Abstract

The use of non-negative matrix factorisation (NMF) on 2D face images has been shown to result in sparse feature vectors that encode for local patches on the face, and thus provides a statistically justified approach to learning parts from wholes. However successful on 2D images, the method has so far not been extended to 3D images. The main reason for this is that 3D space is a continuum and so it is not apparent how to represent 3D coordinates in a non-negative fashion. This work compares different non-negative representations for spatial coordinates, and demonstrates that not all non-negative representations are suitable. We analyse the representational properties that make NMF a successful method to learn sparse 3D facial features. Using our proposed representation, the factorisation results in sparse and interpretable facial features.

1. Introduction

A key problem in computer vision and pattern recognition is to define a representation for a set of data suitable for the task at hand. Two popular linear methods aim to find the axes along which the data has maximum variance (PCA) or that maximally discriminate between classes of objects (LDA). Applied to face images these result in the well known Eigenfaces [14] and Fisherfaces [2] approaches respectively. The principle behind these methods is that maximisation of the signal should improve the signal-to-noise ratio. As a consequence of the maximisation, feature vectors in such methods exploit the correlation between all variables jointly, and thus result in “holistic” feature vectors, i.e., every basis vector is a function of the whole face. This, however, complicates the interpretation of images as linear combinations of the feature vectors: if we consider a small part of the face it is still represented by the same linear combination (the same coefficients) as the whole image. In other words, there is no distinction between the whole face and its parts. The eyebrows are encoded identically to the nose, the mouth, or the whole face.

In contrast with the holistic feature learning methods above, non-negative matrix factorisation (NMF) learns feature vectors that are naturally sparse, i.e., most of the elements are zero. Furthermore, the nonzero elements in each feature vector group the most strongly correlating variables from the input domain. Of particular interest in this paper is the application of NMF to face images. In this domain, but limited to single channel (e.g. monochromatic) 2D images, the nonzero elements in the feature vectors have been shown to reflect local parts of the face [5, 8, 9, 17]. Following on the example above, this means the eyebrows can be encoded separately from the nose and the mouth. Consequently, the encoding of the whole face is formed from the juxtaposition of the individual parts.

Sparsity of the feature vectors is beneficial in many situations, such as when dealing with occlusion and pose variations. Sparse decomposition, compared to holistic analysis, has also shown to generalise better to faces outside the training data. Of particular interest to this work is the objective methodology of arriving at those sparse features, compared to manually selected regions, and the semantic interpretability of the resultant feature vectors. This suggests that those features somehow reflect specific natural variations of human faces, which is of particular interest to gene expression analysis. Being able to link facial features to genetic markers opens up a wide range of possibilities for biometrics and forensics related applications.

The topic of non-negative matrix factorisation has attracted considerable research efforts. Different error functions have been proposed for different purposes. For example in [5], a spatial term was introduced to force the parts to mark physically coherent regions. This has been used to deal with occlusion and other noise by focussing on well detected regions [10, 17]. Other areas of extending NMF have focussed on embedding discriminative models akin to Fisherfaces [16, 19], and on kernel spaces to learn non-linear basis functions [3, 11, 18]. Also efforts have been made in proving near optimal solutions [1, 15]. One of the keys to
its success is the natural sparsity of the feature vectors, i.e., there are no sparsity inducing terms.

In the field of face images, NMF results in feature vectors that represent interpretable parts of the face, as has been demonstrated in [8] and in the context of dictionary learning in [9]. The feature vectors therefore provide a statistical basis for the segmentation of faces into separate components, in contrast with, for example, methods such as watersheds segmentation [7] and Randomised Cuts [6] that operate on single images. Although the latter methods operate on 3D data, the lack of a statistical basis for the segmentation impedes the extension to other images, and thus does not serve as a good basis for facial analysis. In [12], a simplistic attempt to extend NMF to colour images advocates the application of NMF to each colour channel separately and then fuse the matching scores obtained in each factored channel.

Indeed, for face analysis the use of NMF has only focussed on 2D images, in which the pixel intensities are naturally non-negative. By contrast, the 3D shape space is a continuum with no obvious bounds. One method of note is [4], which provides a statistical method to learn a part based decomposition of 3D face images. The idea therein is that a set of test images is encoded as a linear combination of separate training images, where the mapping is mediated by a set of vectors that mask specific areas (parts) of the face. In turn, the method updates its estimate for the basis that maps from training to test images, followed by updating its estimate for the masks. This is repeated until some stopping condition is met. The method has a few caveats though. It requires initialisation and it is not clear how the final result depends on this. Also, each vertex is encoded as its distance from the mean, which loses a lot of information. It restricts the interpretation of the parts to amounts of displacement without knowing the direction. A last argument brought up by De Smet and Van Gool is that convergence of their method is extremely slow. In contrast, current NMF methods are very fast.

