Pattern Recognition 52 (2016) 260–273

Contents lists available at ScienceDirect

Pattern Recognition

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

Semantic segmentation of images exploiting DCT based features and random forest

D. Ravì [a,*], M. Bober [b], G.M. Farinella [a], M. Guarnera [c], S. Battiato [a]

[a] Image Processing Laboratory, Dipartimento di Matematica e Informatica, University of Catania, Italy
[b] Center for Vision, Speech and Signal Processing, University of Surrey, UK
[c] Advanced System Technology - Computer Vision, STMicroelectronics, Catania, Italy

ARTICLE INFO

Article history:
Received 29 November 2014
Received in revised form
15 October 2015
Accepted 31 October 2015
Available online 7 November 2015

Keywords:
Semantic segmentation
Random forest
DCT
Textons

ABSTRACT

This paper presents an approach for generating class-specific image segmentation. We introduce two novel features that use the quantized data of the Discrete Cosine Transform (DCT) in a Semantic Texton Forest based framework (STF), by combining together colour and texture information for semantic segmentation purpose. The combination of multiple features in a segmentation system is not a straightforward process. The proposed system is designed to exploit complementary features in a computationally efficient manner. Our DCT based features describe complex textures represented in the frequency domain and not just simple textures obtained using differences between intensity of pixels as in the classic STF approach. Differently than existing methods (e.g., filter bank) just a limited amount of resources is required. The proposed method has been tested on two popular databases: CamVid and MSRC-v2. Comparison with respect to recent state-of-the-art methods shows improvement in terms of semantic segmentation accuracy.

1. Introduction and motivations

Nowadays a wide range of applications including medical, robotics and automotive, require the ability to automatically understand the real world. Examples of these applications are a smart cars able to recognize and eventually help a careless driver, to detect a pedestrian crossing the street. Another example is a smart system that during a surgery operation is able to drive the surgeon on the localization of the tumour area and steer him in the removal process of that area. Last but not least a surveillance system that can analyze and recognize automatically what is going in the world from a recorded video. Electronic devices with the ability to understand the real world from images are called intelligent systems with semantic segmentation. The semantic segmentation, can be thought as an extension of the popular scene classification problem where the entity to classify is not anymore the whole image but single group of pixels [1]. It aims at pixel-wise classification of images according to semantically meaningful regions (e.g., objects). A precise automated image segmentation is still a challenging and an open problem. Among others, local structures, shape, colour and texture are the common features deployed in the semantic segmentation task. Colour or gray level information is essential core features used to segment images into regions [2,3]. An efficient and computationally light descriptor to build on colour features is the colour histogram. The histogram ignores the spatial organization of the pixels, which is generally an advantage as it supports rotation and scale invariance. When spatial organization is required a second order statistics can be used. An example is image correlograms [4] that describes the correlation of the image colours as a function of their spatial distance. Local structures (i.e., edges, corners, and T-Junctions) are also useful features that are detected by differential operators commonly applied to the luminance information. The shape is one of the most important characteristic of an object and allows us to discriminate different objects. Finally texture is a visual cue that describes the luminosity fluctuations in the image, which let us interpret a surface as a whole part. Textures can be characterized using properties such as regularity, coarseness, contrast and directionality and contain also important information about the structural arrangement of the surface. It also describes the relationship of the surface to the surrounding environment. One immediate application of image texture is the recognition of image regions using texture properties. Texture features can be extracted by using various methods. Gray-level occurrence matrices (GLCMs) [5], Gabor Filter [6], and Local Binary Pattern (LBP) [7] are examples of popular methods to extract texture features. Other

* Corresponding author. Tel.: 095 7337219
E-mail addresses: ravi@dmi.unict.it (D. Ravì), m.bober@surrey.ac.uk (M. Bober), gfarinella@dmi.unict.it (G.M. Farinella), mirko.guarnera@st.com (M. Guarnera), battiato@dmi.unict.it (S. Battiato).

http://dx.doi.org/10.1016/j.patcog.2015.10.021
0031-3203/

methods to obtain texture features are the fractals representation [8] and Textons [9].

The key step to obtain a reliable semantic segmentation system is the selection and design of robust and efficient features that are capable of distinguishing the predefined pixels' classes, such as grass, car, and people. The following criteria should be taken into account while considering the design of a system and the related features extraction method:

- Similar low-level features response can represent different objects as part of objects. Each single feature cannot be hence adequate for segmenting, in a discriminative way, the object that they belong to. A spatial arrangement of low-level features increases the object discrimination.
- A semantic segmentation approach cannot be generic because is strongly related to the involved application both in terms of requirements and input data types. Some examples of different domains include the segmentation of images obtained from fluorescence microscope, video surveillance cameras and photo albums. Another important parameter that is application dependent is for example the detail coarseness of required segmentation.
- The information needed for the labelling of a given pixel may come from very distant pixels. The category of a pixel may depend on relatively short-range information (e.g., the presence of a human face generally indicates the presence of a human body nearby), as well as on very long-range dependencies [10].
- The hardware miniaturization has reached impressive levels of performance stimulating the deployment of new devices such as smart-phones and tablets. These devices, though powerful, do not have yet the performance of a typical desktop computer. These devices require algorithms that perform on board complex vision tasks including the semantic segmentation. For these reasons, the segmentation algorithms and related features should be designed to ensure good performance for computationally limited devices [11].

