A Handcrafted Normalized-Convolution Network for Texture Classification
Vu-Lam Nguyen, Ngoc-Son Vu, Philippe-Henri Gosselin

To cite this version:
Vu-Lam Nguyen, Ngoc-Son Vu, Philippe-Henri Gosselin. A Handcrafted Normalized-Convolution Network for Texture Classification. Compact and Efficient Feature Representation and Learning in Computer Vision, Oct 2017, Venice, Italy. hal-01681182

HAL Id: hal-01681182
https://hal.archives-ouvertes.fr/hal-01681182
Submitted on 11 Jan 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
A Handcrafted Normalized-Convolution Network for Texture Classification

Vu-Lam Nguyen
lam.nguyen@ensea.fr
Ngoc-Son Vu
son.vu@ensea.fr
Philippe-Henri Gosselin
philippe-henri.gosselin@ensea.fr
ETIS/Université Paris Seine, Université Cergy-Pontoise, ENSEA, CNRS
95000-Cergy, France
Tel.: +33 01 30 73 66 10
Fax: +33 01 30 73 66 27

Abstract

In this paper, we propose a Handcrafted Normalized-Convolution Network (NmzNet) for efficient texture classification. NmzNet is implemented by a three-layer normalized convolution network, which computes successive normalized convolution with a predefined filter bank (Gabor filter bank) and modulus non-linearities. Coefficients from different layers are aggregated by Fisher Vector aggregation to form the final discriminative features. The results of experimental evaluation on three texture datasets UIUC, KTH-TIPS-2a, and KTH-TIPS-2b indicate that our proposed approach achieves the good classification rate compared with other handcrafted methods. The results additionally indicate that only a marginal difference exists between the best classification rate of recent frontiers CNN and that of the proposed method on the experimented datasets.

1. Introduction

Texture provides important clues for identifying materials and objects, especially when their shapes are not available. A wide range of applications such as industrial inspection, image retrieval, medical imaging, remote sensing, object and facial recognition can be developed depend upon analyzing the textures of their surfaces. Hence, texture analysis includes segmentation, shape extraction, synthesis, and classification is an active field.

There are variety texture analysis approaches have been proposed. They can be ranged from a simple to sophisticated computation strategy methods. Simple yet efficient feature extraction approaches can be a long list. They can be: 1) local binary pattern (LBP) method [32] and its variants; 2) the representation based on co-occurrence matrix [16]; 3) the filter-based approaches such as works in [2, 30]; 4) the wavelet transform method as works in [28, 20, 4]; 5) the textron dictionary-based [22, 43, 44]; 6) the use of bidirectional features [8, 9]; and so on. In addition, many sophisticated computation strategy methods have been introduced to improve the feature robustness and performance. Scattering Network (ScatNet) [41], and Convolutional neural network (CNNs) with Fisher vector CNN (FV+CNN) delegation [6] belong to this category.

Among those approaches, the LBP-family can be considered as a popular feature method which extracts well local micro-structure information from images. Ojala et al. [31] first introduced the LBP method in 1996, then a multi-resolution version [32] in 2002. After that, several extensions on LBP have been conducted. In 2007, Tan et al. extended LBP to three-valued codes to become the local ternary pattern (LTP) [42]. Liao et al. proposed dominant LBP (DLBP) [23] which combines the most frequently occurred patterns with the Gabor filter responses for features. Later Guo et al. introduced completed LBP (CLBP) [12], which merges three components the sign (CLBP_S), magnitude (CLBP_M), and center pixel intensity (CLBP_C) together to form features. This enhances discriminative power compared to the original version. Variance in LBP (LBPV) [13] is utilized to encode local contrast information without requiring a quantization process, and rotation invariance is implemented by estimating principal orientations and aligning LBP histograms. By constructing a cascading spatial pyramid of LBP, Qian et al. [37] introduced pyramid transformed LBP (PLBP), the robustness of PLBP was compared with those of other LBP variants in this works. Further, Liu et al. suggested extended LBP [26] by a combination of pixel intensities and local differences. In this way, the pixel intensity part is divided into a central pixel’s component and neighbor’s component. Likewise, the local difference consists of two components: radial differences and an an-
gular difference. At the end, those four were combined to form features. In addition, Zhao et al. in [51] presented local binary pattern histograms of Fourier features (LBP-HF) which implement rotation invariance by computing discrete Fourier transforms of LBP histograms. In [11], moreover, Guo et al. presented a three-layered learning framework in which LTP and CLBP were used as raw features to train and select the discriminative features.

