Introducing a weighted non-negative matrix factorization for image classification

D. Guillamet a,*, J. Vitrià a, B. Schiele b

a Centre de Visió per Computador (CVC), Departament d’Informàtica, Edifici O. Catalunya, Universitat Autònoma de Barcelona (UAB), 08193 Bellaterra, Barcelona, Spain
b Perceptual Computing and Computer Vision Group, Computer Science Department, ETH Zurich, Switzerland

Received 18 September 2002; received in revised form 20 February 2003

Abstract

Non-negative matrix factorization (NMF) technique has been recently proposed for dimensionality reduction. NMF is capable to produce region or part based representations of objects and images. Also, a direct modification of NMF, the weighted non-negative matrix factorization (WNMF) has also been introduced to improve the NMF capabilities of representing positive local data (as color histograms). A comparison between NMF, WNMF and the well-known principal component analysis (PCA) in the context of image patch classification has been carried out and it is claimed that all these three techniques can be combined in a common and unique classifier. This contribution is an extension of a previous study and we introduce the use of the WNMF as well as a probabilistic approach to compare all the three techniques noticing a great improvement in the final recognition results.

Keywords: Non-negative matrix factorization (NMF); Principal component analysis (PCA); Color histogram classification

1. Introduction

Principal component analysis (PCA) is a very popular technique for dimensionality reduction. It is a well-known fact that PCA is optimal in terms of the reconstruction error but not for the separation and recognition of classes. Nevertheless, PCA is often used directly for pattern and object recognition tasks. In the computer vision community for example it has been used for recognition of faces (Turk and Pentland, 1991) and 3D objects (Murase and Nayar, 1995) as well as dealing with partial occlusions by using robust estimation techniques (Black and Jepson, 1998).

Recently, Lee and Seung (1999) proposed a new technique, called Non-negative matrix factorization (NMF), to obtain a reduced representation of data only using positive restrictions. NMF differs from other methods by its use of non-negativity constraints. They demonstrated with a set of face images (Lee and Seung, 1999) that NMF can be used to obtain a set of bases of localized features in an unsupervised way. Many of those localized features...
correspond to the intuitive notion of face parts such as eyes and mouth. Guillamet et al. (2001) presented a study of a weighted version of the original NMF (weighted non-negative matrix factorization, WNMF) where they demonstrated that for local data representations NMF can generate redundant bases when the dimensionality of the feature space is high and they presented a modification of the original technique that minimizes this behaviour. WNMF is also based on the same principles as NMF but with a new weighting matrix that is able to weigh each training feature vector minimizing the presence of possible repeated bases.

A recent study (Guillamet et al., 2002) compares both PCA and NMF techniques in an image patch classification framework where the main claim is that both techniques can be merged in a common classifier. Since both PCA and NMF techniques are of different nature, they cannot be compared in a direct way (Buchsbaum and Bloch, 2002) and is for that reason that a comparison based on the reconstruction error has been carried out in (Guillamet et al., 2002). The final classifier that merges PCA and NMF is based on the L2-norm of the reconstruction error but it remains unclear whether this is the best way to merge both techniques.

This paper presents a comparative study of PCA, NMF and WNMF techniques in the same color classification scheme used in (Guillamet et al., 2002). One of the main goals is to validate that WNMF can also be introduced in a classification framework as an improvement of the original NMF. Additionally, since L2-norm seems not to be an appropriate metric to use in conjunction with NMF (Guillamet and Vitrià, 2002), we introduce a reliable probabilistic framework to compare all techniques. For each method, we define a probabilistic density function that is used to merge all techniques in a simple and unique classifier that outperforms each of them and the L2-norm based results.

