Improving Vision Substitution Using Self-Organizing Feature Maps

ECE 5524

Michael Hopkins
Virginia Polytechnic Institute and State University
Department of Electrical and Computer Engineering
Blacksburg, VA 24060

***Abstract*—This paper investigates the possibility of using an image compression scheme to enhance the performance of vision substitution systems for the blind. As a case scenario, we propose a system which allows the blind to perceive human faces using an artificial vision system and a tactile interface. Traditional vision substitution systems encode visual information by mapping pixel intensities directly to the lattice of a tactile array. We examine the various issues with this approach and investigate methods to encode a source image as a linear combination of basis images where the activations of nodes in a tactile array represent the weights of each basis. The basis images are trained using machine learning techniques such as k-means, self-organizing maps and non-negative matrix factorization.**

I. Introduction

Vision substitution systems aim to restore visual perception in the blind by channeling optical information through an alternative sense such as touch or sound. Traditionally, electro- and vibrotactile arrays have been used to map pixel intensities from pre-processed images onto sensitive areas of the skin or tongue. Audio encodings have also been proposed that transform two dimensional pixel maps into time-varying signals, allowing the blind to process near real-time video streams. Since the 1960's, a number of researchers have investigated the merit of such systems [9],[10],[11],[12]. Users have achieved limited success, in some cases being able to identify common objects in controlled environments [8]. As of today, however, these systems have not been widely adopted by the blind community [13].

The primary problem encountered in vision substitution systems is the low resolution and dynamic range of the sensory input devices. There a roughly 100 million rods in the human retina, which form a dense field of highly sensitive photoreceptors. Current vision substitution devices have very limited spatial resolutions that are orders of magnitude lower than the human visual system. State of the art systems can encode roughly 600 grayscale pixels using a tactile array. Additionally, the use of the tactile interfaces presents a number of unique challenges.

Tactile interfaces usually consist of a set of electro- or vibrotactile nodes arranged in a two-dimensional grid (see fig. 1). It is common for each node to represent a single pixel in an image. When a node is activated (by vibration or other means) the user interprets the intensity as the brightness of the pixel at that location. In this manner, the brightness pattern in an image can be mapped directly to the tactile array through simultaneous activation of many nodes.

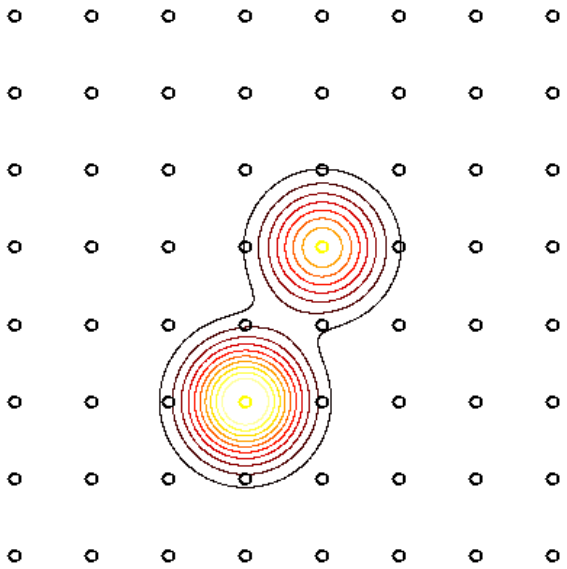

Fig. 1: An 8×8 tactile array with two active (vibrating) nodes

A problem with this design is that activations propagate through the skin, resulting in blurring between neighboring nodes. This decreases the effective spatial resolution of the input signal. Even with a single node activated, it may be difficult for the user to perceive a precise location in the grid due to the limited sensitivity of the somatosensory system. As a result, experiments have shown that users have difficulty processing detailed tactile-visual information and can even experience sensory overload [8]. To compensate for these effects, some vision substitution systems have tried pre-processing the input images using edge filters in order to improve sensitivity, but this further reduces the amount of information available to the user.

A. Objectives

To improve the effectiveness of vision substitution systems, we propose the use of an encoding which compresses visual information into a more efficient tactile representation. Figure 2 shows the related system architecture. Traditional approaches rely on the somatosensory system to process large

amounts of tactile information in the same way that the vision system processes optical information. Instead of encoding tactile-visual information in the format of a two-dimensional image, we investigate methods to encode a source image as a linear combination of basis images where the activations of nodes in a tactile array represent the weights of each basis.