In this work we propose a novel non-negative encoding of the 3D spatial data, so that we can use well established NMF implementations for learning the decomposition. To the authors’ knowledge NMF has not been performed on 3D face data before.

The layout of this paper is as follows: An intuition for the sparsity of NMF is provided in Sec. 2, which is a more theoretical aspect of NMF itself. In Sec. 3 we then focus on the mapping of continuous data (such as 3D face shape) to the non-negative domain. In Sec. 4 we evaluate the different representations and demonstrate significant differences in the results obtained. Section 5 concludes with a short summary of the paper and a view to future directions for part based face analysis. But before we continue, let us first provide a context for learning part based representations from face images in the following section.

1.1. 3D shape representation

A prerequisite for any analysis is that a measurement system is established that allows a consistent extraction of information. Under such a measurement system the topology of the facial surface is fixed so that variation in the j-th point, \( v_j \), measures the variation of that specific point in all images. Thus we can write each image in vector form as \( \mathbf{v} = [v_1, v_2, v_3, \ldots, v_n]^T \), which allows us to write all \( m \) images together in a matrix as

\[
\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3, \ldots, \mathbf{v}_m]
\]

Similar to PCA, NMF is a matrix factorisation method which aims to write the data matrix, \( \mathbf{V} \), as a product of two smaller matrices, \( \mathbf{W} \) and \( \mathbf{H} \):

\[
\mathbf{V} = \mathbf{WH}
\]

of size \((n \times m) = (n \times k)(k \times m)\), where each image, \( \mathbf{v}_i = \mathbf{Wh}_i \), is written as the linear combination of the columns of \( \mathbf{W} \)—the feature vectors—weighted by the coefficients in the associated column, \( \mathbf{h}_i \), in \( \mathbf{H} \). In case of PCA, the constraint on this product is that the columns of \( \mathbf{W} \) are orthogonal and point in the direction of maximum (remaining) variance. For NMF the only constraint is that none of the matrices contain negative elements. In Sec. 2 we will provide an intuition for why this constraint induces sparsity on both \( \mathbf{W} \) and \( \mathbf{H} \). The choice of \( k \) is a subject of active research. Formally, non-negative factorisation can only be exact if \( k \geq \text{rank}_+(\mathbf{V}) \), i.e., when \( k \) is chosen greater than or equal to the positive rank of \( \mathbf{V} \), determination of which has been shown to be NP-hard [15]. In practice though, \( k \) is usually chosen much smaller and depends on the intended purpose. In this work we experiment with a few different values, ranging from 10 to 35 to show that the results do not critically depend on a precise value.

However, the main contribution of this paper is the definition of a suitable non-negative encoding of \( \mathbf{V} \). Whereas the pixel values in 2D images are commonly in the range of \((0, 255)\), or in some cases on the unit interval, the vertex positions of 3D face images are in Cartesian coordinates and have no obvious bounds. Another challenge is that the coordinate of a vertex consists of 3 values, \((x, y, z)\), and thus spans three rows in \( \mathbf{V} \) and in \( \mathbf{W} \). Correlation between these three values will be reflected in non-zero values in the feature vectors, and thus decrease sparsity.

2. Sparsity of NMF Bases

In PCA the orthogonality constraint ensures that the maximum amount of variance is captured in the least number of basis vectors. For two orthogonal vectors \( \mathbf{a} \) and \( \mathbf{b} \)
it holds that their dot product \( \mathbf{a} \cdot \mathbf{b} = \sum_i a_i b_i = 0 \). Consequently, if either \( \mathbf{a} \) or \( \mathbf{b} \) is strictly positive, then the other must have negative elements. However, the definition of NMF disallows negative elements in the basis vectors, leaving two options: if we insist on the orthogonality then for any \( i \), if \( a_i > 0 \) then \( b_i \) must be zero. In this case sparsity is directly apparent. Also note that \( \mathbf{a} \) and \( \mathbf{b} \) are forced “out” to lie on the quadrant’s bounding surfaces.

