)\frac{1}{N_r})h_{k,j}^{(t-1)} \sum_i z_{i,k}^{(t-1)} \frac{x_{i,j}}{\sum_l z_{i,l}^{(t-1)} h_{l,j}^{(t-1)}}}}{2(2\gamma - (2\gamma + 2\delta)\frac{1}{N_r})}$$

$$T_1 = (2\gamma + 2\delta)(\frac{1}{N_r} \sum_{\lambda, \lambda \neq l} h_{k,\lambda}) - 2\delta\mu_k - 1$$

- Extract the discriminant features of an unknown test sample:

$$\acute{\mathbf{x}}_j = \mathbf{Z}^{\dagger}\mathbf{x}_j$$

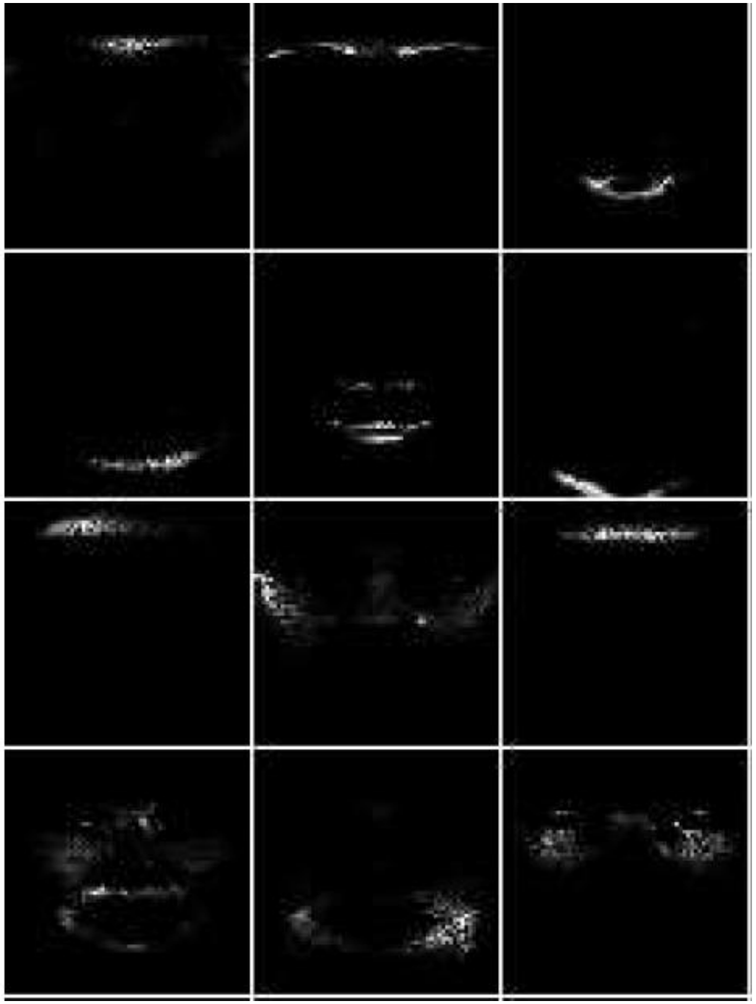
- Z^T can be also used as an appropriate alternative for the pseudo-inverse.

Discriminant Non-negative Matrix Factorization

- DNMF achieves to decompose a facial image in its discriminant parts.
- The resulting basis images correspond to salient facial features as: eyes, nose, mouth, eyebrows, etc.
- DNMF has been successfully applied for face verification, facial expression recognition and frontal facial view recognition.

Discriminant Non-negative Matrix Factorization

- DNMF basis images.
  - The resulting basis images correspond to salient facial features as: eyes, nose, mouth, eyebrows, etc.



Projected Gradients DNMF

- Multiplicative update rules only guarantee a non increasing behavior of the objective function.
- Convergence to a stationary limit point is not guaranteed.
- To assure stationarity, the constrained optimization problem is solved using projected gradients.