2. PCA, NMF and WNMF techniques

2.1. Principal component analysis

Due to the high dimensionality of data, similarity and distance metrics are computationally expensive and some compaction of the original data is needed. PCA is an optimal linear dimensionality reduction scheme with respect to the mean squared error (MSE) of the reconstruction. For a set of \( N \) training vectors \( X = \{x^1, \ldots, x^n\} \) the mean (\( \mu = (1/N) \sum_{i=1}^{N} x^i \)) and covariance matrix (\( \Sigma = (1/N) \sum_{i=1}^{N} (x^i - \mu)(x^i - \mu)^T \)) can be calculated. Defining a projection matrix \( E \) composed of the \( K \) eigenvectors of \( \Sigma \) with highest eigenvalues, the \( K \)-dimensional representation of an original, \( n \)-dimensional vector \( x \), is given by the projection \( y = E^T(x - \mu) \).

Once a PCA subspace (\( \Omega \)) is created with \( K \ll n \) dimensions, an efficient density estimation method can be considered. As explained in (Moghaddam and Pentland, 1997), if we consider only the PCA subspace (\( \Omega \)) description defined by the first \( K \) eigenvectors, the likelihood estimate of a new vector \( \Delta \) can be approximated by a Gaussian density function:

\[
P(\Delta|\Omega) = \left( \frac{\exp \left(-\frac{1}{2} \sum_{i=1}^{K} \frac{\lambda_i^2}{\lambda_i^2} \right)}{(2\pi)^{K/2} \prod_{i=1}^{K} \lambda_i^{1/2} \rho^{n-K/2}} \right) \left( \frac{\exp \left(-\frac{c^2(\Delta)}{2\rho} \right)}{(2\pi\rho)^{(n-K)/2}} \right)
\]

where \( y_i \) are the principal components, \( \lambda_i \) their eigenvalues, \( c^2(\Delta) \) is the residual (or DFFS). The optimal value for the weighting parameter \( \rho \) is found by averaging \( \rho = (1/n - K) \sum_{i=K+1}^{n} \lambda_i \) (see (Moghaddam and Pentland, 1997) for more detailed information).

2.2. Non-negative matrix factorization

NMF is a method to obtain a representation of data using non-negativity constraints. These constraints lead to a part-based representation because they allow only additive, not subtractive, combinations of the original data (Lee and Seung, 1999). Given an initial database expressed by a \( n \times m \) matrix \( V \), where each column is an \( n \)-dimensional non-negative vector of the original database \( m \) vectors, it is possible to find two new matrices \( (W \) and \( H) \) in order to approximate the original matrix \( V_{ij} \approx (WH)_{ij} = \sum_{a} W_{ai} H_{aj} \). The dimensions of the factorized matrices \( W \) and \( H \) are \( n \times r \) and \( r \times m \), respectively. Usually, \( r \) is chosen...
so that \((n + m)r < nm\). Each column of matrix \(W\) contains a basis vector while each column of \(H\) contains the weights needed to approximate the corresponding column in \(V\) using the bases from \(W\).

In order to estimate the factorization matrices, an objective function has to be defined. A possible objective function is given by \(F = \sum_{i=1}^{n} \sum_{\mu=1}^{m} [V_{i\mu} \log(WH)_{i\mu} - (WH)_{i\mu}]\). This objective function can be related to the likelihood of generating the images in \(V\) from the bases \(W\) and encodings \(H\). An iterative approach to reach a local maximum of this objective function is given by the following rules (Lee and Seung, 1999): \(W_{ia} \leftarrow W_{ia} \sum_{\mu} (V_{i\mu}/(WH)_{i\mu})H_{a\mu}, \quad W_{ia} \leftarrow W_{ia}/\sum_{j} W_{ja}, \quad H_{a\mu} \leftarrow H_{a\mu} \sum_{l} W_{il} (V_{i\mu}/(WH)_{i\mu})\). Initialization is performed using positive random initial conditions for matrices \(W\) and \(H\). The convergence of the process is also ensured (see Lee and Seung, 1999, 2000 for more information).

It can be seen that the objective function comes out if we assume that \(V\) is drawn from a Poisson distribution with mean \(WH\). In this case, we can obtain a probability density function that takes the following form:

\[
p(V_{i\mu}|(WH)_{i\mu}) = e^{-(WH)_{i\mu}} \frac{((WH)_{i\mu})^{V_{i\mu}}}{V_{i\mu}!}
\]  

(2)

Thus, once a set of bases \(W\) is correctly described through the above defined iterative process and a set of encodings \(H\) is also determined, it is possible to obtain the probability of new data \((V)\) to belong to this set of bases \(W\).