For simplicity, we denote each basis image by a column vector $\boldsymbol{w}_i \in \mathbb{R}^d$ where d is the dimensionality (i.e. the number of pixels) of the image. Given a set of k nodes $\{\eta_1, \eta_2, \dots, \eta_k\}$ in a tactile array and a set of k basis images, we aim to approximate a source image $\boldsymbol{v} \in \mathbb{R}^d$ by the weighted sum

$$\boldsymbol{v} \approx \sum_{i=1}^{k} h_i \boldsymbol{w}_i = \boldsymbol{W}\boldsymbol{h} \qquad (1)$$

where $\boldsymbol{h} \in \mathbb{R}^k$ is the vector of node activations $[h_1\ h_2 \dots h_k]^T$. The $d \times k$ matrix $\boldsymbol{W} = [\boldsymbol{w}_1 | \boldsymbol{w}_2 | \dots | \boldsymbol{w}_k]$ is sometimes referred to as a dictionary or codebook, and its range represents the subspace of images which can be perfectly reconstructed from the bases. If $k < d$ then the vector $\boldsymbol{h}$ represents a compressed version of the source image $\boldsymbol{v}$. By reducing the dimensionality of the incoming visual information, we hope to improve the brain's ability to process the encoded tactile signal.

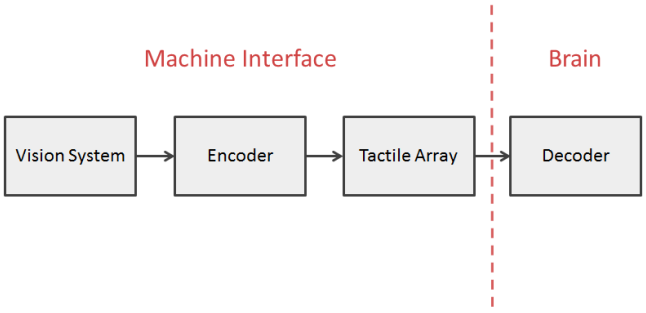

Fig. 2: System architecture of the proposed vision substitution system

Having specified the general format of the encoding in (1), the question becomes how to choose the best set of basis images $\boldsymbol{W}$ to maximize the performance of a vision substitution system. The choice and arrangement of bases will directly dictate the characteristics of the encoded tactile signal $\boldsymbol{h}$. By introducing a non-trivial encoding scheme into the human machine interface, it is assumed that the human brain will learn to decode the signal back into a useful representation (see fig. 2). Based on this assumption, we define the following properties of a good encoding:

1. **Simplicity** - The encoding should be intuitive, easy to understand and easy to learn.
2. **Sparsity** - For any given image, the number of active nodes should be minimized.
3. **Locality** - Adjacent nodes should convey similar information.
4. **Generality** - The encoding should approximate a large variety of images.

The motivation in designing a simple encoding is to maximize the likelihood that the human brain will learn to decode the signal properly. The sparsity constraint is intended to reduce the complexity of the tactile information being presented to the user by limiting the majority of the terms in $\boldsymbol{h}$ to zero. The locality constraint is intended to reduce the effects of blurring between adjacent nodes in the tactile array. Ideally, if basis images $\boldsymbol{w}_i$ and $\boldsymbol{w}_j$ are nearest neighbors in d-dimensional space then nodes η_i and η_j should be neighbors within the tactile array.

The final desirable property of generality, is at odds with the first three. This stems from the no free lunch theorem; it is not possible to design a compression scheme that will perform well for the set of all possible images. For this reason, we will assume that the target vision substitution system is restricted to a specific domain. By designing an encoding that represents only a certain type of visual information, we can choose the most efficient dictionary $\boldsymbol{W}$ given any available *a priori* knowledge concerning the format the input image.

As a case scenario, we will consider designing a system that will allow the blind to perceive human faces using a tactile array. An external vision system performs face detection, preprocesses and crops an input image to the frontal view of a face. An encoder then compresses the information by choosing the set of activations $\boldsymbol{h}$ which minimize the error between the source image and the decoded image. Finally the information is presented to the user through a tactile array. In the following sections we investigate methods for obtaining an optimal dictionary $\boldsymbol{W}$ for the encoder in this system given the unique constraints of the human-machine interface specified in fig. 2. A number of possible encoding schemes are reviewed, including vector quantization, self-organizing maps and non-negative matrix factorization.

II. Encodings

A. Vector Quanization (VQ)

In vector quantization (VQ) the encoded signal $\boldsymbol{h}$ in (1) is constrained to have exactly one non-zero element, thereby restricting each image to be approximated by a single basis, which we will denote $\boldsymbol{w}_c$. For a given source image, the optimal choice of $\boldsymbol{w}_c$ is the dictionary vector which best matches the input vector $\boldsymbol{v}$:

$$\|\boldsymbol{v} - \boldsymbol{w}_c\|^2 = \min_i \{\|\boldsymbol{v} - \boldsymbol{w}_i\|^2\} \qquad (2)$$

VQ is analogous to nearest neighbor classification, where the columns of $\boldsymbol{W}$ represent the set of possible classes and $\boldsymbol{h}$ represents the zero-one loss. In order to minimize the mean approximation error, the basis vectors $\{\boldsymbol{w}_1, \boldsymbol{w}_2, \dots, \boldsymbol{w}_k\}$ are typically selected through an iterative training process referred to as competitive learning, described briefly below.