Contrary to the micro-structure descriptors of LBP family, several broader range feature methods have been developed. Bovik et al. applied the Gabor filters to compute the average filter responses for features [2]. Mallat proposed the multi-resolution wavelet decomposition method [29], which generates coefficients from the high-low (HL), low-high (LH), and low-low (LL) channels for subsequent classification tasks. Porter et al. [34] removed the high-high (HH) wavelet channels and combined the LH and HL wavelet channels to obtain rotation invariance wavelet features. Haley et al. [15] calculated isotropic rotation invariance features from Gabor filter responses. More recent, scattering transform is considered as a high-performance approach based on cascading wavelet transform layers [41] compared to previous wavelet-based methods.

The paper is organized as follows: we start with a review of the normalized convolution in Section 2.1, Fisher vector aggregation presented in Section 2.2, followed by details of the derivation of the proposed approach (Section 2.3). Section 2.4 presents the proposed method. In Section 3, we verify the approach with experiments on popular texture datasets and comparisons with various state-of-the-art texture classification techniques. Section 4 provides concluding remarks and possible extensions of the proposed method.

2. Proposed Handcrafted Network

In this section, we propose a handcrafted Normalized-Convolution Network with its derivation described in 2.1, 2.1, and 2.2 for efficient texture classification.

2.1. Normalized Convolution

Normalized convolution was introduced by Knutsson and Westin in [19]. It is a method for performing general convolution operations on data of signal/certainty type. More specifically, it takes into account the uncertainties in signal values and allows spatial localization of possibly unlimited analysis functions. The conceptual basis for the method is the signal/certainty philosophy, i.e. separating the values of a signal from the certainty of the measurements. The separation of both data and operator into a signal part and a certainty part. Missing data is simply handled by setting the certainty to zero. In the case of uncertain data, an estimate of the certainty must accompany the data. Localization or windowing of operators is done using an applicability function, the operator equivalent to certainty, not by changing the actual operator coefficients. Spatially or temporally limited operators are handled by setting the applicability function to zero outside the window.

Formally, Knutsson et al. define normalized convolution in form of standard convolution as follows,

\[ U(\xi) = \sum a(x)B(x) \odot c(\xi - x)T(\xi - x) \]  

where \( \xi \) is the global spatial coordinate. \( x \) is the local spatial coordinate. \( T(\xi) \) is a tensor corresponding to the input signal. \( c(\xi) \) is a function which represents the certainty of \( T(\xi) \). \( B(x) \) is a tensor corresponding to the operator filter basis. \( a(x) \) is the operator equivalent to certainty, a function which represents the applicability of \( B(x) \).

\( \odot \) denotes multi-linear operations (in standard convolution this operation is scalar multiplication). When the basic operation is understood explicitly indicating the dependence on the global spatial coordinates \( \xi \), the local spatial coordinate \( x \) plays no role. Then, the expression (1) can be written as

\[ U = \{aB\hat{\otimes}cT\} \]  

where the “hat” over the multilinear operation symbol acts as a marker of the operation involved in the convolution (it is useful when more than one operation symbol appears within the brackets).

Knutsson et al. also draw another definition of normalized convolution by the means of general convolution operations on data of signal/certainty type. Normalized convolution of \( aB \) and \( cT \) can be defined and denoted by:

\[ U_N = \{aB\hat{\otimes}cT\}_N = N^{-1}D \]  

where:

\[ D = \{aB\hat{\otimes}cT\} \]

\[ N = aB \odot B^\ast c \]

The star, \( \ast \), is the complex conjugate operator.