### 2.3. Weighted non-negative matrix factorization

When using a local representation, similarity between the training vectors can introduce redundancy in the \(W\) bases. This redundancy is manifested by the presence of repeated bases as demonstrated in (Guillamet et al., 2001). A possible solution to this problem is to introduce a weight on each training vector, giving more weight to those vectors with low probability of appearing in the training set. This weighted model can be seen as the result of multiplying both sides of the factorization with a \(m\) by \(m\) diagonal weight matrix \(Q\) and to estimate the bases and encodings for the new factorization model, \(VQ \approx WHQ\). Where the diagonal element \(q_{\mu}\) corresponds to the weight of training vector \(\mu\), with \(1 \leq \mu \leq m\). It is also assumed that all the weights sum to unity. The modified objective function in this case takes the form \(F_Q = \sum_{\mu=1}^{m} q_{\mu} \sum_{i=1}^{n} [V_{i\mu} \log((WH)_{i\mu}q_{\mu}) - (WH)_{i\mu}].\) Now, the iterative update rules to obtain the new matrices subject to this new objective function are defined by: \(W_{ia} \leftarrow (W_{ia}/\sum_{\mu} q_{\mu} H_{a\mu}) \times \sum_{\mu} (q_{\mu} V_{i\mu}/(WH)_{i\mu}) H_{a\mu}, \quad W_{ia} \leftarrow W_{ia}/\sum_{j} W_{ja}, \quad H_{a\mu} \leftarrow H_{a\mu} \sum_{l} W_{il} (V_{i\mu}/(WH)_{i\mu}).\)

As global NMF finds out some redundant bases corresponding to the most frequent training vectors when applied to local data, a good choice to define the weighted matrix \(Q\) is giving more weight to the less frequent training vectors. By obtaining the probability of each training vector with respect to the training database and assuming that this probability will hold on the test stage, we invert these probabilities and we take them as the \(q_{\mu}\) coefficients. In this way, we equalize the importance of the training vector classes. Thus, the obtained bases will contain a wide variety of features: from the most to the least frequent ones, improving the global NMF capacity of representation that only takes into account the most frequent ones.

The introduction of a weighting into NMF does not overfit spurious patterns. One problem of NMF is that we obtain some redundancy in the \(W\) bases that comes out from the most repeated patterns in the training set. If we introduce some weight to the training vectors, we reduce this redundancy by adding new patterns to the obtained representation. But, it is impossible to introduce spurious patterns because we are also minimizing an objective function that takes into account the global behaviour of the method. Thus, the iterative process guarantees that spurious patterns do not appear in the \(W\) bases.

Since WNMF is a direct modification of the original NMF, the probability density function takes the same form as the one shown in expression (2). Since WNMF is a constrained scheme of NMF, once WNMF arrives to a stable solution, it can be iterated again using the original NMF. As described in (Guillamet et al., 2001), this final iteration process provides a better solution for the
3. Experimental results

These experimental results are the continuation of the ones described in (Guillamet et al., 2002) with the introduction of the WNMF technique and a probabilistic approach to compare all methods. In (Guillamet et al., 2002), we selected 932 color images from the Corel Image database. These images were selected in order to obtain data for 10 different classes of image patches namely: Clouds, Grass, Ice, Leaves, Rocky Mountains, Sand, Sky, Snow mountains, Trees and Water. Image patches from each of these classes contain different color tonalities rather than one unique and global color. Each image is automatically divided in $10 \times 10$ local regions or image patches of 3456 (48 $\times$ 72) RGB pixels. Whenever possible, each of these image patches is labeled according to the 10 classes mentioned above. Each image patch is represented using a color histogram of 512 dimensions (8 bins per color). Thousand of those are randomly chosen per class for training and another 1000 for testing. Experiments are based on using local color histograms since we are evaluating two techniques (NMF and WNMF) that work with positive representations.