The alternative option is that \( \mathbf{a} \) and \( \mathbf{b} \) are not orthogonal. This reduces sparsity of the basis vectors in exchange for sparsity of the data: As the basis vectors contract, forming a tighter hull around the data, the data coefficients in the projected space will be closer to zero. Notice that any data outside the convex cone is counted towards the error (see Fig. 1). Minimising the error thus involves finding a suitable balance between sparsity of the basis and sparsity of the data.

3. NMF on 3D Registered Images

As described in Sec. 1, the registration of 3D images to a face model allows us to write all images consistently in vector form, and thus to write the whole data set of \( m \) face images as a matrix \( \mathbf{V} \in \mathbb{R}^{n \times m} \). It stores the Cartesian coordinates of all \( n/3 \) vertices \( (x, y, z) \) measured on \( m \) face images. Hence, \( \mathbf{V} \) may contain positive as well as negative elements. A function, \( f: \mathbb{R}^{(n \times m)} \to \mathbb{R}^+(n \times m) \), can be defined that maps \( \mathbf{V} \) to the non-negative domain. The following sections present two alternative functions for this purpose.

3.1. Subtract the minimum

In analogy to PCA, in which the mean is subtracted from all images, an obvious mapping to qualify for NMF is perhaps subtraction of the minimum. More formally, let us denote by \( \mathbf{V}_{ij} \) the cell value in \( \mathbf{V} \) at the \( i \)-th row and \( j \)-th column. Then the “minimum image” can be defined as

\[
\nu = [\nu_1, \nu_2, \nu_3, \ldots, \nu_n]^T
\]

where

\[
\nu_i = \min_j \mathbf{V}_{ij}
\]

and so

\[
f(\mathbf{V}) = [\mathbf{v}_1 - \nu, \mathbf{v}_2 - \nu, \mathbf{v}_3 - \nu, \ldots, \mathbf{v}_m - \nu]
\]

removes any negative elements from \( \mathbf{V} \). Note that replacing min in (4) by the mean centres the data as done in PCA.

3.2. Vertex shape space

The following approach, which we shall term the “vertex shape space”, considers the spatial distribution of each vertex separately. In \( \mathbf{V} \) the position of all vertices is expressed in the same shared \( (x, y, z) \) coordinate system. However, when we consider the spatial distribution of one specific vertex across all images, we can, for this vertex, rotate the coordinate system to align with the directions of maximum variance. This has several advantages. The principal axis captures the most significant shape variation, and subsequent axes capture smaller variations. Also, this decorrelates the distribution of data along the axes which (generally) results in only one dimension receiving non-zero weights in NMF, making it easier to use the coefficients directly as weights on the vertices.

To this end, let us define the matrix \( \mathbf{M}^i \in \mathbb{R}^{3 \times m} \) formed of the rows in \( \mathbf{V} \) that describe the \( x, y \), and \( z \) coordinates of the \( i \)-th vertex. The modes of variation of \( \mathbf{M}^i \) reveal the primary directions in which that small bit of the face varies. This can be computed by a principal components decomposition,

\[
\Sigma^i \mathbf{T}^i = \mathbf{T}^i \Lambda^i
\]

where

\[
\Sigma^i = \frac{1}{m-1} (\mathbf{M}^i)^T \mathbf{M}^i
\]

is the \( 3 \times 3 \) covariance matrix with eigenvalues \( \lambda_1^i, \lambda_2^i, \lambda_3^i \) along the diagonal of \( \Lambda^i \) and eigenvectors \( \mathbf{T}^i = [t_{1i}, t_{2i}, t_{3i}] \) ordered in decreasing value of \( \lambda^i \). The eigenvectors correspond to directions in Euclidian space (the 3D space of images) and represent the basis directions of shape variation.
The three images together visualise $\lambda^i$ for all vertices $i$ in the face. From left to right the images correspond to the first, second and third eigenvalues. The grey scale ranges between 0 and 1 and is computed as $\lambda^i_j / \sum \lambda^i$, with $j = 1 \ldots 3$. The level of grey around the cheeks in the second image appearing brighter than in the first is an illusory effect: these values lie around 0.42 with the smallest difference between them 0.0045. (The perception of depth from shading is also an optical illusion by the way, as the tint solely reflects eigenvalues.)

on a vertex level (independent on its neighbourhood). The eigenvalues represent the spread of the points along each axis, or more precisely their variance.

