Information Sciences 429 (2018) 37–48

Contents lists available at ScienceDirect

Information Sciences

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

Perceptual multi-channel visual feature fusion for scene categorization

Xiao Sun[a], Zhenguang Liu[b], Yuxing Hu[c,*], Luming Zhang[a], Roger Zimmermann[b]

[a] School of Computer and Information, Hefei University of Technology, China
[b] School of Computing, National University of Singapore, Singapore
[c] School of Aerospace Engineering, Tsinghua University, China

ARTICLE INFO

Article history:
Received 9 December 2016
Revised 21 September 2017
Accepted 28 October 2017

Keywords:
Image kernel
Feature fusion
Scene categoriztion
Perception

ABSTRACT

Effectively recognizing sceneries from a variety of categories is an indispensable but challenging technique in computer vision and intelligent systems. In this work, we propose a novel image kernel based on human gaze shifting, aiming at discovering the mechanism of humans perceiving visually/semantically salient regions within a scenery. More specifically, we first design a weakly supervised embedding algorithm which projects the local image features (*i.e.*, graphlets in this work) onto the pre-defined semantic space. Thereby, we describe each graphlet by multiple visual features at both low-level and high-level. It is generally acknowledged that humans attend to only a few regions within a scenery. Thus we formulate a sparsity-constrained graphlet ranking algorithm which incorporates visual clues at both the low-level and the high-level. According to human visual perception, these top-ranked graphlets are either visually or semantically salient. We sequentially connect them into a path which mimics human gaze shifting. Lastly, a so-called gaze shifting kernel (GSK) is calculated based on the learned paths from a collection of scene images. And a kernel SVM is employed for calculating the scene categories. Comprehensive experiments on a series of well-known scene image sets shown the competitiveness and robustness of our GSK. We also demonstrated the high consistency of the predicted path with real human gaze shifting path.

1. Introduction

Scene categorization is an important module in a large body of modern intelligent systems [8,42–44,49,64], *e.g.*, image annotation and multimedia management. This technique is helpful to support likely scene configurations and reject unlikely ones. As an example example, a real-world scene annotation system should encourage the occurrence of a parking lot in an aerial photograph belonging to “downtown area” while suppressing the occurrence of an airport in an aerial photograph belonging to “residential area”. Nevertheless, effectively predicting sceneries from a vast number of categories remains a tough task due to the following technical challenges:

- Experiments in eye tracking have demonstrated the fact that humans allocate gazes selectively and sequentially to visually/semantically important regions within a scenery. As can be seen from Fig. 1, most observers will start with viewing

* Corresponding author.
E-mail address: sunx@hfut.edu.cn (X. Sun).

https://doi.org/10.1016/j.ins.2017.10.051
0020-0255/

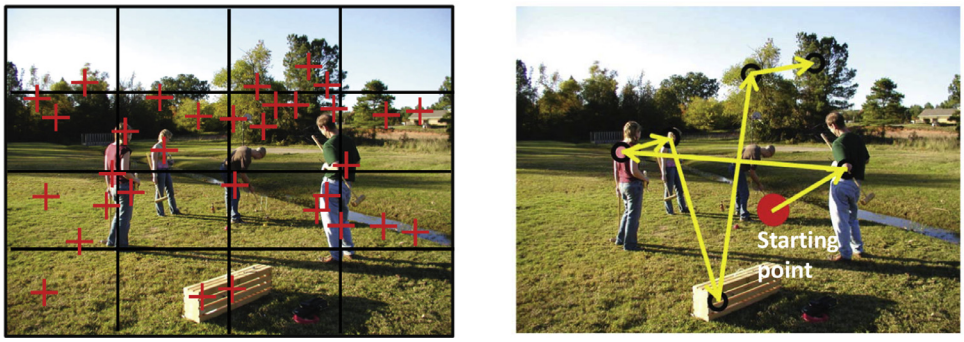

Fig. 1. A comparison between the image kernels calculated based on spatial pyramid matching (left) and gaze shifting path (right), wherein the scattered red crosses indicate the local image features. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