3.1. PCA, NMF and WNMF models

In our previous work (Guillamet et al., 2002), the way to compare two different models was to take into account the reconstruction error of both PCA and NMF models using the $L_2$-norm as a metric. In a PCA context, this means to consider the residual error $e^2(A)$ considered in expression (1) but the reconstruction error obtained using NMF was an heuristic way. It is not clear that the reconstruction error obtained using the $L_2$-norm in the NMF context is a completely justified and reliable framework. Introducing a probabilistic approach for both techniques (PCA and NMF) it is clear that both methods can be compared directly because each probability density function is obtained according to each technique nature. Table 1 reflects the confusion matrices obtained when using PCA, NMF and WNMF with a probability density function. A first improvement is the introduction of a probabilistic framework because both PCA and NMF results are improved with respect to our previous work (Guillamet et al., 2002). PCA is slightly improved from an initial classification of 56.43% to a 57.41% but NMF is considerably improved because the initial recognition rate of 59.89% has changed to 62.11%. Also, WNMF is, at the same time, slightly better than NMF. From the confusion matrices of Table 1 we can deduce that each technique is suitable for some specific classes and this is the main interesting feature to consider because we can combine all the three techniques in order to build an unique classifier.

3.2. Combination of an unique classifier

As in our previous work (Guillamet et al., 2002), combination of models can be a good solution in order to take advantage of each positive aspect of our three techniques (PCA, NMF and WNMF). The central idea is that each class would be represented for the best technique and, as explained in (Guillamet et al., 2002), the way to select the best method for each class would be based on the matrices defined in Table 1 and considering expression (3). Table 2 contains a set of $\alpha_{CLASS}$ per each class and technique. The best technique that would be used for each class is the one that is in bold face.

$$
\alpha_{CLASS} = \frac{\text{Correct classification vectors of CLASS}}{\text{No. of vectors of other classes classified as CLASS}}
$$

(3)

Table 2 reveals that nearly all the classes that were selected in our previous work (Guillamet et al., 2002) to be represented with PCA are also selected in this experiment to be represented with PCA. In our previous work, classes Clouds, Leaves, Rocky, Snow, Tree and Water were represented using PCA. But, since WNMF is introduced as an improvement of the original NMF, nearly all the classes that were represented using NMF in our
previous work turned to be represented with WNMF. Finally, we present in Table 3 the confusion matrix of this combined classifier where classes Clouds, Leaves, Rocky Mountains, Snow
and Trees are represented using PCA, classes Sky and Water using NMF, and Grass, Ice and Sand using WNMF.

### 3.3. Hierarchical classifier

Again, as in our previous work, a hierarchical division of classes would be the best solution to our problem since we can classify our data using specific classifiers and take advantage of each of our three techniques at each level of division. It is natural to think that we can separate blue classes against other classes at the first level of our hierarchical classifier and use the best technique to represent each class. This classification scheme is useful because we can use specific and specialized classifiers at different levels of our division. The best distribution of classes and techniques for such a classification scheme is described in Fig. 1. Given a node in this new tree representation and having a local vector to classify between the right and left leaves, we obtain a probability measure for each technique and we choose the leaf that containing the highest probability. Under the name of each class in Fig. 1, the employed technique for each discrimination task is shown. Finally, considering this optimal hierarchical distribution of models, we obtain the results presented in Table 4 noticing a significant improvement of the recognition rate of 70.23%.