Projected Gradients DNMF

- The modified optimization problem minimizes the following cost function:

$$\mathcal{O}(\mathbf{X}||\mathbf{ZH}) \triangleq \frac{1}{2}||\mathbf{X} - \mathbf{ZH}||_F^2 + \frac{\alpha}{2}\mathrm{tr}[\acute{\mathbf{S}}_w] - \frac{\beta}{2}\mathrm{tr}[\acute{\mathbf{S}}_b]$$

- Two sub problems are defined considering one variable is kept fixed and optimization is performed for the other.

Projected Gradients DNMF

- We successively optimize the following sub problems
  - $$\min_{\mathbf{Z}} \mathcal{O}_1(\mathbf{Z}) \quad \text{subject to:} \quad z_{i,k} \geq 0 \quad , \quad \forall i, k$$
  - $$\min_{\mathbf{H}} \mathcal{O}_2(\mathbf{H}) \quad \text{subject to:} \quad h_{k,j} \geq 0 \quad , \quad \forall k, j.$$
- Considering the first sub problem, at a given iteration round *t* the following update rule is applied:

$$\mathbf{Z}^{(t)} = P[\mathbf{Z}^{(t-1)} - \alpha_t \nabla \mathcal{O}_1(\mathbf{Z}^{(t-1)})]$$

Projected Gradients DNMF

- Operator P[.] guarantees that no negative values are assigned to the updated elements.
- α_t is the learning step at iteration round *t*. Crucial since it determines convergence speed.
- Iterating this update rule a sequence of minimizers $\{\mathbf{Z}^{(t)}\}_{t=1}^{\infty}$ is generated where it is guaranteed to find a stationary point.

Projected Gradients DNMF

- Stationarity condition check step to terminate optimization:

$$||\nabla^P \mathcal{O}_1(\mathbf{Z}^{(t)})||_F \leq e_{\mathbf{Z}} ||\nabla^P \mathcal{O}_1(\mathbf{Z}^{(1)})||_F$$

- $\nabla^P \mathcal{O}_1(\mathbf{Z}^{(t)})$ is the projected gradient:

$$[\nabla^P \mathcal{O}_1(\mathbf{Z}^{(t)})]_{i,k} = \begin{cases} [\nabla \mathcal{O}_1(\mathbf{Z}^{(t)})]_{i,k} & \text{, if } z_{i,k} > 0 \\ \min\left(0, [\nabla \mathcal{O}_1(\mathbf{Z}^{(t)})]_{i,k}\right) & \text{, if } z_{i,k} = 0 \end{cases}$$

Projected Gradients DNMF

- $e_{\mathbf{Z}}$ is a predefined stopping tolerance.
  - A small value leads to a termination after a large number of iterations.
  - A value close to 1 results in a premature termination.
- A similar optimization process is followed for the weights matrix.

Projected Gradients DNMF

- Discriminant constraints are only involved during optimization of the weights matrix.

- Projected gradients advantages:
  - Well established optimization properties.
  - Achieve faster convergence.
  - Achieve better performance.