Let $V = \{\boldsymbol{v}_1, \boldsymbol{v}_2, \dots, \boldsymbol{v}_n\}$ be a set of training images sampled from the range of all possible source images. Then

assuming the basis vectors $\{\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_k\}$ have been initialized to some arbitrary values, the optimal bases can be approximated by iterative steepest descent using the following stochastic update rules:

$$\boldsymbol{w}_c(t+1) = \boldsymbol{w}_c(t) + \alpha(t)[\boldsymbol{v}(t) - \boldsymbol{w}_c(t)] \quad (3)$$
$$\boldsymbol{w}_j(t+1) = \boldsymbol{w}_j(t) \quad , \quad j \neq c \quad (4)$$

where t is the index of iteration, $\alpha(t)$ is the learning rate and $\boldsymbol{v}(t)$ is sampled randomly from the training set V [2]. The algorithm is competitive in the sense that at each iteration only the "winning" basis, i.e. the dictionary vector that best matches the input image, is updated [3]. A closely related clustering technique is the k-means algorithm, which employs a batch update instead of the stochastic updates given in (3) and (4) [2].

We have applied VQ to the MIT CBCL face database [14]. K-means clustering was performed on a set of 2429 images of faces with $k = 64$. Fig. 3 shows the VQ coding for an example face, with the resulting basis images mapped to the lattice of an 8×8 tactile array. The encoded signal, representing the node activations in the tactile array, was calculated using nearest neighbor classification as described in (2). The decoded image was calculated using (1).

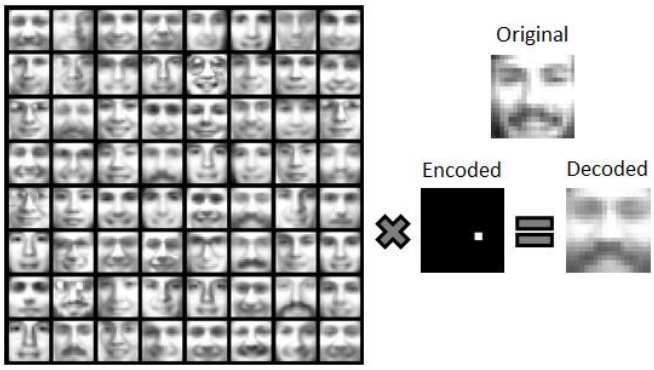

Fig. 3: An 8×8 mapping of basis images trained using k-means ($k = 64$) on a set of $2,429$ 19×19 images of faces from the MIT CBCL database. An example face is shown with its VQ encoding and resulting decoded image.

In the context of the proposed vision substitution system, VQ is a very simple encoding scheme, which could make it easy for the human brain to learn. VQ coding also offers the advantage of maximum sparsity. Since each source face image is encoded by a single node activation, the corresponding basis images represent prototypical faces.

Unfortunately, VQ utilizes only a very small percentage of the available bandwidth of a tactile interface. In effect, an 8 x 8 tactile array encoded using VQ is capable of representing only 6 bits of information per image. Furthermore, there is no inherent organization in the mapping of basis images to the tactile array. As seen in fig. 3, neighboring faces are generally unrelated in appearance. This violates the desirable locality properties discussed in section I. As a result, spatial uncertainty could greatly degrade the decoded image (see the results). The coding scheme discussed in the following section aims to correct this problem.

B. Self-Organizing Map (SOM)

Under the proposed system, each basis image $\boldsymbol{w}_j$ maps to a single node η_j in a tactile array, as shown in fig. 3. In VQ the mapping of basis images to physical nodes is arbitrary; the underlying topology of the tactile array is not taken into account during the training of the bases. The Self-Organizing Map (SOM) [2] is an extension to competitive learning which aims to preserve local and global neighborhoods when basis features are mapped to some target space (in our case, the lattice of the tactile array).