The concept of normalized convolution has a theory behind, the least squares problem.

To see this, express a given neighbourhood, \( t \), in a set of basis functions given by a matrix \( B \) and the coefficients \( u \). (standard matrix and vector notation are used.)

\[ t' = Bu \]  

It is proved that for a given set of basis functions, \( B \), the least square error, \( \|t' - t\| \), is minimized by choosing \( u \) to be:

\[ u = [B^TB]^{-1}B^Tt \]  

With the introduction of a diagonal matrix, \( W \), a weighted least square problem can be solved by

\[ Wt' = WBu \]
Then the minimum of $||W(t' - t)||$ is obtained by choosing coefficients $u$ to:

$$u = \left( (WB)^T W B \right)^{-1} (WB)^T W t$$  \hspace{1cm} (7)

which can be rewritten and split into two parts, $N^{-1}$ and $D$.

$$u = \left[ \frac{B^T W^2 B}{N} \right]^{-1} \frac{B^T W^2 t}{D}$$  \hspace{1cm} (8)

It can be shown that $N$ and $D$ are identical to the corresponding quantities used in normalized convolution, equation (3). In normalized convolution, the diagonal weight matrix is, for a neighbourhood centered on $\xi$, given by:

$$W^2_{ii}(\xi_0) = a(x_i) e(\xi_0 - x_i)$$  \hspace{1cm} (9)

Therefore, normalized convolution can be considered as a method used to obtain a local weighted least square error representation of the input signal. The input signal is described in terms of the basis function set, $B$, while the weights are adaptive and given by the data certainties and the operator applicability.

One of the most striking applications of normalized convolution might be the interpolation of Lena image (Figure 1) using applicability function illustrated in Figure 1.

![Figure 1. An applicability function.](image1)

**2.2. Fisher Vectors**

This section describes the Fisher Vector (FV) of [33]. The FV is an image representation obtained by pooling local image features. It is frequently used as a global image descriptor in visual classification. Given an input image $I$, the Fisher Vector (FV) formulation of [33] starts by extracting local descriptors $X = \{x_t, t = 1 \cdots T\}$, they can be densely and at multiple scales. It is assumed that $X$ can be modeled by a probability density function $u_\lambda$, with $\lambda$ is the set of parameters of $u$ as well as the estimation of those. Then, $X$ can be described by the gradient vector as

$$G^X\lambda = \frac{1}{T} \sum_{t=1}^{T} \nabla \log u_\lambda(x_t).$$  \hspace{1cm} (11)

Soft assignment ($\gamma_t(i)$) of descriptor $x_t$ to Gaussian $i$ is computed as

$$\gamma_t(i) = \frac{w_i u_i(x_t)}{\sum_{j=1}^{K} w_j u_j(x_t)}.$$  \hspace{1cm} (12)

Let $G^X_{\mu_i}$ be the gradient vector computed on the mean $\mu_i$, and $G^X_{\sigma_i}$ be the gradient vector computed on the standard deviation $\sigma_i$ of Gaussian $i$, then derivative of those will lead to formulations for computing the first and second order

![Figure 2. Top left: The famous Lena-image, top right the image has been degraded to a grey-level test image only containing 20% of the original information. Bottom left: Interpolation using standard convolution with a normalized smoothing filter (see Figure 1). Bottom right: Interpolation using normalized convolution with the same filter as applicability function.](image2)
statistics of the local descriptors.

\[
\Gamma_{\mu,i}^X = \frac{1}{T \sqrt{w_i}} \sum_{t=1}^{T} \gamma_t(i) \left( \frac{x_t - \mu_i}{\sigma_i} \right), \quad (13)
\]

\[
\Gamma_{\sigma,i}^X = \frac{1}{T \sqrt{2w_i}} \sum_{t=1}^{T} \gamma_t(i) \left[ \frac{(x_t - \mu_i)^2}{\sigma_i^2} - 1 \right]. \quad (14)
\]

The final Fisher vector is the concatenation of the \(\Gamma_{\mu,i}^X\) and \(\Gamma_{\sigma,i}^X\) vectors for \(i = 1...K\), with the dimensionality is \(2KD\), \(D\) is the dimension of a descriptor \(x_i\).