Figure 2 visualises the eigenvalues relative to their sum at each individual vertex. Interestingly, the amount of variance in regions around the cheeks appears to differ very little between the first and second principal components.

The columns of $T^i$ form an orthonormal basis for the $i$-th vertex, which permits the projection of Euclidean vertex coordinates into coordinates specific to that vertex, $L^i = (T^i)^T M^i$ (7)

where $L^i$ is of size $3 \times n$. Now we can define the mapping from $V$ to vertex shape space as follows

$$g(V) = \left| (L^1)^T, (L^2)^T, (L^3)^T, \ldots, (L^{n/3})^T \right|^T$$ (8)

where $|$ denotes the absolute of its argument. In words: in each vertex, $x$, $y$, and $z$ are decorrelated by projection onto the principal components (this also centres the data on the origin). In the decorrelated space, we take the absolute of each variable.

It should be noted that, along the axis of maximum variation, normally distributed data is symmetric. The operation of taking the absolute, therefore, does not alter the distribution in the positive octant. It only increases the number of samples. On top of that, one should note that the sign of PCA axes is not fixed. This is particularly relevant when considering the variation between two vertices, as it means that no meaningful distinction can be made between positive and negative correlation. Provided that two variables correlate, but with unknown sign, taking their absolute values will cause them to correlate with positive sign.

An additional benefit of the decorrelation is that, because NMF feature vectors ($w_j \in W$, see (2)) combine correlated variables, it is expected that the non-negative coefficients in each $w_j$ select at most one of the three variables per vertex. Hence, the feature vectors will be sparser.

Further to that, it is a common principle of PCA to select the first few components that explain a sufficient level of variance. One could thus, rather than concatenating the full $L^i$, select the first one or two rows. A visual representation of the amount of variance in the first, second, and third variable of all vertices is given in Fig. 2. Such a selection could be called the “vertex-reduced shape space”.

4. Experiments

In this section we compare the basis $W$ obtained by the non-negative matrix factorisation $V' = WH$, where $V'$ is the non-negative encoding of data matrix $V$ using either $f$ or $g$ as defined in (5) and (8) respectively. We set the number of feature vectors for this factorisation to $k = 35$. The algorithm used for NMF is described in [9]. It is a stochastic implementation that minimises the sum of squares of the residual matrix $U = WH - V'$. For each factorisation, the algorithm is run for 60 seconds which equates to roughly 40 iterations.

The data set comprises 3D images of 1230 people (male and female) taken with a 3dMD Face camera system. Fourteen fiducial points have been marked on each face prior to dense registration with the Morphable Model using the algorithm described in [13]. The dense registration samples

---

1This data set was collected by a third party with the aim of establishing a link between DNA and face shape. The data collection agreement with the donors does not provision for a stand alone use of the face data for other biometric studies, which, unfortunately, means that the 3D face database is not publicly available.
each face over 29,658 vertices, from which we selected a subset of 16,656 vertices based on an assessment of their accuracy on a separate data set. The resulting face shapes, and their mirror images (reflecting horizontally), were included in V.

4.1. Learning parts

The question is, what are good qualities of the basis vectors, \( w_i \in W \)? Visually speaking, we aim to select local parts of the face. A sensible criterion therefore would be to see which feature vectors describe more locally constrained segments. Figures 3 and 4 visualise the coefficients in the first twenty feature vectors, \( w_i, 1 \leq i \leq 20 \), for \( f \) and \( g \) respectively. It is clearly observable that the latter method (the vertex shape space representation) produces more clearly and locally defined parts.

Although mapping \( g \) produces visually more pleasing results, we wish to provide an objective assessment of the difference the two methods. To this end we measure the sparsity, \( s \), of \( W \), expressed as the number of zeros in \( W \) relative to its size. That is,

\[
s(W) = \frac{1}{nk} \sum_{i=1}^{n} \sum_{j=1}^{k} \delta(W_{ij})
\]  

(9)

where \( \delta \) is the Dirac delta function, which is 1 if and only if its argument is zero, and is zero otherwise. It is worth noting that this function does not provide a measure of “locality” of the non-zero elements. This is not a problem since the spatial assessment of the non-zero elements is much easier done visually than through some numerical metric. A second point of note is that the non-zero elements in \( W \) cover all variables. In other words, \( W \) contains at least \( n \).
non-zero elements. For \( k = 35 \), the maximum sparsity is thus \( 1 - \frac{1}{k} = 0.97 \), which implies an exact segmentation of the facial surface.