The SOM algorithm is almost identical to the competitive learning algorithm described in the previous section. Instead of the winner-take-all updates in (3-4), however, SOM updates the basis vectors at each node according to the following stochastic update rule

$$\boldsymbol{w}_j(t+1) = \boldsymbol{w}_j(t) + \boldsymbol{\Phi}_{cj}(t)[\boldsymbol{v}(t) - \boldsymbol{w}_c(t)] \quad (5)$$

$\forall j \in [1, k]$, where $\boldsymbol{\Phi}_{\mathbf{cj}}(t)$ is a scalar kernel function that determines the gain of the update for each basis vector $\boldsymbol{w}_j$. The values of $\boldsymbol{\Phi}_{\mathbf{cj}}(t)$ depend on the distance between η_j and η_c, the node with the "winning" basis $\boldsymbol{w}_c$. For our experiments, the $k \times k$ matrix $\boldsymbol{\Phi}(t)$ is defined as

$$\boldsymbol{\Phi}_{ij}(t) = \alpha(t) e^{-d_{ij}^2/\sigma^2(t)} \quad (6)$$

where d_{ij} is the Euclidean distance between nodes η_i and η_j within the tactile array. The σ parameter determines the attenuation of the update gain as this distance increases. The key to the SOM algorithm is that basis vectors that map to nodes nearer to the winning node are updated more heavily than nodes that are further away. This has the effect of organizing similar basis features into local neighborhoods in the topology of the target space.

Fig. 4 shows the mapping of basis images for the MIT CBCL face database trained using the SOM algorithm. The resulting encoded and decoded signals are shown for the same source image as in fig. 3. As in VQ, the encoded signal is constrained to have one non-zero element, and is calculated using nearest-neighbor classification.

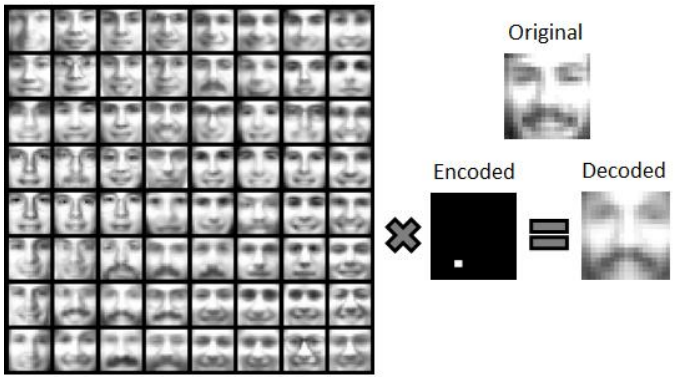

Fig. 4: An 8×8 mapping of basis images trained using SOM ($k = 64$) on the MIT CBCL face database. The SOM algorithm was run for 100000 iterations with a linear decaying $\alpha(t)$ and exponentially decaying $\sigma(t)$.

Notice that the faces of neighboring nodes within the tactile array are similar in appearance. In comparison to VQ, SOM reduces the complexity of the mapping by grouping similar basis images. This helps reduce the effects of spatial uncertainty and blurring between nodes (see the results). Additionally, we hypothesize that it will be easier for the brain to learn this mapping due to its organized structure.

SOM fulfills three of the four properties of a good encoding scheme for a vision substitution system defined in section I. The encoding is simple and sparse like VQ, and also local due to the topology preserving effects of the SOM. By using nearest neighbor classification, however, SOM suffers the same low bandwidth as VQ. As a result, neither encoding scheme meets the fourth requirement of generality. Given the limited number of nodes in a tactile array, it is not possible to accurately encode a large variety of facial images using holistic basis images. In the next section we discuss how the expressiveness of the encoding can be improved using a parts based representation.

C. Non-negative Matrix Factorization(NMF)

In VQ and SOM, the encoded signal $\boldsymbol{h}$ was constrained to have exactly one non-zero element, resulting in a dictionary of holistic basis images. In non-negative matrix factorization (NMF) [1], this condition is relaxed; the encoded signal is required only to have all non-negative elements. This leads to multiple activations per source image, increasing the complexity of the encoding and utilizing more of the available bandwidth of the tactile array.

As the name suggests, NMF is formulated as a matrix factorization problem. Given a source matrix $\boldsymbol{V}$, NMF aims to find the most accurate factorization

$$\boldsymbol{V} \approx \boldsymbol{WH} \qquad (7)$$

subject to the constraint $\forall_{ij}$: $\boldsymbol{W}_{ij} \geq 0$, $\boldsymbol{H}_{ij} \geq 0$. Given a set of training images $\{\boldsymbol{v}_1, \boldsymbol{v}_2, \dots, \boldsymbol{v}_n\} \in \mathbb{R}^d$, we can construct a dictionary of basis images $\boldsymbol{W}$ by applying NMF to the $d \times n$ matrix $\boldsymbol{V} = [\boldsymbol{v}_1 | \boldsymbol{v}_2 | \dots | \boldsymbol{v}_n]$. The columns of the resulting $k \times n$ matrix $\boldsymbol{H} = [\boldsymbol{h}_1 | \boldsymbol{h}_2 | \dots | \boldsymbol{h}_n]$ represent the encoded signals for each of the source images in $\boldsymbol{V}$.

