Computer Vision and Image Understanding 142 (2016) 13–22

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

Local Higher-Order Statistics (LHS) describing images with statistics of local non-binarized pixel patterns☆

Gaurav Sharma[a,b,1,*], Frédéric Jurie[a]

[a] GREYC CNRS UMR 6072, Université de Caen Basse-Normandie, France
[b] Max Planck Institute for Informatics, Germany

ARTICLE INFO

Article history:
Received 21 January 2015
Accepted 20 September 2015
Available online 28 September 2015

Keywords:
Local features
Texture categorization
Face verification
Image classification

ABSTRACT

We propose a new image representation for texture categorization and facial analysis, relying on the use of higher-order local differential statistics as features. It has been recently shown that small local pixel pattern distributions can be highly discriminative while being extremely efficient to compute, which is in contrast to the models based on the global structure of images. Motivated by such works, we propose to use higher-order statistics of local non-binarized pixel patterns for the image description. The proposed model does not require either (i) user specified quantization of the space (of pixel patterns) or (ii) any heuristics for discarding low occupancy volumes of the space. We propose to use a data driven soft quantization of the space, with parametric mixture models, combined with higher-order statistics, based on Fisher scores. We demonstrate that this leads to a more expressive representation which, when combined with discriminatively learned classifiers and metrics, achieves state-of-the-art performance on challenging texture and facial analysis datasets, in low complexity setup. Further, it is complementary to higher complexity features and when combined with them improves performance.

1. Introduction

Categorization of textures and analysis of faces under multiple and difficult sources of variations like illumination, scale, pose, expression and appearance are challenging problems in computer vision with many important applications. Texture recognition is beneficial for applications such as mobile robot navigation or biomedical image processing. It is also related to facial analysis e.g. facial expression categorization and face verification (two faces are of same person or not), as the models developed for textures are generally found to be competitive for face analysis. Analysis of faces, similarly, has important applications especially in human computer interaction and in security and surveillance scenarios. This paper proposes a new model for obtaining a powerful and highly efficient representation for textures and faces, with such applications in mind.

Initial success on texture recognition was achieved by the use of filter banks [4–8], where the distributions of the filter response coefficients were used for discrimination. The focus was on evaluating appropriate filters, selective for edge orientation and spatial-frequencies of variations, and better capturing the distributions of such filter responses. However, later works e.g. by Ojala et al. [9] and Varma and Zisserman [10], showed that it is possible to discriminate between textures using pixel values directly (with pixel neighborhoods as small as 3×3 pixels), discounting the necessity of filter banks. It was demonstrated that despite the global structure of the textures, very good discrimination could be achieved by exploiting the distributions of such small pixel neighborhoods. More recently, exploiting such small pixel neighborhoods or *micro-structures* in textures by representing images with distributions of local descriptors has gained much attention and has led to state-of-the-art performances for systems with low complexity, e.g. Local Binary Patterns (LBP) [1,2], Local Ternary Patterns (LTP) [11] and Weber Local Descriptor (WLD) [12]. Most of such local pixel neighborhood based descriptors were shown to be highly effective for facial analysis [2,11] as well. However these methods suffer from important limitations–the use of fixed hard quantization of the feature space (the space of small pixel patterns) and the use of heuristics to prune uninteresting regions in the feature space. In addition, they use histograms to represent the feature distributions. Histograms, or count statistics, are zeroth order statistics of distributions and thus give a quite restrictive representation.

In contrast, we propose a model that represents images with higher-order statistics of small local pixel neighborhoods. Fig. 1

☆ This paper has been recommended for acceptance by Aleix M Martinez.

* Corresponding author at: Max Planck Institute for Informatics, Germany.
E-mail addresses: grvsharma@gmail.com (G. Sharma), frederic.jurie@unicaen.fr (F. Jurie).

[1] GS is currently with Max Planck Institute for Informatics. The majority of the work was done when GS was with the Université de Caen Basse-Normandie.