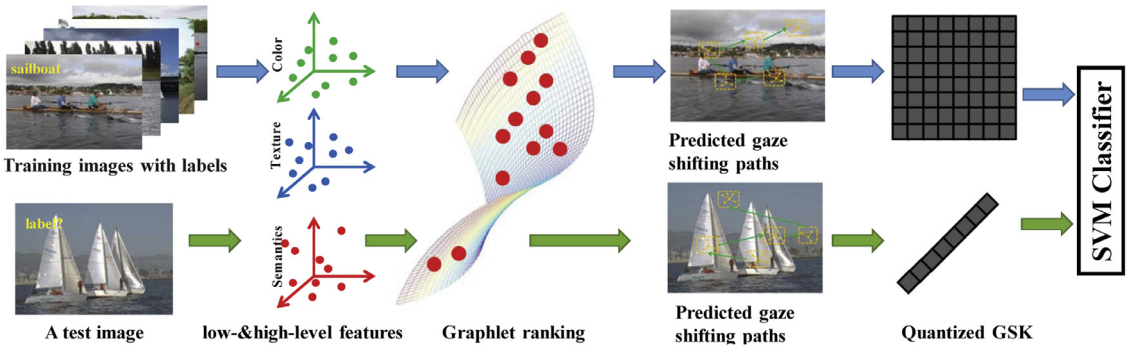

Fig. 2. The flowchart of our proposed gaze shifting kernel (GSK).

the player in the center, afterward shifting their gazes to the remaining three players , and finally to the bench as well as background trees. It is obvious that such directed viewing path is informative to distinguish scene images from different categories. But this clue has not been well modeled in the existing models. As shown on the left of Fig. 1, existing recognition models usually encode local image descriptors unorderly for scene categorization.

- Both biological and psychological researches have illustrated that the bottom-up and top-down visual features draw the attention of human eye. In practice, a successful scene categorization model should seamlessly encode both the low-level and the high-level visual features. But, typically multiple features are fuzed in a linear or nonlinear manner. The underlying cross-feature information may not exploited optimally. Even worse, such fusion schemes are not designed with the objective of maximally capturing human visual perception.

To solve or at least alleviate the above problems, this paper proposes gaze shifting kernel (GSK) to reflect human gaze allocation process. We focus on formulating a sparsity-constrained ranking algorithm which jointly optimizes the graphlet weights from multiple visual channels. An overview of our proposed GSK is elaborated in Fig. 2. In particular, we first transfer the semantics of image labels into a number of graphlets by designing a manifold embedding algorithm, wherein each graphlet is represented by a combination of low-level and high-level visual features. Afterward, we formulate a sparsity-constrained algorithm to encode the multiple features for deriving the saliency of each graphlet. More specifically, we establish the matrices describing the visual/semantic features of graphlets inside each scene image. Thereby, we propose a ranking algorithm which seeks the consistently sparse elements by jointly decomposing the multiple-feature matrices into pairwise low-rank and sparse matrices. In comparison with the conventional algorithms linearly/non-linearly combining multiple global features, multiple visual/semantic features are seamlessly integrated for discovering salient graphlets. We link the detected graphlets into a path to mimic the process of human gaze shifting. Noticeably, the obtained paths are 2-D planar features which cannot be directly fed into a conventional classifier like SVM. Therefore, we quantize them into a kernel, which is subsequently integrated into a kernel SVM for scene categorization.

The key contributions of this article can be summarized as follows. 1) We proposed a novel image kernel which well describes human gaze allocation in order to distinguish sceneries from different categories. 2) We developed a sparsity-constrained ranking algorithm which can intelligently detect visually/semantically significant graphlets capturing the attention of human eye. 3) We carried out extensive experimental validations on five popular scene image sets to demonstrate the superiority of our method.