The factorization can be estimated using iterative methods which search for a local minimum in the objective function:

$$D(\boldsymbol{V} \| \boldsymbol{WH}) = \sum_{i=1}^{d} \sum_{j=1}^{n} \left[\boldsymbol{V}_{ij} \log \left(\frac{\boldsymbol{V}_{ij}}{(\boldsymbol{WH})_{ij}} \right) - \boldsymbol{V}_{ij} + (\boldsymbol{WH})_{ij} \right] \qquad (8)$$

where $D(\boldsymbol{V} \| \boldsymbol{WH})$ is a measure of the divergence of $\boldsymbol{V}$ from $\boldsymbol{WH}$ [5]. Assuming $\boldsymbol{W}$ and $\boldsymbol{H}$ are initialized to arbitrary non-negative values, the divergence can be monotonically decreased using Lee and Seung's multiplicative update rules [1]:

$$\boldsymbol{W}_{ij} \leftarrow \boldsymbol{W}_{ij} \sum_{\mu} \frac{\boldsymbol{V}_{i\mu}}{(\boldsymbol{WH})_{i\mu}} \boldsymbol{H}_{j\mu} \qquad (9)$$

$$\boldsymbol{H}_{j\mu} \leftarrow \boldsymbol{H}_{j\mu} \sum_{i} \frac{\boldsymbol{V}_{i\mu}}{(\boldsymbol{WH})_{i\mu}} \boldsymbol{W}_{ij} \qquad (10)$$

Fig. 5 shows the mapping of basis images for the MIT CBCL face database trained using NMF. The resulting encoded and decoded signals are shown for the same source image as in Figs. 3 and 4. The encoded vector $\boldsymbol{h}$ is calculated by solving the non-negative least squares problem $\boldsymbol{Wh} = \boldsymbol{v}$ given $\boldsymbol{h} \geq \boldsymbol{0}$. The decoded image is calculated using (1).

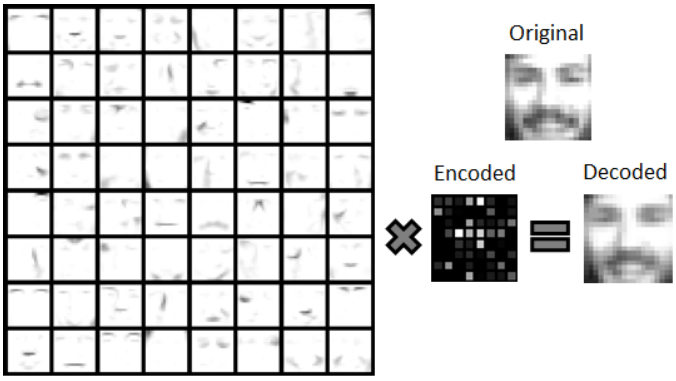

Fig. 5: An 8×8 mapping of basis images trained using NMF ($k = 64$) on the MIT CBCL face database. The NMF algorithm was run for 250 iterations.

As seen in Fig. 5, the basis images resulting from NMF are not holistic as in VQ and SOM; individual bases represent parts of faces, resulting in multiple node activations for each source image. These results mirror the findings in [1], where the authors argued that NMF is well suited for learning the parts of objects due to its non-negative constraints. We can conceptualize the decoded face as being formed by the sum of its individual features, a view that is consistent with the weighted summation in (1). While the use of multiple activations increases the complexity of the encoding, it also decreases the approximation error.

By enforcing non-negativity in both the encoding and the bases, NMF offers two main benefits over alternative factorizations such as PCA, ICA and sparse coding (SC). First, a non-negative encoding is easier to represent using a tactile array since only one direction of intensity is required. Second,

when negative weights and bases are allowed, the encoding can be unintuitive due to complex cancellations among bases [1]. By utilizing only non-negative combinations, the NMF encoding is easier to understand and may be easier to learn.

While NMF is more expressive than VQ, it is necessarily less sparse and therefore more complex. Although NMF can be argued to promote inherently sparse encodings, algorithms have been proposed to further increase the sparsity of the factorization. Non-negative sparse coding (NNSC) [6] is a variation of NMF which optimizes the modified cost function

$$\boldsymbol{J}(\boldsymbol{W},\boldsymbol{H}) = \frac{1}{2}\|\boldsymbol{V}-\boldsymbol{W}\boldsymbol{H}\|^2 + \lambda\sum_{ij}\boldsymbol{H}_{ij} \qquad \textbf{(11)}$$

Here the first term represents the mean squared error of the approximation, which is substituted for the divergence criteria defined in (8). The second term supports sparsity in columns of $\boldsymbol{H}$ by penalizing the sum of its elements. Other variations such as LMNF [7] have been designed to support sparsity in both the bases and the encodings, which may prove beneficial for learning bases that correspond to intuitive features.