http://dx.doi.org/10.1016/j.cviu.2015.09.007
1077-3142/

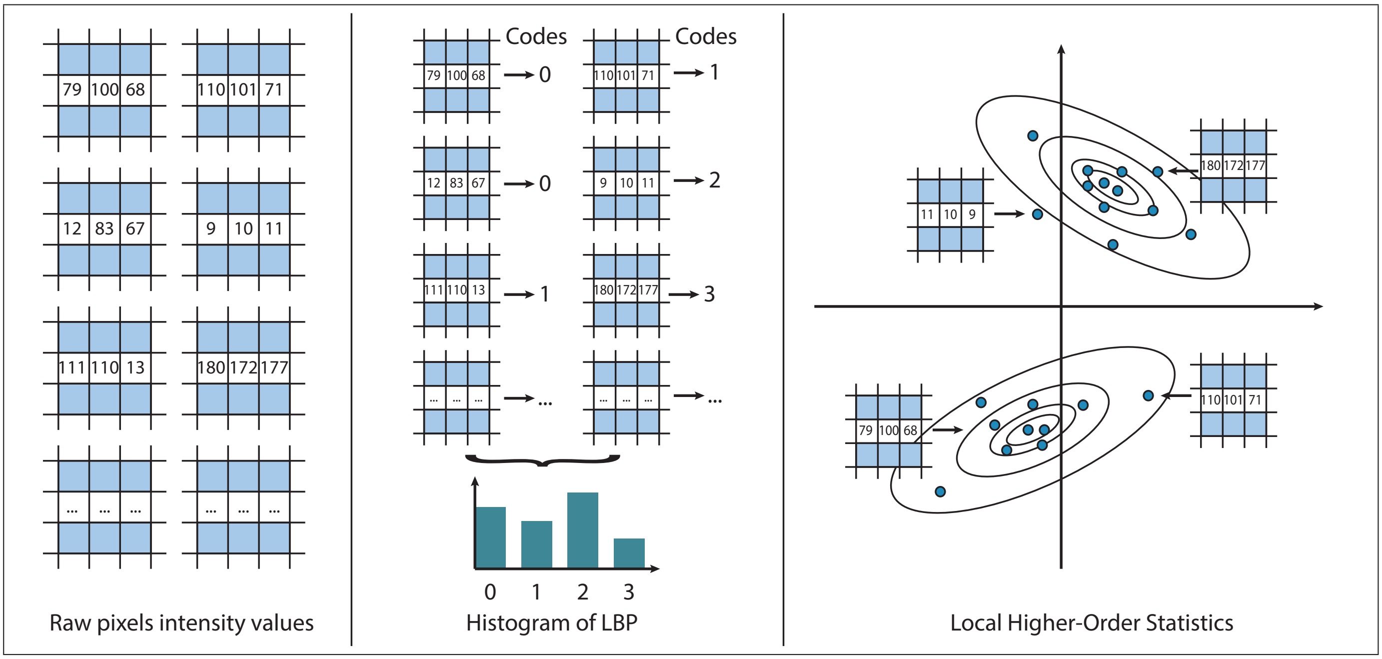

Fig. 1. Illustration of the proposed Local Higher-order Statistics (LHS) representation. The left-hand side of the figure represents a collection of pixel-centered raw pixel intensity values (image patches). For making the figure simple, we consider only horizontal 2-neighborhood (in practice we use a 3 × 3 neighborhood). The middle of the figure shows how these patches can be turned into LBP codes [1,2] – 4 different codes in this case – and then represented as an histogram of LBP. The proposed representation, illustrated on the right-hand side of the figure, is much richer, as the distribution of the local patches is represented by a Gaussian Mixture model, encoded as Fisher scores [3].

shows an illustration of this representation. We obtain a data driven soft partition of the feature space using parametric mixture models, to represent the distribution of the vectors, with the parameters learnt from the training data. Hence, in the proposed method, the coding of vectors is intrinsically adapted to the data and the computations involved remain very simple despite the strengths. This helps us avoid the above mentioned limitations of the previous methods – (i) instead of a fixed quantization, we learn a data driven, and hence, adaptive quantization using Gaussian mixture models (GMM), (ii) quantizing using GMM also avoids any heuristic pruning as any low occupancy region in the feature space will be automatically ignored by the GMM learning and (iii) learning GMM allows us to use Fisher vectors [3] which are higher-order statistics of the feature distribution. We discuss in more detail on this in the following sections. A preliminary version of this work appeared in Sharma et al. [13].

We validate the proposed representation by extensive experiments on four challenging datasets: (i) Brodatz 32 texture dataset [14,15], (ii) KTH TIPS 2a materials dataset [16], (iii) Japanese Female Facial Expressions (JAFFE) dataset [17], and (iv) Labeled Faces in the Wild (LFW) dataset [18]. Two dataset, Brodatz-32 and JAFFE, are relatively easier with limited variations while the other two, KTH TIPS 2a and LFW, are more challenging with realistic high levels of variation in illumination, pose, expressions etc. We show that using higher-order statistics gives a more expressive description and lead to state-of-the-art performance in low complexity settings, for the above datasets. Further, with the challenging LFW dataset as the experimental testbed, we also show that the proposed representation is complementary to the recent high complexity state-of-the-art representations. However, in case of challenging variations, like in LFW, unsupervised approach is not sufficient and hence we show that when used with supervised metric learning the performance of the proposed representation improves substantially. When combined with higher complexity methods, the proposed representation achieves the state-of-the-art performance on the challenging LFW dataset in the supervised protocol, when no external labeled data is used.

2. Related works

Texture analysis was initially addressed using filter banks and the statistical distributions of their responses e.g. [4–8]. Most of the initial works proposed appropriate directionally and frequency-adapted multiscale filter banks and/or methods to better capture the statistical distributions of their responses. Later, Ojala et al. [9] and, more recently, Varma and Zisserman [10] showed that statistics of small pixel neighborhoods, as small as 3 × 3 pixels, are capable of achieving high discrimination. This was in contrast to first convolving the local patches with filter banks and then taking their responses. The success of using raw pixel patches without any processing discounted the use of filter banks for texture recognition. Since then many methods working directly with local pixel neighborhoods have been used successfully in texture and face analysis e.g. Local Binary Patterns (LBP) [1,2], Local Ternary Patterns (LTP) [11] and Weber Law Descriptor (WLD) [12].

Local pixel pattern operators, such as Local Binary Patterns (LBP) by Ojala et al. [9], have been very successful for image description. LBP based image representation aims to capture the joint distribution of pixel intensities in a local neighborhood as small as 3 × 3 pixels. LBP makes two approximations, (i) it takes the differences between the center pixel and its eight neighbors and (ii) then considers just the signs of the differences. The first approximation lends invariance to gray-scale shifts and the second to intensity scaling. As an extension to LBP, Local Ternary Patterns (LTP) were introduced by Tan and Triggs [11] to add resistance to noise. LTP adds an additional parameter t, which defines a tolerance for similarity between different gray intensities, allowing for robustness to noise. Doing so lends an important strength: LTPs are capable of encoding pixel similarity information modulo noise using the simple rule that any two pixels within $\pm t$ intensity of each other are considered similar. This is accompanied by a clever split coding scheme to control the size of the descriptor. However, LTP (and LBP) coding is still limited due to its hard and fixed quantization. In addition, both LBP and LTP representations usually use the so-called *uniform* patterns: patterns with at most one 0–1 and