Pattern Recognition 68 (2017) 1–13

Contents lists available at ScienceDirect

Pattern Recognition

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

Content based image retrieval with sparse representations and local feature descriptors : A comparative study

Ceyhun Celik[a], Hasan Sakir Bilge[b,*]

[a] Department of Computer Engineering, Gazi University, Ankara, Turkey
[b] Department of Electrical-Electronics Engineering, Gazi University, Ankara, Turkey

ARTICLE INFO

Article history:
Received 23 August 2016
Revised 10 February 2017
Accepted 3 March 2017
Available online 6 March 2017

Keywords:
Content based image retrieval
Local feature descriptor
Sparse representation
Dictionary learning
Coefficient learning

ABSTRACT

Content Based Image Retrieval (CBIR) has been widely studied in the last two decades. Unlike text based image retrieval techniques, visual properties of images are used to obtain high level semantic information in CBIR. There is a gap between low level features and high level semantic information. This is called semantic gap and it is the most important problem in CBIR. The visual properties were extracted from low level features such as color, shape, texture and spatial information in early days. Local Feature Descriptors (LFDs) are more successful to increase performance of CBIR system. Then, a semantic bridge is built with high level semantic information. Sparse Representations (SRs) have become popular to achieve this aim in the last years.

In this study, CBIR models that use LFDs and SRs in literature are investigated in detail. The SRs and LFD extraction algorithms are tested and compared within a CBIR framework for different scenarios. Scale Invariant Feature Transform (SIFT), Speeded-Up Robust Features (SURF), Histograms of Oriented Gradients (HoG), Local Binary Pattern (LBP) and Local Ternary Pattern (LTP) are used to extract LFDs from images. Random Features, K-Means and K-Singular Value Decomposition (K-SVD) algorithms are used for dictionary learning and Orthogonal Matching Pursuit (OMP), Homotopy, Lasso, Elastic Net, Parallel Coordinate Descent (PCD) and Separable Surrogate Function (SSF) are used for coefficient learning. Finally, three methods recently proposed in literature (Online Dictionary Learning (ODL), Locality-constrained Linear Coding (LLC) and Feature-based Sparse Representation (FBSR)) are also tested and compared with our framework results. All test results are presented and discussed. As a conclusion, the most successful approach in our framework is to use LLC for Coil20 data set and FBSR for Corel1000 data set. We obtain 89% and 58% Mean Average Precision (MAP) for Coil20 and Corel1000, respectively.

1. Introduction

Content Based Image Retrieval (CBIR) became more important and also practical with the increasing number of digital images stored on media devices and need of semantic search on these images. With developing technologies and the widespread usage of Internet, shared and stored data are substantially increased. A great majority of these data are multimedia and images. Accessing to these multimedia and images is an important problem. At first, Query By Text (QBT) methods are used to access these data [1–3]. The user searches the images with text queries in this case. Annotations are required for all images to apply QBT. First challenge on using QBT is specifying the annotations of all images and this is a difficult task. Second challenge is ambiguity in annotations [3]. Query By Image (QBI) is more useful than QBT because of these difficulties [1,3]. QBI methods are also called as CBIR in literature. User generates image queries to search images on CBIR systems. It is aimed to obtain High Level Semantic Information (HLSI) using extracted Low Level Features (LLFs) of the images with generated queries [4,5]. Extracted LLFs from images are color, shape, texture and spatial layout informations. CBIR methods assume that semantic similarity is introduced with these features, i.e., a bridge could be built between LLFs and HLSI. Nevertheless, how the bridge between LLFs and HLSI will be built is the primary problem CBIR approaches have to solve [6]. This problem is called semantic gap.

The CBIR models consist of three main steps. First, the feature extraction and feature selection are done. Then, similarity measurements are calculated. Finally, indexing and retrieval are performed. In literature, there are many surveys and reviews that analyze these steps in detail [1–11]. In this study, just the Local

[*] Corresponding author.
E-mail addresses: celik.ceyhun@gmail.com (C. Celik), bilge@gazi.edu.tr (H.S. Bilge).

http://dx.doi.org/10.1016/j.patcog.2017.03.006
0031-3203/© 2017 Elsevier Ltd. All rights reserved.

Feature Descriptors (LFDs) for the first step and the Sparse Representation (SR) algorithms for the second step are analyzed and compared.

Feature extraction and selection are the first steps of the CBIR. These processes are performed as region based or whole image based [11]. Image descriptors are obtained with the help of color, shape, texture and spatial layout information, then these global descriptors are used for image retrieval. Usage of local descriptors have become more popular in last years. Local descriptors could be consistent with exactly the same features except these descriptors are extracted from regions of the image instead of the whole image [3]. Furthermore, there are local descriptors that also are extracted from the whole image. Especially following LFDs are used extensively in CBIR: Scale Invariant Feature Transform (SIFT) [12], Speeded-Up Robust Features (SURF) [13], Histograms of Oriented Gradients (HoG) [14], Local binary Pattern (LBP) [15]. In addition to these LFDs, Local Ternary Pattern (LTP) is recently proposed as the improved version of LBP [16].