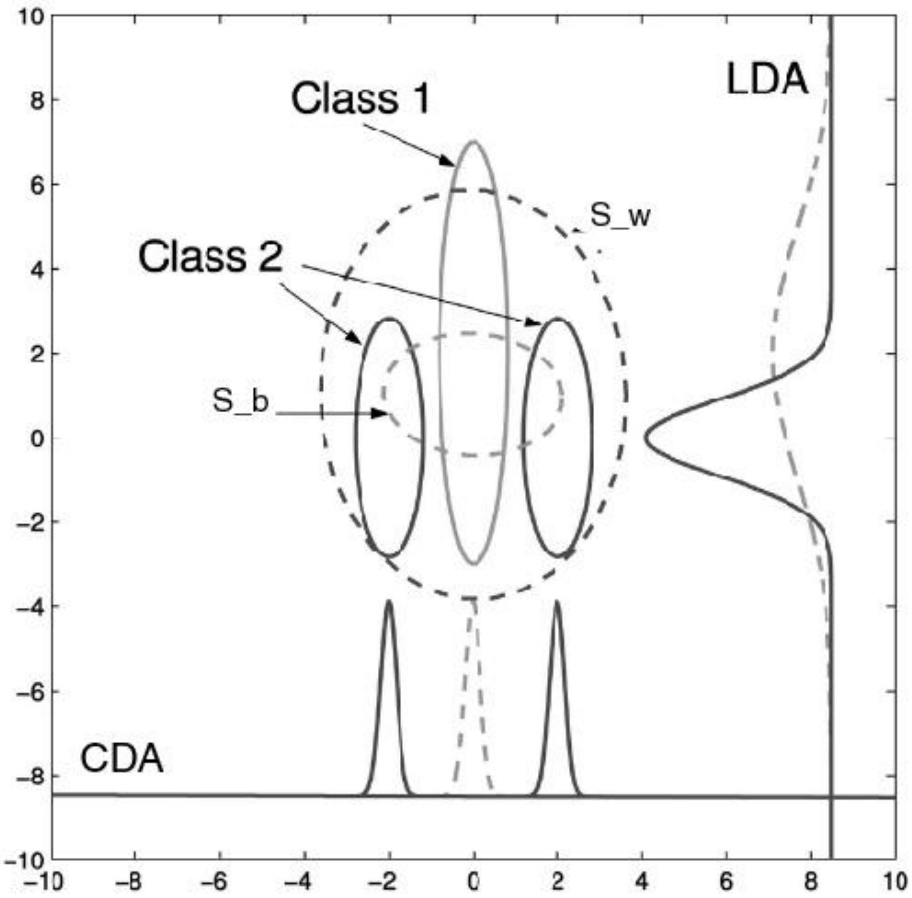
Subclass Subspace Techniques

- LDA limitations:
  - LDA assumes that the sample vectors of each class are generated from underlying multivariate Gaussian distributions having a common covariance matrix but with different class means.
  - Assuming that each class is represented by a single compact data cluster, the problem of nonlinearly separable classes can not be solved.

Subclass Subspace Techniques

- In this two class dimensionality reduction problem LDA will fail to reduce the dimensionality of the original feature space to one because the second class corresponds to two disjoint distributions.

- One can solve this problem by dividing the second class into two subclasses.



Subclass Subspace Techniques

- Typically, in real world applications, data usually do have a subclass structure.
- Common case in facial expression recognition, since there is no unique way that people express certain emotions, hence leading to expression subclasses.
- Other factors such as facial pose, texture and illumination variations, enhance the subclass structure of facial expressions

Subclass Subspace Techniques

- Clustering based Discriminant Analysis (CDA) regards that data inside each class form various subclasses, where each one is approximated by a Gaussian distribution.
- Approximate the underlying distribution of each class by a mixture of Gaussians

Subclass Discriminant NMF (SDNMF)

- SDNMF is a supervised learning algorithm.
- Requires class and subclass labels.
- Attempts to find discriminant projections by imposing discriminant criteria that assume multimodality of the available train data.

Subclass Discriminant NMF (SDNMF)

- The decomposition cost function imposes CDA inspired discriminant criteria that aim to enhance class separability in the reduced dimensional projection subspace by achieving better discrimination of the respective subclasses.

$$D_{SDNMF}(\mathbf{X}||\mathbf{ZH}) = \sum_{j=1}^{L} KL(\mathbf{x}_j||\mathbf{Zh}_j) + \frac{\alpha}{2}\mathrm{tr}[\mathbf{\Sigma}_w] - \frac{\beta}{2}\mathrm{tr}[\mathbf{\Sigma}_b]$$

Subclass Discriminant NMF (SDNMF)