### 2.3. Scattering Network

Scattering Network (ScatNet) [27] is a handcrafted deep convolution network, in which cascade of wavelet transform and modulus non-linearities operators are consecu-

The average \(S_m x\) carries only the low frequencies of \(U_m x\) while the high frequencies are captured by roto-translation convolutions with wavelets. \(|W_m|\) transforms \(U_m x\) into the average \(S_m x\) and a new layer \(U_{m+1} x\) of wavelet amplitude coefficients: \(|W_m|U_m x = (S_m x, U_{m+1} x)\). Repetitively computing this wavelet modulus transform would generate multiple layers of scattering invariant coefficients. For \(m = 0\) \(U_0 x = x\), in case the network has three layers, the scattering feature vector \(S_x\) would be a concatenation of three \(S_i x\) coefficients such that \(S_x = (S_0 x, S_1 x, S_2 x)\). A filter bank of low-pass and high-pass filters for implementing Morlet wavelet \((W_m\) operator) is illustrated in Figure 4.

### 2.4. Normalized-Convolution Network

This section introduces a handcrafted network which inherits ScatNet [41] for texture classification. In this novel handcrafted network, so called Normalized-Convolution Network (NmzNet), we propose two important changes to the ScatNet (section 2.3) to enhance its robustness.

i) We modify the wavelet modulus operator, the fundamental operator of ScatNet, by a substitution of the normal-

![Figure 4. Complex Morlet wavelets with Gaussian kernel (top left corner), different scales (along rows) and orientations (along columns). The real and imaginary parts are shown on the left and on the right, respectively.](image)

**Figure 4.** Complex Morlet wavelets with Gaussian kernel (top left corner), different scales (along rows) and orientations (along columns). The real and imaginary parts are shown on the left and on the right, respectively.
3.1. Experimental Settings

We analyze the effectiveness of our method by doing experiments on three popular texture databases, and their testing protocols are strictly followed. The SVM classifier is used to produce texture classification results.

Arguments of NmzNet are selected such that scaling number is \( J = 4 \) for image size \( 300 \times 300 \) or smaller, and \( J = 5 \) otherwise. Here, the principle is that the smaller image size, the smaller scaling number chosen as recommended in \([41]\). Orientations of the filter bank are set to 8 (\( L = 8 \)), the number of NmzNet layers is set to 3 (\( M = 3 \)). Since if the layer number exceeds this threshold, both feature dimension and feature extraction time increase with minor improvement in classification accuracy.

Experiments were conducted on datasets, with samples of those represented in Figure 5, and the summary is in Table 1.

![Figure 5](image_url) Images in the same class from the 3 experimented datasets: Rows 1) UIUC, 2) KTH-TIPS-2a, 3) KTH-TIPS-2b.

Table 1. Summary of datasets for the experiments.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Images</th>
<th>Image size</th>
<th>Classes</th>
<th>Splits</th>
</tr>
</thead>
<tbody>
<tr>
<td>UIUC</td>
<td>1000</td>
<td>640 \times 480</td>
<td>25</td>
<td>100</td>
</tr>
<tr>
<td>KTH-TIPS2a</td>
<td>4608</td>
<td>200 \times 200</td>
<td>11</td>
<td>pre-defined</td>
</tr>
<tr>
<td>KTH-TIPS2b</td>
<td>4752</td>
<td>200 \times 200</td>
<td>11</td>
<td>pre-defined</td>
</tr>
</tbody>
</table>

UIUC [21] has 25 classes of texture, each class contains 40 different images which has the resolution of \( 640 \times 480 \) include changes of viewing angle, scale, and illumination conditions. The mean classification accuracy, 100 random splits between training and testing with, a half of samples (20 images) per class chosen for training and the rest for testing, is reported.