NMF results in an unorganized mapping of basis images to the tactile array. Like VQ, there is no correlation between the topology of the nodes in the tactile array and the characteristics of their associated bases, which violates the locality constraint discussed in section I. In the next section we will discuss possible extensions to NMF which will introduce locality into the encoding in the same way that SOM did for VQ.

D. Self-Organizing NMF

NMF offers the advantage of generating a parts-based dictionary of basis images, resulting in an intuitive and expressive encoding scheme. However, as shown in fig. 5, the mapping of basis images to the tactile array is unorganized. To reduce the complexity of the encoding, we propose an extension to NMF which organizes parts-based features into global and local neighborhoods, analogous to SOM's organization of holistic features. Fig. 6 shows an idealized mapping of facial features which might be learned by such a method.

In this example, common facial features (e.g. the eyes, nose, mouth, etc.) are organized into global neighborhoods within the tactile array. Within each global neighborhood, are local neighborhoods of features. For instance, in the "Mouth" region of the tactile array, there might be a group of bases for representing smiles and a group for representing frowns. For a typical source image, the encoded signal would exhibit strong localized activations within each of the global neighborhoods. A notable advantage of this type of hierarchical organization is that these activations are spread across the tactile array.

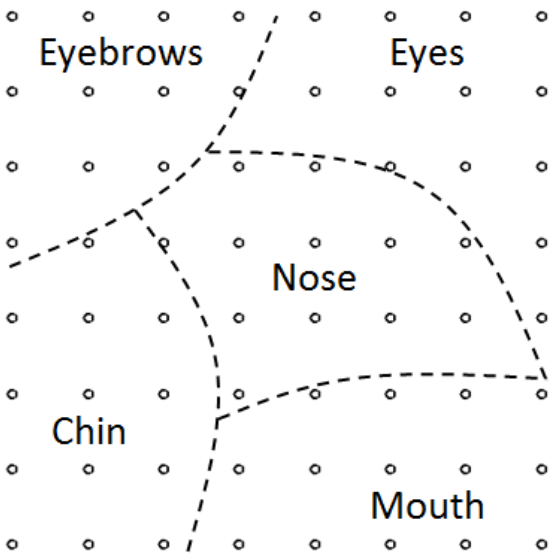

Fig. 6: Idealized mapping of facial features learned by a Self-Organizing NMF algorithm

The spatial organization and parts-based representation of the proposed Self-Organizing NMF encoding draws a number of similarities to the receptive fields of the brain's inferior temporal cortex (IT), which has been shown to play a large role in visual object recognition. In [4] an extension to the NMF algorithm called topographic NMF (TNMF) is proposed to induce topological organization in the learned bases. This is done by modifying the standard cost function in **(8)** to

$$J(\boldsymbol{W},\boldsymbol{H}) = \sum_{i=1}^{d}\sum_{j=1}^{n}\left[\boldsymbol{V}_{ij}\log(\boldsymbol{W}\boldsymbol{\Phi}\boldsymbol{H})_{ij} - (\boldsymbol{W}\boldsymbol{\Phi}\boldsymbol{H})_{ij}\right] \qquad \textbf{(12)}$$

where the $k \times k$ matrix $\boldsymbol{\Phi}$ determines the spread of activations among neighboring nodes and is defined in the same way as $\boldsymbol{\Phi}(t)$ in the SOM algorithm (5).

An alternative approach would consist of training an over-complete dictionary of features using NMF, and then applying SOM on the resulting basis images in order to obtain a final organized mapping of bases. Future work will investigate the performance of TNMF and other possible implementations of the proposed Self-Organizing NMF encoding.

III. RESULTS

Although the proposed vision substitution system has not been implemented as a physical system, preliminary tests have been designed to compare the performance of the encoding schemes discussed in the previous sections. First, to observe the effects of blurring between nodes in a tactile array, we applied Gaussian filtering to the encoded signals attained using VQ, SOM and NMF. For each method, a dictionary of basis images was trained on the MIT CBCL face database with $k = 64$. The resulting encodings for an example source image were mapped to an 8×8 grid and convolved with Gaussian filter kernels of various sized. The results are shown in fig. 7,

along with the mean squared errors of the decoded images (measured in units of intensity in the range [0,1]).

