Pattern Recognition 48 (2015) 1174–1184

Contents lists available at ScienceDirect

Pattern Recognition

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

A unified framework for local visual descriptors evaluation

Olivier Kihl [a,*], David Picard [a], Philippe-Henri Gosselin [a,b]

[a] ETIS/ENSEA – Université Cergy-Pontoise, CNRS, UMR 8051, 6 avenue du Ponceau, CS 20707 CERGY, F 95014 Cergy-Pontoise Cedex, France
[b] INRIA Rennes Bretagne Atlantique, Campus de Beaulieu, 35042 Rennes Cedex, France

ARTICLE INFO

Article history:
Received 29 November 2013
Received in revised form
7 August 2014
Accepted 20 November 2014
Available online 28 November 2014

Keywords:
Image processing and computer vision
Vision and scene understanding
Video analysis
Image/video retrieval
Object recognition
Feature representation

ABSTRACT

Local descriptors are the ground layer of recognition feature based systems for still images and video. We propose a new framework for the design of local descriptors and their evaluation. This framework is based on the descriptors decomposition in three levels: primitive extraction, primitive coding and code aggregation. With this framework, we are able to explain most of the popular descriptors in the literature such as HOG, HOF or SURF. This framework provides an efficient and rigorous approach for the evaluation of local descriptors, and allows us to uncover the best parameters for each descriptor family. Moreover, we are able to extend usual descriptors by changing the code aggregation or adding new primitive coding method. The experiments are carried out on images (VOC 2007) and videos datasets (KTH, Hollywood2, UCF11 and UCF101), and achieve equal or better performances than the literature.

1. Introduction

Most multimedia retrieval systems compare multimedia documents (image or video) thanks to three main stages: extract a set of local visual descriptors from the multimedia document; learn a mapping of the set of descriptors into a single vector to obtain a signature; compute the similarity between signatures. In this paper, we focus on the computation of visual descriptors. The main goal of local visual descriptors is to extract local properties of the signal. These properties are chosen so as to capture discriminative characteristic atoms in images or videos. Since local descriptors are the ground layer of recognition systems, efficient descriptors are necessary to achieve good accuracies. Such descriptors have become essential tools in still image classification [1,2] and video action classification [3–5].

The main contribution of this paper is a unified framework for visual descriptors evaluation that includes all the usual descriptors from the literature such as SIFT (scale-invariant feature transform) [6], SURF (speeded up robust features) [7], HOG (histogram of oriented gradient) [8], HOF (histogram of oriented flow) and MBH (motion boundary histogram) [9]. This framework is based on the decomposition of the descriptor in three levels: primitive extraction, primitive coding and code aggregation. Each popular descriptor is composed by a given primitive, a given coding and a given aggregation. Following these principles, we are able to perform a rigorous evaluation of many common descriptors and pinpoint which of the primitive, coding or aggregation is the source of their effectiveness. The consequence is that this evaluation allows us to improve our understanding of local descriptors. For example, we are able to show that the best coding method so far for the motion and gradient of motion primitives is a rectification, contrarily to the widely used orientation coding like in the well known HOF and MBH. Moreover, our framework allows the design of novel, more efficient and complementary descriptors, which we take as our second contribution. Using the framework as a method to explore the possible combinations of primitive, coding and aggregation allows us to precisely know the gain of each changed step compared to existing descriptors. For example, we are able to propose new descriptors based on oscillating functions aggregation which achieve the best performances for single descriptors once combined with the relevant primitive and coding steps. By knowing precisely the improvement caused by each changed step, we believe subsequent research can efficiently be focused on the steps where much of the gain is expected.

The paper is organized as follows. In Section 2, we present the most popular descriptors in the literature, for still images and for human action videos. Then, in Section 3, we present our framework, explain the most popular descriptors, and extend them by modifying some of these three steps. In Section 4, we propose an evaluation of the framework hyperparameters and parameters on one still image classification dataset and on two action classification datasets. Finally, in Section 5, we compare our results with the literature on one still image classification dataset and four action

* Corresponding author. Tel.: +33 1 30 73 66 10; fax: +33 1 30 73 66 27.
E-mail addresses: olivier.kihl@ensea.fr (O. Kihl), picard@ensea.fr (D. Picard), philippe-henri.gosselin@ensea.fr (P.-H. Gosselin).

http://dx.doi.org/10.1016/j.patcog.2014.11.013
0031-3203/

classification datasets according to the bests descriptors of our evaluation.

2. Related work

In this section, we present the most popular descriptors in the literature, first for still image and then for human action video.

2.1. Still image descriptors

In the past 10 years, several descriptors have been proposed for key-points matching and successfully used for still image classification. The most commonly used are SIFT [6], SURF [7] and histogram of oriented gradient (HOG) [9]. These descriptors are something referred to as "edge descriptors" since they mostly consider the spatial repartition of gradient vectors around the keypoint. SIFT and SURF are both interest points detector and local image descriptor. In this paper, we only consider the descriptors.

Several descriptors have been proposed with the aim to decrease the computation time without loss of performance, for example SURF [7] and GLOH [10]. Similarly, Daisy [11] is a SIFT like descriptor designed to be faster to compute in the case of dense matching extraction.