The material databases KTH-TIPS2a, KTH-TIPS-2b [3], with three viewing angles, four illuminants, and nine different scales, produce 432 images per class, with the image size of \( 200 \times 200 \) and 11 classes in total. Regarding these databases, we follow the common testing and training protocols. Only unseen data is used for testing, with three out of four samples used for training and the remaining for testing.

3.2. Experimental Analysis

This section discusses how the proposed method enhances texture representation when normalized convolution used. As illustrated in (Figure 6), the first row presents (from left to right) a Gaussian window used as applicability function, the real and imaginary part of one Gabor filter, windowed filters respectively which can be used for computing normalized convolution. The second row is a lettuce leaf image and its corresponding convolution results. It is obvious that normalized convolution results (right) retain more details from the image than those of standard convolution. This clue inspires us to build a convolution handcrafted network based on ScatNet with normalized convolution used instead of standard convolution.

We built a convolution network based on ScatNet model with normalized convolution substituted for standard convolution, so called normalized-convolution-ScatNet, then extracted feature vectors of images (Figure 7), taken from the KTH-TIPS-2b dataset which has the same class (cork). We next computed \( \chi^2 \) distance between those, it is 0.45. The same procedures are done with ScatNet, and the \( \chi^2 \) distance is 14.89. It is obvious that our approach has a better discriminative representation compared with its derivation (ScatNet).

Finally, We use off-the-shelf method Fisher vector as an aggregation method for our approach. Classification results are presented in the next section, with SVM classifier is used.

3.3. Classification Results

Experiments were conducted on three texture datasets, the results are compared with well-known and state-of-the-art of those, we chose the highest results reported by relevant articles for the comparison.

Results in Table 2 shows that our accurate rate on UIUC dataset is similar to that of works in \([49]\), roundly 1% lower than the recent frontier FV-SIFT+ FC+FV-VD \([6]\) of this dataset while beating all others.

KTH-TIPS-2a \([3]\) is a very challenging dataset for the texture classification task. State-of-the-art performance on this dataset was reported approximately 78%. KTH-TIPS-2a is challenging because it has very intense intra-class variations, as evident in Figure 5. Table 3 shows the experimental results obtained on KTH-TIPS-2a. NmzScat achieved a 82.37% classification rate, which was superior to the results obtained by handcrafted conventional methods with a wide margin.

Results on KTH-TIPS-2b: As can be seen from Table 4, our proposed method gets a significant improvement over the original features it inherits from, the enhancement in classification accuracy is approximate 10% compared with
Figure 6. Normalized Convolution compared with Standard Convolution, lettuce leaf image from KTH-TIPS-2b dataset

Figure 7. Images in the same class (cork) from KTH-TIPS-2b dataset whose NmzNet descriptors have smaller $\chi^2$ distances compared with those of ScatNet.

Table 2. Classification accuracy comparisons with state-of-the-art on the UIUC database.

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MDLBP [40]</td>
<td>88.05</td>
</tr>
<tr>
<td>LBPHF [1]</td>
<td>89.58</td>
</tr>
<tr>
<td>Hayman et al. [17]</td>
<td>92.00</td>
</tr>
<tr>
<td>CLBP [52]</td>
<td>92.55</td>
</tr>
<tr>
<td>MFS [48]</td>
<td>92.74</td>
</tr>
<tr>
<td>BRINT [25]</td>
<td>93.30</td>
</tr>
<tr>
<td>MRELBP (SVM) [24]</td>
<td>96.88</td>
</tr>
<tr>
<td>ScatNet (PCA) [41]</td>
<td>96.15</td>
</tr>
<tr>
<td>VZ-joint [44]</td>
<td>97.83</td>
</tr>
<tr>
<td>OTF [47]</td>
<td>97.40</td>
</tr>
<tr>
<td>scLBP [39]</td>
<td>98.45</td>
</tr>
<tr>
<td>FV-AlexNet (SVM) [6]</td>
<td>99.0</td>
</tr>
<tr>
<td>FV-VGGM (SVM) [6]</td>
<td>99.7</td>
</tr>
<tr>
<td>FV-VGGVD (SVM) [5]</td>
<td>99.8</td>
</tr>
<tr>
<td>NmzNet (ours)</td>
<td>96.97</td>
</tr>
</tbody>
</table>

Table 3. Classification accuracy comparisons with state-of-the-art on the KTH-TIPS-2a database.

