Pattern Recognition Letters 33 (2012) 1224–1235

Contents lists available at SciVerse ScienceDirect

Pattern Recognition Letters

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

Fast Local Self-Similarity for describing interest regions

Jingneng Liu [a,*], Guihua Zeng [a,*], Jianping Fan [b,*]

[a] State Key Laboratory of Advanced Optical Communication Systems and Networks, Key Lab on Navigation and location-based Service, Department of Electronic Engineering, Shanghai Jiaotong University, Shanghai 200240, China
[b] Department of Computer Science, University of North Carolina-Charlotte, Charlotte, NC 28223, USA

ARTICLE INFO

Article history:
Received 12 August 2010
Available online 1 February 2012

Keywords:
Local Self-Similarity
Fast Local Self-Similarity
SIFT
Region description
Image matching
Object classification

ABSTRACT

Two novel methods for extracting distinctive invariant features from interest regions are presented in this paper. The idea of these methods are associated with that measuring similarity between visual entities from images can be based on matching the internal layout of Local Self-Similarities. The main contributions are two-folds: firstly, two new texture features called Local Self-Similarities (LSS,C) and Fast Local Self-Similarities (FLSS,C) based on Cartesian location grid, are extracted, which are the modified versions of the well-known Local Self-Similarities (LSS,LP) feature based on Log-Polar location grid. To combine the powers of the SIFT and LSS (LP), LSS and FLSS are used as the local features in the SIFT algorithm. Secondly, different from the natural LSS (LP) descriptor that chooses the maximal correlation value in each bucket to get photometric translations invariance, the proposed LSS (C) and FLSS (C) adopt distribution-based representation to achieve more robust geometric translations invariance. In the contexts of image matching and object category classification experiments, the LSS (C) and FLSS (C) both outperform the original LSS (LP), and achieve favorably comparable performance to the SIFT. Furthermore, these descriptors are low computational complexity and simpler than the SIFT.

1. Introduction

A critical aspect of computer vision research involves in detecting and describing of the keypoints or interest regions. Nowadays, the local image features, which are distinctive and yet invariant to many kinds of geometric and photometrical transformations, have been attracting more and more attention because of their promising performance. Usually, there are two different ways for utilizing the local features for image content representation (Li and Allinson, 2008), i.e., the traditional utilization which consists of feature detection, feature description and feature matching, and the Bag-of-Features (Sivic and Zisserman, 2003) which consists of feature detection, feature description, feature clustering and frequency histogram construction. Today, they are the preferred strategy for solving a wide variety of problems, for such tasks as image retrieval (Mikolajczyk and Schmid, 2001), wide baseline matching (Tuytelaars and Van Gool, 2004), object recognition (Lowe, 2004), texture recognition (Lazebnik et al., 2005), object category classification (Hörster et al., 2008) and robot localization (Se et al., 2002).

The process for extracting the local features consists of a feature detector and a feature descriptor. The feature detectors provide the feature points to be matched and determine the neighboring regions to compute the descriptors. Many feature detector methods have been proposed in the literature. Most of the existing detectors can be categorized into three types (Tuytelaars and Mikolajczyk, 2008): (1) corner detectors such as Harris and SUSAN detectors; (2) blob detectors such as Hessian and Hessian-Affine detectors and (3) region detectors such as MSER detector. This paper focuses on the feature descriptors only, with the emphasis on local image features well suited for image understanding applications. Existing detector performance evaluation (Mikolajczyk et al., 2005) has offered more information on interest region detection.

Given invariant interest regions from feature detectors, the remaining process is to describe interest regions around the feature points. Many different descriptors for feature points and interest regions have been developed and proven to be very successful in applications such as object recognition, image retrieval. There are lots of possible descriptors that emphasize a diverse set of image properties such as pixel intensity, gradient, color, texture, contour, edge and so on. In this work, we focus on the descriptors that are computed on the gray-value images. The local descriptors can be categorized as the followings: distribution-based, spatial-frequency techniques, and differential-based descriptors. Many of the proposed descriptors are distribution-based, i.e., they use histograms to represent different characteristics of appearance or shape. Lowe (2004) developed a Scale-Invariant Feature Transform (SIFT) descriptor based on the gradient distribution in the detected regions. SIFT is invariant to image scaling and rotation, and partially invariant to change in illumination and 3D camera viewpoint.

* Tel./fax: +1 8621 3420 4361.
E-mail addresses: bzljn@163.com (J. Liu), ghzeng@sjtu.edu.cn (G. Zeng), jfan@uncc.edu (J. Fan).

0167-8655/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.patrec.2012.01.013

The SIFT has been proven to be the most robust among the local invariant feature descriptors with respect to different geometric changes (Mikolajczyk and Schmid, 2005). One extension of the SIFT descriptor is the GLOH descriptor (Mikolajczyk and Schmid, 2005), which replaces the Cartesian location grid used by the SIFT with a log-polar one, and applies PCA to reduce the descriptor dimension. Similar to GLOH that reduces the descriptor dimension by PCA is the PCA-SIFT (Ke and Sukthankar, 2004). The up-to-date local feature descriptor is the center-symmetric local binary pattern (CS-LBP) (Heikkilä et al., 2009). It combines the strengths of the SIFT and LBP, which uses the CS-LBP as the local feature in the SIFT algorithm. The descriptor performs favorably compared to the SIFT. Further, the CS-LBP descriptor is computationally simpler than the SIFT. Belongie et al. (2002) proposed the shape context descriptor by starting with a collection of shape points, and, for each point, building a histogram to describe the relative distribution of the other points in log-polar space. The shape context is scale and rotation invariant. The SURF descriptor (Herbert et al., 2006) builds on the strengths of the leading existing detectors and descriptors. It uses a Hessian matrix-based measure for the detector and Haar wavelet responses for the descriptor. Relying on integral images for image convolutions, computation time is significantly reduced. In earlier research (Ling and Jacobs, 2005), geodesic sampling is used to get neighborhood samples for interest points and then a geodesic-intensity histogram (GIH) is used as a deformation invariant local descriptor.