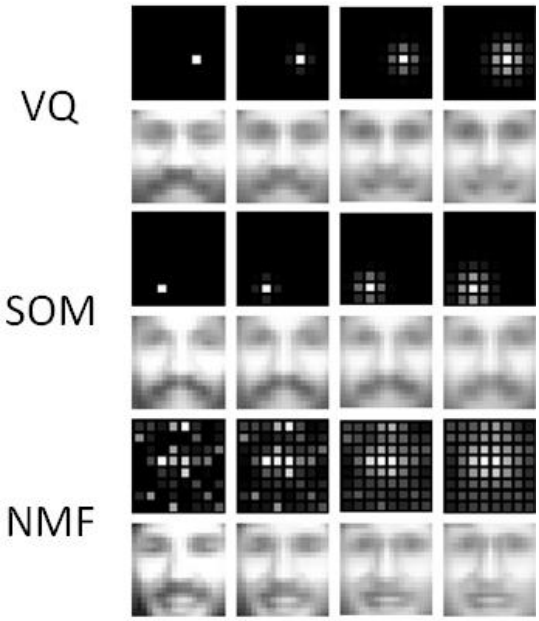

Fig. 7: Node activations and decoded images for an example face encoded by VQ, SOM and NMF. Encodings are shown in order of increasing Gaussian blurring ($\sigma = 0, 0.5, 0.75$ and 1.0). MSE of decoded images:

VQ	**(0.0129,**	**0.0139,**	**0.0186,**	**0.0215)**
SOM	**(0.0099,**	**0.0115,**	**0.0159,**	**0.0198)**
NMF	**(0.0025,**	**0.0066,**	**0.0176,**	**0.0241)**

As shown in fig. 7, NMF yielded the lowest mean squared error at low levels of blurring ($\sigma = 0$ and $\sigma = 0.5$). This is due to the higher expressive power of NMF, which results from its flexible parts-based representation. It is clear that as the variance of the Gaussian filter increases, however, the features of the original image become heavily distorted in both NMF and VQ. This is due to the lack of locality in these encodings. On the contrary, the spatial organization of the SOM encoding helps to minimize the effects of blurring. Since neighboring nodes convey similar information, their weighted contributions do not heavily distort the original signal. As a result, in this example the performance of SOM surpassed NMF at $\sigma = 0.75$ and $\sigma = 1$. In this regard, we predict that a Self-Organized NMF encoding will outperform both SOM and NMF.

To test the generalization capabilities of the various encodings, we trained VQ and NMF dictionaries using a variable number of training images from the MIT CBCL face database, and tested their performance on an independent set of 500 facial images. Fig. 8 shows the training and testing curves for a VQ dictionary trained using k-means on a set of 250, 500 and 1000 images. The mean squared errors of the decoded images are given for dictionaries ranging in size from 10 to 100 bases. Fig. 9 shows an identical plot for dictionaries trained using NMF. The results of SOM training are not shown, due to its close similarity to VQ.

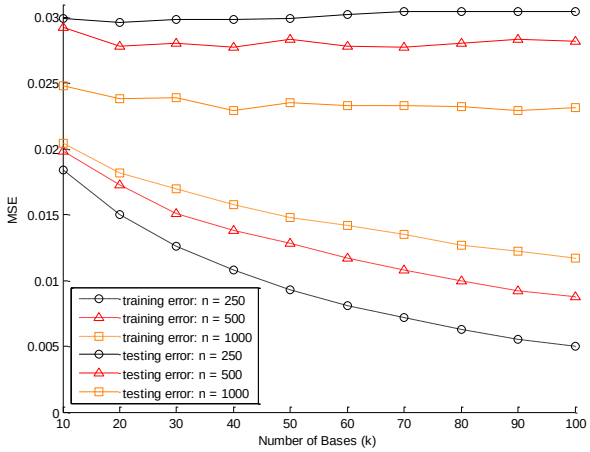

Fig. 8: Training/Testing MSE Curves for VQ

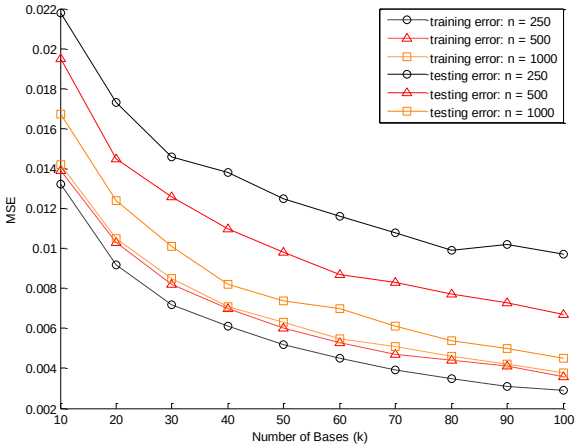

Fig. 9: Training/Testing MSE Curves for NMF

In both VQ and NMF, the testing error decreased with the number of training samples, while the training error increased. The error curves show that, in general, larger dictionaries are better able to approximate the training and testing data. We also note that the approximation errors of the NMF encodings are consistently lower than VQ. The results show that, given a large training set of images, an NMF based encoding should be able to encode a variety of facial images with relatively low approximation error, so long as careful preprocessing is performed by the front end of the system.