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>LBP [32]</td>
<td>58.10</td>
</tr>
<tr>
<td>CLBP [53]</td>
<td>66.58</td>
</tr>
<tr>
<td>VZ-MR8 [43]</td>
<td>62.35</td>
</tr>
<tr>
<td>VZ-joint [44]</td>
<td>61.93</td>
</tr>
<tr>
<td>Multi-Scale-BIF [7]</td>
<td>71.56</td>
</tr>
<tr>
<td>ScatNet [41]</td>
<td>72.57</td>
</tr>
<tr>
<td>COV-LBPD [18]</td>
<td>74.86</td>
</tr>
<tr>
<td>MLEP [50]</td>
<td>75.57</td>
</tr>
<tr>
<td>scLBP [39]</td>
<td>78.39</td>
</tr>
<tr>
<td>NmzNet (ours)</td>
<td>82.37</td>
</tr>
</tbody>
</table>

ScatNet. Although our classification result on this database is inferior to FV-SIFT+ FC+FV-VD [6], it is on the high ranking.

4. Conclusions

In this paper, we have proposed NmzNet, a handcrafted convolution network which leverages off-the-shelf methods such as Normalized Convolution, ScatNet, Fisher Vector aggregation for efficient texture classification. Our approach extracts features which are invariant to scale, rotation, and illumination changes.

Future works can be drawn on other categories of texture data such as noise and occlusion data with other types of feature aggregation.
<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>VZ-MR8</td>
<td>55.70</td>
</tr>
<tr>
<td>NRLBP[38]</td>
<td>57.00</td>
</tr>
<tr>
<td>NTLBP[10]</td>
<td>58.78</td>
</tr>
<tr>
<td>MBP[14]</td>
<td>60.29</td>
</tr>
<tr>
<td>LBP[32]</td>
<td>60.35</td>
</tr>
<tr>
<td>VZ-Path [44]</td>
<td>60.70</td>
</tr>
<tr>
<td>LTP[62]</td>
<td>61.17</td>
</tr>
<tr>
<td>PRICoLBp[36]</td>
<td>63.47</td>
</tr>
<tr>
<td>COV-LBPD[18]</td>
<td>64.84</td>
</tr>
<tr>
<td>MWLD[5]</td>
<td>65.70</td>
</tr>
<tr>
<td>ELBP[26]</td>
<td>68.48</td>
</tr>
<tr>
<td>MSJLBp[35]</td>
<td>65.51</td>
</tr>
<tr>
<td>MDLBP$^{323}$ [40]</td>
<td>66.52</td>
</tr>
<tr>
<td>BRINT[25]</td>
<td>66.67</td>
</tr>
<tr>
<td>LBPFFH[1]</td>
<td>67.51</td>
</tr>
<tr>
<td>CLBPFFH[52]</td>
<td>68.10</td>
</tr>
<tr>
<td>ScanNet(PCA)[61]</td>
<td>68.92</td>
</tr>
<tr>
<td>MRELBP[23]</td>
<td>68.98</td>
</tr>
<tr>
<td>MRELBP (SVM)[24]</td>
<td>77.91</td>
</tr>
<tr>
<td>FV-AlexNet (SVM) [6]</td>
<td>77.9</td>
</tr>
<tr>
<td>FV-VGGM (SVM) [6]</td>
<td>77.9</td>
</tr>
<tr>
<td>FV-VGGVD[6]</td>
<td>88.2</td>
</tr>
<tr>
<td>NmzNet (ours)</td>
<td>78.10</td>
</tr>
</tbody>
</table>

References