Similarity measurement is the second step. This is the key step of CBIR models since semantic gap is tried to be reduced with this step [3]. In the early years of CBIR, it is emphasized that different similarity measurements are required for each feature [4]. Similarity measurement is basically the measurement of distance between image descriptors. Essential measurements such as Euclidean and Mahalanobis distances were the first measurements at the early years of CBIR. However, similarity learning have become popular instead of similarity measurement recently [3,17]. Furthermore, machine learning techniques are also used in CBIR approaches, so complex models are proposed [18–20]. Performance of the system could be improved with Relevance Feedback (RF) which helps for training of the system with user feedbacks [21,22]. In the last years, SR is used extensively in CBIR approaches for the second step. Publishing dates of the studies about this topic are between 2009 and 2016. The aim of SR is representing the input signals with a simple combination. This representation is done with sparse coefficient set based on a dictionary. Usage of SR can be divided into two categories: traditional way and the improved way.

SR is used in traditional way in the following studies [23–28]. When the CBIR models with traditional SR are analyzed, it is observed that the usage of LFDs are common [23–26,28]. Localization of images on the dictionary, learning strategy and similarity measurement are the novelties of CBIR models with traditional SR in literature. In [23], [24] and [25], the sum of the coefficients that are obtained for each LFD are tried to minimize. In [26] and [28], max pooling method is used to combine coefficients that are obtained from different LFDs. SR is used for building graphs in [27]. In general, the image labels are not utilized [23,24,26,28].

Besides traditional SR with CBIR, CBIR models with improved SR are presented in the following studies [29–40].

In CBIR with improved SR, image labels are also included to dictionary and coefficient learning. While the sparse vectors are constructed, similar parts or similar images are tried to be on the same region of dictionary [29–31,34–36]. Thus, similar images are featured to each other. Usage of Bag of Features (BoFs) and Bag of Words (BoWs) models in CBIR with SR is common. Local descriptors are used in BoF models. On the other hand, shared space of images is used in BoW models. Feature extraction strategy, combining different dictionary or coding algorithm and improved Coefficient Learning (CL) algorithms are novelties of the studies [32,33,37–40]. Both LFDs and the others are tried to be as input space in CBIR with improved studies, since feature extraction is the key point of retrieval system.

Extracted features from images, dictionary and CL algorithms are shown on Table 1 for each CBIR study in literature. SIFT, SURF, HoG and LBP are commonly used for LFDs. Random Features, K-Means and K-Singular Value Decomposition (K-SVD) [41] give effective solutions for Dictionary Learning (DL). Orthogonal Matching Pursuit (OMP) [42], Homotopy [43], Least Angle Regression (LARS) [44] and Elastic Net [45] algorithms are well-studied in image processing including CBIR.

In this study, the LFDs and SR algorithms are used to build CBIR model and they are analyzed and tested in detail. First, a framework for CBIR is built to analyze the performance of LFDs and SR algorithms. The framework is visualized in Fig. 1. SIFT, SURF, HoG, LBP and LTP LFDs are used for feature extraction step. Random Features, K-Means and K-SVD algorithms build the dictionary on DL step. OMP, Homotopy, Lasso, Elastic Net, PCD and SSF are used to obtain sparse coefficients on CL step.

Then, the SR algorithms Feature-based Sparse Representation (FBSR), Online Dictionary Learning (ODL) and Locality-constrained Linear Coding (LLC) that are presented in [24,48] and [49] are compared. FBSR is proposed not only for image retrieval but image similarity assessment, image copy detection and image recognition, too [24]. LLC algorithm is used in [29,30] and ODL is used in [36] for CBIR. Coil20 [50] and Corel1000 [51] data sets are used to evaluate the performance of all tests.

2. Local feature descriptors

In the simplest term, methods used to solve computer vision problems such as object recognition, classification and retrieval are comparing images. The information called features in images are extracted to realize this comparison. LLFs of images were used at the early years of the computer vision. These features are color, shape, texture and spatial layout information and called global features [52]. Nevertheless, these are invariant to translation, rotation, scaling and affine deformation [53]. Furthermore, these are insufficient for recognition of iterant objects. Therefore, local descriptors were proposed to overcome these problems in the later years [54]. These descriptors are in interest with local structures in the images. These structures are robust to process on image and preferred for computer vision problems [55]. These descriptors are invariant to scaling, rotation and partial invariant to illumination and 3D camera view. Hence illumination changes are still an open area to study. Color-based invariants could solve this problem [56]. A number of novel color-based invariants are proposed to investigate this issue [56]. The results show that they increase the performance of gray-scale based matching algorithms. Object recognition has a key role on CBIR especially when the model is region-based. The studies about this topic are usually texture based. But these methods fail when the salient point detection does not produce robust results for untextured objects. Although the Bag of Boundaries (BoB) is proposed to solve this problem, there are some weaknesses of this method [57]. Therefore, Arandjelovic points to weakness of BoB and proposes a sparse method called Bag of Normals (BoN) [57].

Local descriptors are discussed in three subsections named distribution-based, spatial-frequency and other descriptors [58]. Distribution-based local descriptors represent the images with using the mean histogram of the regions in image. SIFT, SURF, HoG, LBP and LTP descriptors are in this group.

2.1. SIFT

SIFT features are proposed by Lowe [12,59]. SIFT extracts features that are invariant to scaling, rotation and partial invariant to illumination and 3D camera view [12]. Furthermore, these are well oriented at spatial and frequency domain. With this, deterioration probability is lowered. Extracting of SIFT feature is done in four steps [12]. These are:

- Scale-space extrema detection