IV. Future Work

Future work will focus on the development of a sparse, self-organizing NMF algorithm which can be used to train a dictionary of basis images that maximizes the performance the proposed vision substitution system. Although we have focused on a system that allows the blind to recognize human faces, the target encoding scheme should be general enough to be applied to other constrained image domains as well. For instance, a system that encodes the information from low-resolution occupancy maps could serve as a new type of navigational aid for the blind.

Future research will also investigate means to train the human brain to properly decode the encoded tactile signal. For face recognition, the learning process might be based on visual-tactile association, where the subject feels the structure of different faces while simultaneously observing the corresponding encodings. This may prove problematic, however, if a large number of training faces are required. As an alternative, it may prove sufficient to provide verbal descriptions of example faces. Although the training process offers some unique challenges, we hope that the amazing learning capacity of the human brain will make up for these difficulties. Finally, given that the proposed encoding scheme draws some inspiration from networks within the human vision system, this research may lead to results which are also applicable to field of computational neuroscience and/or biologically inspired machine vision.

REFERENCES

[1] D. Lee and S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788-791, 1999.

[2] T. Kohonen, "The Self-Organizing Map," Proc. IEEE, vol. 78, no. 9, pp. 1464-1480, 1990.

[3] Richard O. Duda, Peter E. Hart, David G. Stork (2001) *Pattern Classification* (2nd edition), Wiley, New York, ISBN 0-471-05669-3

[4] K Hosoda, M Watanabe, H Wersing, E K¨orner, H Tsujino, H Tamura,and I Fujita. A model for learning topographically organized parts-based representations of objects in visual cortex: Topographic nonnegative matrix factorization. Neural Computation, 21(9):2605–2633, 2009.

[5] Daniel D. Lee and H. Sebastian Seung (2001). "Algorithms for Non-negative Matrix Factorization". *Advances in Neural Information Processing Systems 13: Proceedings of the 2000 Conference*. MIT Press. pp. 556–562.

[6] P. O. Hoyer. Non-negative sparse coding. In *Neural Networks for Signal Processing XII (Proc. IEEE Workshop on Neural Networks for Signal Processing)*, pages 557–565, Martigny, Switzerland, 2002.

[7] T. Feng, S. Z. Li, H. Shum and H. J. Zhang, Local non-negative matrix factorization as a visual representation, in *Proc. 2nd Int. Conf. Development and Learning*. Washington DC (June, 2002), pp. 178–183.

[8] P. Bach-y-Rita and S. Kercel, "Sensory substitution and the human machine interface," *Trends in Cognitive Sciences*, vol. 7, pp. 541–546, 2003.

[9] P. Bach-Y-Rita, C. C. Collins, F. A. Saunders, B. White, and L. Scadden, "Vision substitution by tactile image projection," Nature, vol. 221 pp. 963-964, Mar. 8, 1969.

[10] J. C. Bliss, M. H. Katchner, C. H. Rogers, and R. P. Shepard, "Optical to-tactile image conversion for the blind," *IEEE Transactions on Man-Machine Systems*, vol. 11, 1970.

[11] Ptito M, Moesgaard SM, Gjedde A, Kupers R (2005) Cross-modal plasticity revealed by electrotactile stimulation of the tongue in the congenitally blind. Brain 128(Pt 3):606–614

[12] Capelle C, Trullemans C, Arno P, Veraart C. *A real time experimental prototype for enhancement of vision rehabilitation using auditory substitution.* IEEE T Bio-Med Eng 1998*;*45*:* 1279–93*.*

[13] Lenay et al., 2003 C. Lenay, O. Gapenne, S. Hanneton, C. Genouëlle and C. Marque, Sensory substitution: limits and perspectives. In: Y. Hatwell, A. Streri and E. Gentaz, Editors, *Touching for knowing*, John Benjamins, Amsterdam (2003), pp. 275–292.

[14] CBCL Face Database #1, MIT Center For Biological and Computation Learning. http://www.ai.mit.edu/projects/cbcl