- Within subclass scatter matrix represents the scatter of the projected sample vector coefficients around their subclass mean.

$$\Sigma_w = \sum_{r=1}^{n} \sum_{\theta=1}^{C_r} \sum_{\rho=1}^{N_{(r)(\theta)}} \left(\boldsymbol{\eta}_\rho^{(r)(\theta)} - \boldsymbol{\mu}^{(r)(\theta)}\right) \left(\boldsymbol{\eta}_\rho^{(r)(\theta)} - \boldsymbol{\mu}^{(r)(\theta)}\right)^T$$

  - Minimizing its trace will result in more compact subclasses formation.

Subclass Discriminant NMF (SDNMF)

- Between subclass scatter matrix defines the scatter of the mean vectors between all subclasses that belong to different classes.

$$\Sigma_b = \sum_{i=1}^{n} \sum_{r, r \neq i}^{n} \sum_{j=1}^{C_i} \sum_{\theta=1}^{C_r} \left(\boldsymbol{\mu}^{(i)(j)} - \boldsymbol{\mu}^{(r)(\theta)}\right) \left(\boldsymbol{\mu}^{(i)(j)} - \boldsymbol{\mu}^{(r)(\theta)}\right)^T$$

  - Maximizing its trace will enhance separability between subclasses belonging to different classes.

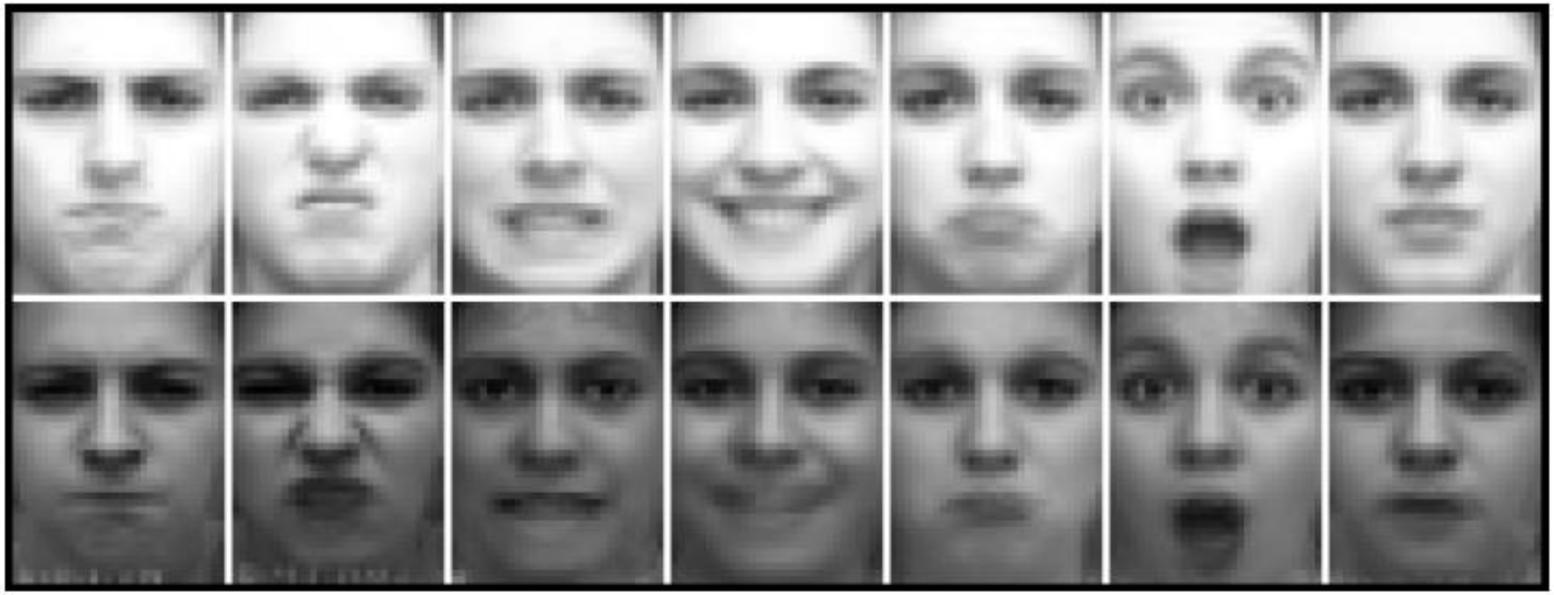
Subclass Discriminant NMF (SDNMF)