The first contribution of this paper is the design of new texture features pipeline, which combine colour and texture clues in more efficient manner with respect to other methods in literature (e.g., convolutional network). Secondly, we propose texture features based on DCT coefficients selected through a greedy fashion approach and suitably quantized. These DCT features have been exploited in [12] and successfully applied for the scene classification task making use of their capability to describe complex textures in the frequency domain maintaining a low complexity. Other approaches usually compute similar features using bank of filter responses that drastically increases the execution time. As in [12] our texture information is extracted using the DCT module that is usually integrated within the digital signal encoder (JPEG or MPEG based). The proposed features are then used to feed a Semantic Texton Forest [13] that has been showed to be a valid baseline approach for the semantic segmentation task.

The rest of the paper is organized as follows: Section 2 discusses the state-of-the-art approaches, whereas Section 3 describes the random forest algorithm and how to add the novel features in the STF system. Section 4 presents the pipeline of the proposed approach. Section 5 introduces the extraction pipelines for each proposed features. Section 6 describes the experimental settings and the results. Finally, Section 7 concludes the paper.

2. Related works

To address the challenges described above, different segmentation methods were proposed in literature. Some basic approaches segment and classify each pixel in the image using a region-based methodology as in [14–22]. Other approaches use a multiscale scanning window detector such as Viola-Jones [23] or Dalal-Triggs [24], possibly augmented with part detectors as in Felszenszwalb et al. [25] or Bourdev et al. [26]. More complex approaches as in [27,28] unify these paradigms into a single recognition architecture, and leverage on their strengths by designing region-based specific object detectors and combining their outputs. By referring to the property that the final label of each pixel can be dependent by the labels assigned to other pixels in the image, different methods use probabilistic models such as the Markov Random Field (MRF) and the Conditional Random Fields (CRF) that are suitable to address label dependencies. As example, the nonparametric model proposed in [29] requires no training and can easily scaled to datasets with tens of thousands of images and hundreds of labels. It works by scene-level matching with global image descriptors, followed by superpixel-level matching with local features and efficient MRF based optimization for incorporating neighbourhood context. In [30], instead, a framework is presented for semantic scene parsing and object recognition based on dense depth maps. Five view independent 3D features that vary with object class are extracted from dense depth maps at a superpixel level for training a randomized decision forest. The formulation integrates multiple features in the MRF framework to segment and recognize different object classes. The results of this work highlight a strong dependency of accuracy from the density of the 3D features. In the TextonBoost technique [31] the segmentation is obtained by implementing a CRF and features that automatically learn layout and context information. Similar features were also proposed in [32], although Textons were not used, and responses were not aggregated over a spatial region. In contrast with these techniques, the shape context technique in [14] uses a hand-picked descriptor. In [33] a framework is presented for pixel-wise object segmentation of road scenes that combines motion and appearance features. It is designed to handle street-level imagery such as that on Google Street View and Microsoft Bing Maps. The authors formulate the problem in the CRF framework in order to probabilistically model the label likelihoods and a prior knowledge. An extended set of appearance-based features is used, which consists of Textons, colour, location and Histogram of Gradients (HOG) descriptors. A novel boosting approach is then applied to combine the motion and appearance-based features. The authors also incorporate higher order potentials in the CRF model, which produce segmentations with precise object boundaries. In [34] a novel formulation is proposed for the scene-labelling problem capable to combine object detections with pixel-level information in the CRF framework. Since object detection and multi-class image labelling are mutually informative dependent problems, pixel-wise segmentation can benefit from the powerful object detectors and vice versa. The main contribution of [34] lies in the incorporation of top-down object segmentations as generalized robust potentials into the CRF formulation. These potentials present a principled manner to convey soft object segmentations into a unified energy minimization framework, enabling joint optimization and thus mutual benefit for both problems. A probabilistic framework is presented in [35] for reasoning about regions, objects, and their attributes such as object class, location, and spatial extent. The proposed CRF is defined on pixels, segments and objects. The authors define a global energy function for the model, which combines results from sliding window detectors and low-level pixel-based unary and pairwise relations. It addresses the problems of what, where, and how many by recognizing objects, finding their locations and spatial extent and segmenting them. Although the MRF and the CRF are adequate models to deal with the semantic segmentation problem in terms of performance, they represent a bottleneck in the computation, because the inference is a highly resources consuming process. A powerful approach with