![Fig. 1. Optimal hierarchical representation to classify the initial 10 classes of data.](image-url)

<table>
<thead>
<tr>
<th>Model</th>
<th>Clouds</th>
<th>Grass</th>
<th>Ice</th>
<th>Leaves</th>
<th>Rocky</th>
<th>Sand</th>
<th>Sky</th>
<th>Snow</th>
<th>Tree</th>
<th>Water</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clouds</td>
<td>517</td>
<td>1</td>
<td>169</td>
<td>0</td>
<td>8</td>
<td>25</td>
<td>175</td>
<td>57</td>
<td>1</td>
<td>47</td>
</tr>
<tr>
<td>Grass</td>
<td>0</td>
<td>825</td>
<td>0</td>
<td>53</td>
<td>19</td>
<td>15</td>
<td>0</td>
<td>0</td>
<td>86</td>
<td>2</td>
</tr>
<tr>
<td>Ice</td>
<td>53</td>
<td>0</td>
<td>707</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>100</td>
<td>41</td>
<td>4</td>
<td>93</td>
</tr>
<tr>
<td>Leaves</td>
<td>2</td>
<td>73</td>
<td>7</td>
<td>849</td>
<td>7</td>
<td>6</td>
<td>0</td>
<td>1</td>
<td>36</td>
<td>19</td>
</tr>
<tr>
<td>Rocky</td>
<td>37</td>
<td>17</td>
<td>9</td>
<td>0</td>
<td>476</td>
<td>186</td>
<td>28</td>
<td>35</td>
<td>167</td>
<td>45</td>
</tr>
<tr>
<td>Sand</td>
<td>55</td>
<td>7</td>
<td>10</td>
<td>0</td>
<td>87</td>
<td>801</td>
<td>1</td>
<td>2</td>
<td>29</td>
<td>8</td>
</tr>
<tr>
<td>Sky</td>
<td>57</td>
<td>3</td>
<td>236</td>
<td>0</td>
<td>1</td>
<td>17</td>
<td>582</td>
<td>46</td>
<td>1</td>
<td>57</td>
</tr>
<tr>
<td>Snow</td>
<td>198</td>
<td>0</td>
<td>242</td>
<td>0</td>
<td>13</td>
<td>6</td>
<td>42</td>
<td>370</td>
<td>6</td>
<td>123</td>
</tr>
<tr>
<td>Tree</td>
<td>2</td>
<td>39</td>
<td>1</td>
<td>24</td>
<td>15</td>
<td>7</td>
<td>0</td>
<td>2</td>
<td>890</td>
<td>20</td>
</tr>
<tr>
<td>Water</td>
<td>43</td>
<td>4</td>
<td>149</td>
<td>2</td>
<td>15</td>
<td>12</td>
<td>34</td>
<td>60</td>
<td>25</td>
<td>656</td>
</tr>
</tbody>
</table>

Total recognition rate: 66.73%
4. Conclusions

In this paper we have experimentally analyzed an alternative technique to PCA, the so-called NMF. Also, since our data is locally defined, we have introduced a direct modification of the NMF, the WNMF, in order to improve our capabilities of representation. In this study, we have demonstrated that a probability based comparison between our techniques is better than considering a $L_2$-norm, as in our previous study (Guillamet et al., 2002). Improvement of introducing this probability based comparison is remarkable when we consider the NMF or WNMF techniques meaning that the $L_2$-norm is a bad metric to use with them. This study also demonstrates that the introduction of the WNMF is important because is able to classify some classes better than NMF and classification results are, in a general point of view, also better. Finally, a combination of PCA, NMF and WNMF techniques in a combined classifier and in a hierarchical classification tree are introduced noticing a great improvement with respect to the results obtained in our previous work meaning that the introduction of WNMF and the probabilistic approach has been successful.

PCA, NMF and WNMF are techniques not suited for classification tasks but it is important to think that if we have different data classes, each of them would be represented better by some technique. This motivates to use a huge number of techniques and combine all of them in a combined classifier taking advantage of their positive aspects but we have to find a way to mix all of them. In this particular study, connection between these three techniques is the probability framework.

In this current study, we worked with local color histograms that are positive descriptions of data. PCA is optimal in terms of the MSE of the reconstruction and NMF/WNMF are two techniques that are able to deal with positive data. Since our problem is divided in 10 different data classes and each class described by a particular set of behaviour defined by its colors, NMF/WNMF are adapted to this problem because $W$ bases can represent these behaviours being able to generalize the results. But it remains an important and interesting question: when and why PCA, NMF or WNMF should be selected for representing a certain class for classification purposes?

References