- Goal of optimization is twofold:
  - Minimize decomposition error.
  - Find that projection matrix that maximizes the CDA inspired criterion.
- Optimization is performed using an auxiliary function.
- Derived multiplicative update rules consider both samples class origin and clusters formation inside each class.

Experimental Results

- Experiments performed on Cohn-Kanade and JAFEE databases.
- Each facial image was isotropically scaled, to a fixed size of 30*40 pixels and converted to grayscale.
- Training set was used to learn the basis images for the low dimensional projection space, while test set to report the facial expression recognition accuracy rates.
- Classification was performed by feeding the projected to the lower dimensional space discriminant facial expression representations to a linear SVM classifier.

Experimental Results



- Mean expressive image for the two more distinct subclasses of each class. (considering 3 subclasses partitioning.)
- The diverse illumination conditions in the Cohn-Kanade database are evident.

Experimental Results

Method	Accuracy Rate	Subspace Dimensionality
SDNMF $C_r = 2$	69.05%	190
SDNMF $C_r = 3$	68.31%	182
DNMF	66.08%	166
NMF	64.85%	134

- An increase by more than 4% has been achieved by incorporating the CDA inspired discriminant constraints in the NMF cost function.

Experimental Results

- Database Enrichment
  - Examine the sensitivity of NMF based algorithms w.r.t. registration errors of the facial ROI .
  - Propose a training set enrichment approach for improving the performance of subspace learning techniques.

Experimental Results

- Database Enrichment
  - Geometrically transformed versions of each initial facial image.
  - Generated 24 different geometrical distortions applied to each initial facial image by varying the eyes center position by a single pixel along a cross shaped shift direction.
  - 24 different translated, scaled and rotated versions of each original facial image in the database.

Experimental Results

- Enriched training facial image samples resulting from a single image of the Cohn-Kanade database

Experimental Results

Database	Kanade	Kanade Enriched	JAFEE	JAFEE Enriched
NMF	64.85%	62.45%	**56.72**%	53.69%
DNMF	**66.08**%	**69.20**%	47.40%	**55.69**%

Experimental Results

Method	JAFFE	JAFFE Enriched
SDNMF $C_r = 2$	48.32%(185)	59.62%(165)
SDNMF $C_r = 3$	49.26%(190)	**62.21**%(175)
DNMF	47.40%(178)	55.69%(160)
NMF	**56.72**%(106)	53.69%(135)

- Experimental Results on the JAFFE database.
  - Classification accuracy increased across all discriminant NMF methods.
  - SDNMF recognition accuracy increased by almost 13% compared with that attained using the original training data.

Conslusions

- Diversity of facial expression problem.
- Discriminant NMF methods successfully decomposed a facial image into its salient parts.
- This decomposition improves performance of subsequent classification algorithms.
- Multimodality of facial expression image samples can be appropriately handled using CDA inspired discriminant constraints.

Thank you

- Information on cited published works:
  - **http://poseidon.csd.auth.gr/LAB_PUBLICATIONS/Journals/index.php**
- Research Projects:
  - MOBISERV FP7-248434 (**http://www.mobiserv.eu**), An Integrated Intelligent Home Environment for the Provision of Health, Nutrition and Mobility Services to the Elderly.
  - i3DPost FP7-211471

  (**http://www.i3dpost.eu/**), Intelligent 3D content extraction and manipulation for film and games.