Image and Vision Computing 42 (2015) 1–12

Contents lists available at ScienceDirect

Image and Vision Computing

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

RSILC: Rotation- and Scale-Invariant, Line-based Color-aware descriptor☆

Sema Candemir *, Eugene Borovikov, K.C. Santosh, Sameer Antani, George Thoma

Lister Hill National Center for Biomedical Communications U. S. National Library of Medicine, National Institutes of Health, Bethesda, MD, USA

ARTICLE INFO

Article history:
Received 26 August 2014
Received in revised form 29 April 2015
Accepted 29 June 2015
Available online 15 July 2015

Keywords:
Image descriptor
Local features
Spatial features
Rotation invariance
Scale invariance
Color aware

ABSTRACT

Modern appearance-based object recognition systems typically involve feature/descriptor extraction and matching stages. The extracted descriptors are expected to be robust to illumination changes and to reasonable (rigid or affine) image/object transformations. Some descriptors work well for general object matching, but gray-scale key-point-based methods may be suboptimal for matching line-rich color scenes/objects such as cars, buildings, and faces. We present a Rotation- and Scale-Invariant, Line-based Color-aware descriptor (RSILC), which allows matching of objects/scenes in terms of their key-lines, line-region properties, and line spatial arrangements. An important special application is face matching, since face characteristics are best captured by lines/curves. We tested RSILC performance on publicly available datasets and compared it with other descriptors. Our experiments show that RSILC is more accurate in line-rich object description than other well-known generic object descriptors.

1. Introduction

Feature matching is an essential component of many modern computer vision applications, including near-duplicate detection [1], stereo correspondence [2], 3D modeling [3], image stitching [4], as well as face alignment and matching [5–7]. Scene and object matching methods in digital images can be roughly divided in the following major groups by the density of the image features they extract and use:

dense descriptor methods [8–10] tend to use all image information and often assume that all pixels in the image are equally important. Hence, they may be computationally expensive and require a high degree of correlation between the probe and gallery images. Typically, such methods are not very accurate given large variations in object pose, scale, and illumination.

sparse descriptor methods use non-dense image features (e.g., edges [11]) and/or various key-spots [12–14]. They are relatively robust to variations in pose, size, orientation, and lighting of the query image with respect to the objects in the gallery. They provide a sparse representation for objects and high-speed matching on the key locations that need to be automatically determined, which calls for some sort of a saliency map [15].

This implicit methodology division prompted some well-performing *hybrid* techniques that include features from both categories and typically fuse them in a weighted features ensemble [16–18], optimized for a particular application [19].

In content-based image retrieval, object detection, and face recognition (FR) in an unconstrained environment (typically outside of the studio), sparse descriptors are generally preferred because they are often more robust to deformations and lighting variations than dense feature methods. Key-spot-based matching involves detecting the key-spots, building local descriptors for each key-spot, and finding an aggregate distance of best matches. For pose- and lighting-robust matching, the descriptors should be robust to geometrical variations such as translation, rotation, scaling (and if possible to affine/projective transformations and photometric variations such as illumination direction, intensity, colors, and highlights. The selection of a robust key-spot (e.g., a point, a line, or a corner) depends on the image collection and the application.

In generic object matching, the coarse features from key-points may be adequate to find suitable matches. However, in line-rich scenes, some dominant lines on objects may provide more stable and discriminative features than key-points. Such line-rich scenes and objects are omnipresent and can be natural (e.g., landscapes, plants, animals, humans) or artificial (e.g., cities, cars, house exteriors and interiors, office spaces). Stable (but not necessarily rigid) characteristic lines in them can be used as good key-spot candidates, promising a more stable object matching ability than key-points can.

Human face matching/recognition (FM/FR) is an important special case of object matching that has been an active research and development area in academia and industry because of the wide variety of real-world applications, such as surveillance, visual authentication, human–machine interface, criminal identification, and commercial applications. It identifies individuals from face images or video sequences

☆ This paper has been recommended for acceptance by Qiang Ji.

* Corresponding author. Tel.: 1301 605 3389; e-mail: candemirsema@gmail.com.
E-mail addresses: sema.candemir@nih.gov (S. Candemir), eugene.borovikov@nih.gov (E. Borovikov), santosh.kc@nih.gov (K.C. Santosh), sameer.antani@nih.gov (S. Antani), george.thoma@nih.gov (G. Thoma).

http://dx.doi.org/10.1016/j.imavis.2015.06.010
0262-8856/

using computer vision and machine learning algorithms. The general procedure for the appearance-based face image retrieval (FIR) systems consists of detecting the faces, extracting the facial information, and comparing a query face descriptor with those in a database [7,20–24].

To deal with the variation in face appearance (e.g., unknown head poses, unexpected facial expressions, and unpredictable lighting), the extracted features are expected to be robust to illumination changes, distortion, and scaling. One can certainly use key-point-based descriptors (e.g., SIFT [12], SURF [13], ORB [14], but because faces are line-rich objects, it may be beneficial to introduce the notion of key-lines and their descriptors for better matching.

According to psychophysics and neuroscience studies, line-rich features, such as face outline, eyes, mouth, and hair, are most important for perceiving and remembering faces [25] by humans. Another study has investigated the importance of facial features for automatic face recognition [26] by extracting facial landmarks. Experimental results indicate that these facial features are indeed important for face identification. Several other studies [6,11,27] showed that the most informative face characteristics appear to come from the face lines that can model face features in a very intuitive, human-perceptible form.