2.2. Action descriptors

In the early work on action recognition, silhouette based descriptors were used. These descriptors are computed from the evolution of a silhouette obtained by background subtraction methods or by taking the difference of frames (DOF). The main silhouette based descriptors are "motion energy image" (MEI) [12], "motion history image" (MHI) [12], the "average motion energy" (AME) and the "mean motion shape" (MMS) [13]. In [14] Kellokumpu et al. use histograms of "local binary patterns" (LBP) [15] to model the MHI and MEI images. As time is an important information in video, Gorelick et al. [16,17] study the silhouettes as space–time volumes. Space–time volumes are modeled with Poisson equations. From these, they extract seven spatio-temporal characteristic components. The main drawback of all these methods is the computation of silhouettes. Indeed, this computation is not robust, making these methods only relevant in controlled environments such as the Weizmann dataset [16] and the KTH dataset [5]. As a result, they tend to fail on more realistic data-sets such as UCF11 [18] or Hollywood2 [4] datasets.

Assuming that action recognition is closely linked to the notion of movement, many authors have proposed descriptors modeling of the optical flow motion field [19–24]. The descriptor proposed by Ali and Shah [24] is based on the computation of many kinematic features on the motion field. Descriptors based on a polynomial approach for modeling global optical flow are proposed in [25,26]. In [27] a local space–time descriptor based on polynomial approximation is proposed. It is named series of polynomial approximation of flow (SoPAF).

Finally, the most successful descriptors developed in recent years are extensions to video of the HOG [8] still image descriptors. The most commonly used are the histogram of oriented flow (HOF) [9] and the motion boundary histogram (MBH) [9]. HOF is the same as HOG but is applied to optical flow instead of gradient. The MBH models the spatial derivatives of each component of the optical flow vector field with a HOG. In this context, several extensions of still image descriptors have been proposed, such as cuboid [28], 3DHOG [29], 3D-SIFT [30], and ESURF [31].

Recently, Wang et al. [3] propose to model these usual descriptors along dense trajectories. The time evolution of trajectories, HOG, HOF and MBH, is modeled using a space time grid following pixels trajectories. The use of dense trajectories for descriptor extraction tends to increase the performances of popular descriptors (HOG, HOF and MBH).

3. Primitive/coding/aggregation framework

In this section, we present the main contribution of this paper. We propose a framework providing a formal description of the steps needed to design local visual descriptors. Our framework splits descriptors extraction in three levels: primitive extraction, primitive coding and code aggregation. These three steps can be seen as hyperparemeters of a descriptor.

3.1. Primitive extraction

At the primitive level, we extract a specific type of low-level information from an image or a video. The objective is to extract local properties of the signal. Generally, it relies on a high frequency filtering, linear for gradient or non-linear in the case of motion (optical flow), filters banks such as Haar (SURF), easy extension of popular filters [32], or non-linear operators. The primitive extraction induces a choice in relevant information and introduces data loss. Such primitives include the gradient (SIFT [6], HOG [8] and Daisy [11]), the responses to 2D Haar-wavelets (SURF [7]), the motion flow (HOF [9], SoPAF [27]), or the gradient of motion flow (MBH [9]). In Fig. 1, we show three examples of primitive used in the literature, the gradient, the motion flow and the gradient of the motion flow.

3.2. Primitive coding

The primitive coding corresponds to a non-linear mapping of the primitive to a higher dimensional space. The objective is to

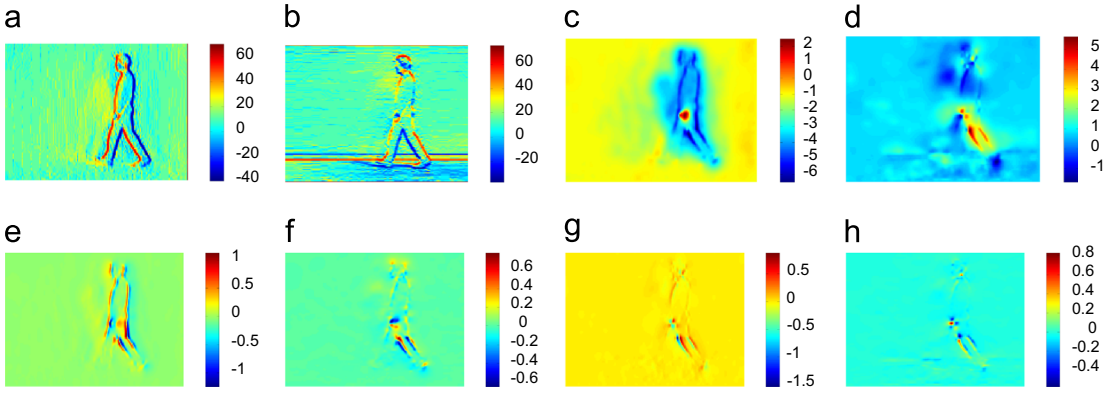

Fig. 1. Example of primitive; (a) horizontal gradient; (b) vertical gradient; (c) horizontal motion flow; (d) vertical motion flow; (e) horizontal gradient of horizontal motion flow; (f) vertical gradient of horizontal motion flow; (g) horizontal gradient of vertical motion flow; and (h) vertical gradient of vertical motion flow.