There are several recent comparative studies on performance evaluation of local region descriptors (Mikolajczyk and Schmid, 2005; Moreels and Perona, 2005). Almost all the experimental comparison results revealed that the best discriminative descriptors are distribution-based descriptors such as GLOH, SIFT and CS-LBP. Recently Hörster et al. (2008) investigated the influence of different types of the local feature descriptors in the context of scene recognition based on PLSA image models. Experimental comparison revealed that the commonly used SIFT descriptor is outperformed by the two other feature descriptors: the geometric blur and the LSS (LP). Although a thorough comparison of many local region descriptors in the contexts of matching and recognizing the same object or scene is presented elsewhere, a detailed evaluation for the advanced local region descriptor LSS (LP) in these contexts is still missing.

For image and video matching, Shechtman and Irani (2007) explored the Local Self-Similarity descriptor based on Log-Polar location grid, namely LSS (LP). The internal geometric layout of Local Self-Similarities (LSS) is introduced in LSS (LP), even though the patterns generating those Local Self-Similarities are quite different or up to some distortions in each of the images or videos. It has been successfully employed for the purpose of object detection, image retrieval, and action detection. The descriptor has some advanced properties, such as invariance to color changes, and computationally simpler than the SIFT. Drawbacks are that this descriptor is only invariance against small local affine and nonrigid deformations, and insensitive to small translations. To address these problems, in this paper, we propose two new LSS based texture features, i.e., Local Self-Similarity and Fast Local Self-Similarity descriptors based on Cartesian location grid, namely LSS (C) and FLSS (C), which are more suitable for different computer vision tasks.

Our approach has a closer relation to the notion of SIFT, CS-LBP. Because the SIFT, CS-LBP and other distribution-based descriptors have shown state-of-the-art performance in different computer vision tasks, we decided to focus on this approach. We are specially desired to see whether the LSS feature and Log-Polar location grid based used in the LSS (LP) algorithm could be replaced by a different feature and location grid that offers better or comparable performance. In this paper, two new interest region descriptors are developed, namely LSS (C) and FLSS (C), which combine the good properties of the SIFT, CS-LBP and LSS (LP). They are achieved by adopting the SIFT descriptor algorithm and using the novel LSS and FLSS features instead of original gradient feature. To the best of our knowledge, we are the first to explore the Log-Polar based LSS (LP) to Cartesian based LSS (C) and FLSS (C) descriptors for image matching and image classification. The new LSS and FLSS features allow simplifications of several steps of the algorithm, which make the resulting descriptors computationally simpler than SIFT and either as simple as or faster than CS-LBP. They also appear to be more robust and stable than the original LSS (LP) descriptor.

The framework of this paper follows the novel CS-LBP (Heikkilä et al., 2009). The rest of the paper is organized as follows. In Section 2, we first briefly describe the starting point for our work, i.e., the SIFT and LSS (LP) methods. Sections 3 and 4 give details for the FLSS (C) operator and the LSS (C) and FLSS (C) descriptors, respectively. The experimental evaluation is carried out in Section 5. Finally, we conclude the paper in Section 6.

2. SIFT and Local Self-Similarity Descriptors

In this work we address LSS based texture feature for different computer vision tasks. The describing methods most closely related to our approach are SIFT, CS-LBP. So before presenting in detail the proposed FLSS operator and the LSS (C) and FLSS (C) descriptors, we give a brief review of the SIFT and the original LSS (LP) methods that form the basis of our work.

2.1. SIFT descriptor

SIFT (Lowe, 2004) descriptors are computed for normalized image interest regions. A SIFT descriptor is a 3D histogram of gradient with location and orientation, where location is quantized into a 4×4 Cartesian location grid and the gradient angle is quantized into eight orientations. The resulting descriptor is of dimension 128. Each orientation plane represents the gradient magnitude corresponding to a given orientation. To obtain illumination invariance, the descriptor is normalized by the square root of the sum of squared components.

2.2. Local Self-Similarity Descriptor

Shechtman and Irani (2007) first proposed a descriptor based on LSS feature. This descriptor has been quickly adopted in the object detection and classification community yet another local descriptor in Bag-of-Visual-Words frameworks (Chatfield et al., 2009; Hörster and Lienhart, 2008; Lampert et al., 2009; Vedaldi et al., 2009) or in nearest-neighbor classifiers (Boiman et al., 2008). Junejo et al. (2008) also performed human action recognition in video by using temporal Self-Similarities extended from LSS (LP).

The LSS (LP) descriptor showed in Fig. 1, captures the internal geometric layout of Local Self-Similarities (LSS) and can be compared across images that appear substantial difference at pixel level. To derive the LSS (LP) descriptor d_q associated with an image pixel q, the surrounding image patch (typically patch size: $P \times P = 5 \times 5$) is compared with a larger surrounding image region centered at q (typically radius 20, region size: $N \times N = 41 \times 41$), using simple sum of square differences (SSD) between patch properties such as pixel intensity and color. The resulting distance surface $SSD_q(x,y)$ is normalized and transformed into a "correlation surface" $S_q(x,y)$:

$$S_q(x,y) = \exp\left(-\frac{SSD_q(x,y)}{\max(var_{noise}, var_{auto}(q))}\right) \quad (1)$$