Table 1 lists the sparsity values for the mappings \( f \) and \( g \), showing that the vertex shape space does result in sparser NMF feature vectors.

### 4.2. Face segmentation

The sparsity in the NMF feature vectors leads naturally to a part-based representation of the face. Despite the sparsity, it is clear from Sec. 4.1 that there is still a significant amount of overlap between the feature vectors. In some situation, however, an exact segmentation is desired. In this section we describe how to construct a hard segmentation and evaluate the influence of the choice for \( k \), the number of feature vectors.

For the sake of simplicity we will assume for now that each vertex is encoded by only one variable. We will extend this to more than one variable per vertex below.

To suppress the lowest values in \( W \), one could define a threshold \( \theta \) and set all values below \( \theta \) to zero. However, the choice for this parameter would be arbitrary and it may even lead to the exclusion of some vertices entirely. Instead, we will assign each vertex to one of the \( k \) components, by selecting the vector \( w_i \) for which it has maximum value. That is, we update \( W \) so that

\[
W_{ij} \left\{ \begin{array}{ll} W_{ij} & \text{if } W_{ij} = \max_j W_{ij} \\ 0 & \text{otherwise} \end{array} \right. \tag{10}
\]

The result is a complete partitioning of the facial surface in disjoint parts.

We will now extend this to multiple variables per vertex. As described in Sec. 3.2, the vertex space can be reduced to one or two dimensions, or kept at three. In each case the axes are orthogonal and therefore measure shape variance in linearly independent directions. However, between neighbouring vertices the orientation of bases is expected to differ slightly, especially on curved parts of the surface, and is not independent. It is not uncommon to find the correlation between the first PC in one vertex and the second PC in a neighbouring vertex to exceed that between both first PCs. Allowing more variables per vertex enables us to extend the otherwise hard boundaries between parts to add some overlap between parts.

Formally, let us generalise our previous definition of \( L^i \) (7) so that it is of size \( d \times m \), with \( 1 \leq d \leq 3 \). Then \( W \) is \( dn \times m \). Since Eq. 10 assigns each row to one component (column), a vertex can be assigned to \( d \) different components. Finding a vertex assigned to the same component multiple times is possible but rare, due to the decorrelation of the basis vectors \( T^i \) (6). If it is required that each part assigns one weight per vertex, instead of \( d \), one can simply take the maximum.

Figure 5 shows the segmentation of \( W \) (10) for different values of \( k \), and with \( d = 2 \). Looking at the first row, i.e., the projections on \( t^1 \), we observe that not all regions of the face are affected similarly by a change of \( k \). For example, the region on the chin remains fairly consistent, as does the nose tip. By contrast the regions around the cheeks break into smaller parts as \( k \) increases. Also the eyebrows are represented as a separate part at higher values of \( k \).

The second row in Fig. 5 shows the segmentation of the face in the projection on \( t^2 \). The results are much less consistent. The idea that hard boundaries of a single variable encoding (\( d = 1 \)) can be expanded by considering the second variable does not appear to hold very well as same-colour segments between the rows do not generally appear to lie close to each other on the face. Values of \( d > 1 \) seem to be better suited when dealing with the parts directly, or for soft-segmentation, rather than for hard segmentation, although the stark differences could also be an indicator that NMF has not fully converged to the right solution.

### 5. Conclusions and Future Work

The aim of this study was to find a suitable non-negative encoding of spatial data, which is naturally unbounded and contains negative values. We have demonstrated a significant dependence between this encoding and the patterns learnt by NMF. Better non-negative mappings result in more local and also sparser feature vectors, similar to the results reported on 2D face images. At the same time the sparser features are more easily interpreted.

We have also evaluated hard segmentation derived from the “soft” feature vectors. The results look particularly promising when using one variable per vertex. Segmentation of the face appears to reflect how people describe faces (e.g. nose, cheeks, eyes, mouth, chin). Increasing the number of parts obviously breaks some parts down, resulting in more smaller features. Interestingly though, this does not happen in equal amounts throughout the face. In particular the chin and nose tip seem to be very consistent segments.

As mentioned at the start of this paper, a part-based representation provides a much more natural description of the face than a holistic representation. In that light, equipped with an objective definition of parts of the face, based on variance measured over a large set of face images, we hope in the future to create a part based face shape model. We
will have to investigate how different parts can be combined, or perhaps from a top-down approach, how parts are exactly split when \( k \) increases.

References