Consider the face images shown in Fig. 1.a–b. The prominent characteristic parts are marked by lines/curves, whose local regions and their spatial arrangements on a face can be used for robust matching. The human-perceptible important face lines overlap very well with the machine-computed edge maps of the faces on the CalTech set [28], whose cumulative distribution is also shown. The lines are mostly located on the prominent face characteristics (landmarks such as eye, mouth, nose, and face shape), which are the discriminative locations of a face (Fig. 1.c.). All these studies and illustrations indicate that lines with their descriptors can provide more stable recognition features for face matching.

We propose a general-purpose key-line descriptor that is color aware, invariant to rotation/scale, and is robust to illumination changes. To increase the discriminative power of the descriptor, we combined color/texture information of the local regions and added the relational information of the other key-lines. We tested our descriptor matching power on publicly available datasets containing unconstrained images of faces and general objects. We compared the new descriptor to well-known descriptors (in the same test-bed system), and our experimental results show that the RSILC descriptor is robust to rotation, scale, reflection and illumination and produces more accurate matches in face and object retrieval applications.

2. Relevant publications

Many different techniques for modeling local image regions have been developed. Scale-Invariant Feature Transform (SIFT) [12,29] is one of the most robust key-point descriptors among the local feature descriptors with respect to different geometrical changes [30]. It detects notable and stable key-points for images at different resolutions and produces scale- and rotation-invariant descriptors for robust matching. Several papers have been published on SIFT-based face recognition [31–33]. Although SIFT originally was designed for gray-scale images, there are several extensions to make use of the color information in the descriptors [34–36]. One of the successful attempts is colored SIFT (CSIFT) [36] which embeds the color information through the gradient of color invariants instead of using gray-scale gradients as in conventional SIFT. Defining the descriptor in color space makes the descriptor more robust with respect to color variations.

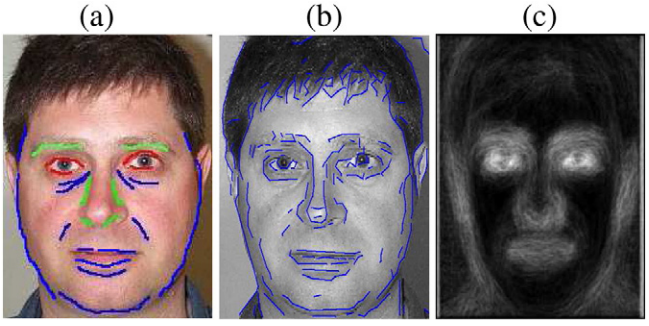

Fig. 1. Line/curve features that may be characteristic of a specific face: (a) human-perceptible, (b) line edge map (LEM) [11], (c) average line map of faces in the CalTech set [28].

Another well-known key-point descriptor is Speeded Up Robust Features (SURF), which provides a quicker way to detect key-points and compute descriptors that are rotation and scale invariant as well as robust to illumination changes [37]. SURF is less accurate than SIFT, but it has been successfully used in many practical applications, including face/components matching [38,39].

The mentioned key-point descriptors (e.g. SIFT and SURF), being robust to various affine transformations and lighting, are widely used for object detection and recognition. However, they typically contain information that is local to their key-points, which prompts some false-positive correspondences when performing many-to-many matches. This problem could be remedied by considering key-point spatial relationships, (e.g., having each descriptor record other key-points' azimuth angles much like shape context [40]), hence capturing not only local context of each key-spot but also their global spatial relationships. Knowing the usefulness of spatial relations in image understanding [41,42], several state-of-the-art methods have been reported together with the use of application-dependent local features. For example, the authors in [43,44] integrate spatial distribution of key-points by using shape context with texture features for food classification. In a similar fashion, spatial relations between the visual primitives (such as circles and corners) are integrated with statistical shape features for graphics recognition [45].

The Pyramid of Histograms of Orientation Gradients (PHOG) descriptor represents an image by its local shape and the spatial layout of shape information [46]. The local shape is modeled by a histogram of edge orientations. The spatial layout is represented by tiling the image into the regions at multiple resolutions. The final descriptor vector is the concatenation of histograms at each resolution. The descriptor is robust to scaling as long as the object position and orientation remain the same, but it is rotation dependent and color-blind.

As an alternative to key-points, another important set of features for object matching can be collected from edges, which provide the advantages of a lesser demand on the storage space and a lower sensitivity to illumination changes. Gao and Leung [11] describe a face recognition method using line edge maps (LEMs). The system extracts the lines from the edge map of face images and compares their similarity using the Hausdorff distance. LEM produces fast and reliable matching on aligned faces but does not use the region around the lines, which contains intensity information that helps to discriminate the lines and reduce mismatches. Gao and Qi [47] extend the LEM approach by considering corner points along the edge lines and show their method's robustness to scale as well as the superior one-image-per-person retrieval capability compared to the eigenfaces [48]. Deboeverie et al. [6] combined the curve edge map with the relative positions and intensity information around the curves. This system uses the orientation of the main axis of the curve segments for the first match. Then, it considers the histograms of inner and outer sides of the curve and relative positions of curve segments to refine the matching stage. However, this method lacks color information for the local regions.

Liu et al. [16] propose SIFT flow to align an image to its nearest neighbors in a photo gallery. This hybrid method matches densely sampled, pixel-wise SIFT features between two images while preserving (sparse) spatial discontinuities, matching a query object located at different parts of the scene. Experiments show that the proposed approach robustly aligns complex scene pairs containing significant spatial differences. The applications include single image motion field prediction/