Neurocomputing 167 (2015) 658–670

Contents lists available at ScienceDirect

Neurocomputing

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

Likelihood-based feature relevance for figure-ground segmentation in images and videos

Mohand Saïd Allili [a,*], Djemel Ziou [b]

[a] Université du Québec en Outaouais, Département d'informatique et d'ingénierie, Gatineau, QC, Canada
[b] Université de Sherbrooke, Département d'informatique, Sherbrooke, QC, Canada

ARTICLE INFO

Article history:
Received 6 October 2014
Received in revised form
25 December 2014
Accepted 9 April 2015
Communicated by Luming Zhang
Available online 6 May 2015

Keywords:
Figure-ground segmentation
Feature relevance
Positive and negative examples
Gaussian mixture models (GMMs)
Level sets

ABSTRACT

We propose an efficient method for image/video figure-ground segmentation using feature relevance (FR) and active contours. Given a set of positive and negative examples of a specific foreground (an object of interest (OOI) in an image or a tracked objet in a video), we first learn the foreground distribution model and its characteristic features that best discriminate it from its contextual background. For this goal, an objective function based on feature likelihood ratio is proposed for supervised FR computation. FR is then incorporated in foreground segmentation of new images and videos using level sets and energy minimization. We show the effectiveness of our approach on several examples of image/video figure-ground segmentation.

1. Introduction

Object segmentation in images/videos (also called figure-ground segmentation) is important for several applications, such as content-based image/video retrieval (CBIVR) [9,40], automatic image/video annotation [30,36], object-based video coding [41], image/video retargeting [6], robotics and activity recognition [1,34]. In CBIVR and image/video annotation, for example, knowing image/video object content is of prominent importance to enhance the accuracy of semantic labeling of images and videos and answering user queries. Also, newly established multimedia standards for video coding (e.g., MPEG) are based on object content of videos. Therefore, efficient figure-ground segmentation is a critical issue for these applications.

Object segmentation in images is a very challenging problem due to several difficulties such as non-uniform illumination, image clutter and variability within object categories [38]. In the past, approaches have been proposed to tackle these difficulties by using either local or global information (or their fusion) for object segmentation. *Bottom-up* approaches group local cues (e.g., contours, color, texture) to form homogenous regions which can be used to build objects. Popular grouping algorithms are finite mixture models (FMM) [2,9] and graph-cuts [17,37,45]. Based on obtained homogenous regions, some approaches identify foreground objects (resp. backgrounds) through an interactive process exploiting the user's feedback [7,14]. However, these approaches suffer from over/under-segmentation where object parts may be merged with the background and vice versa [2,9]. Also, the need for user interaction with each image limits their usage in large-scale image segmentation. Unlike *bottom-up* approaches which consider local image properties regardless of spatial layout of the segmented object, *top-down* approaches rely on the representation of the global form of objects. These include mainly deformable templates [33], which are also applied as part-based representation models. Templates can be either simple geometrical elements (e.g., ellipses, rectangles, arcs, etc.) [18] or active contours (e.g., the *Snake* model [4]), which are evolved using energy minimization [10,33]. The main difficulty in *top-down* approaches lies in segmenting highly deformable objects, in addition to the need for accurate initialization to compensate for a difficult minimization problem [10,17,33].

Recently, several methods have attempted to combine the advantages of *bottom-up* and *top-down* approaches to achieve better object segmentation. In [8,26], for example, overlaps between segmented images and object fragments are used for object segmentation. The object fragments are usually extracted from a learning set of gray-scale or binary segmented images. However, since each fragment is considered independently, the approach is prone to include object parts in the background. Besides, its computational complexity increases exponentially with the number of fragments and explored image positions. Finally, approaches based on figure-ground color

* Corresponding author.
E-mail addresses: mohandsaid.allili@uqo.ca (M.S. Allili), d.ziou@usherbrooke.ca (D. Ziou).

http://dx.doi.org/10.1016/j.neucom.2015.04.015
0925-2312/

statistical modeling (e.g., using Gaussians [33], mixture of Gaussians [35] or kernel methods [25]) have been proposed for object segmentation. In those approaches, however, segmentation success highly depends on how distinguishable an object is from the background. On the one hand, too many dimensions of noise necessarily overwhelm too few dimensions of signal [28,31]. On the other hand, backgrounds can often be highly correlated with the object (e.g., cars and roads, giraffes and grass, swans and water, etc.) [19]. In videos, objects usually lie against the same background over successive frames. Therefore, accurate figure-ground segmentation can be achievable knowing the most discriminative features that best separate objects from their contextual backgrounds.

In this paper, we propose a new framework combining object/background statistical modeling and *feature relevance* (FR) for efficient figure-ground segmentation in images and videos. For images, FR is computed for each object category by using a set of manually segmented images containing instances of that category (i.e., positive examples) and their contextual backgrounds (i.e., negative examples). Local features are automatically extracted from these images and their figure-ground discrimination power is determined by their likelihood ratio. Our object segmentation approach is formulated as an energy minimization problem and implemented using level sets [32]. An energy functional is proposed to fit figure-ground distributions and encode the contribution of each feature according to its discrimination power. The only assumption of our algorithm is that a segmented object lies in the center of attention of the image. A level set function is evolved from its initial position toward the object boundaries using Euler–Lagrange equations. Finally, an extension of our algorithm to video figure-ground segmentation is proposed. We show the performance of our approach on several figure-ground segmentation on real-world images and videos.

Fig. 1a and b summarizes the two steps composing our approach for figure-ground segmentation in images and videos: (1) *a learning step*: computes FR and figure-ground statistical models using training examples and (2) *a segmentation step*: segments objects in new images (resp. videos) using FR and active contours. Early results of this work have been published in [3]. Herein, we give a more in-depth theoretical analysis of the problem and thorough experiments for validation. We have also added a new section containing complexity analysis of the algorithm and a discussion for future improvements.

This paper is organized as follows: Section 2 presents our approach for FR computation. Section 3 presents our segmentation model with FR. Section 4 presents some experiments that validate the proposed approach. We end the paper with a conclusion and some future work perspectives.

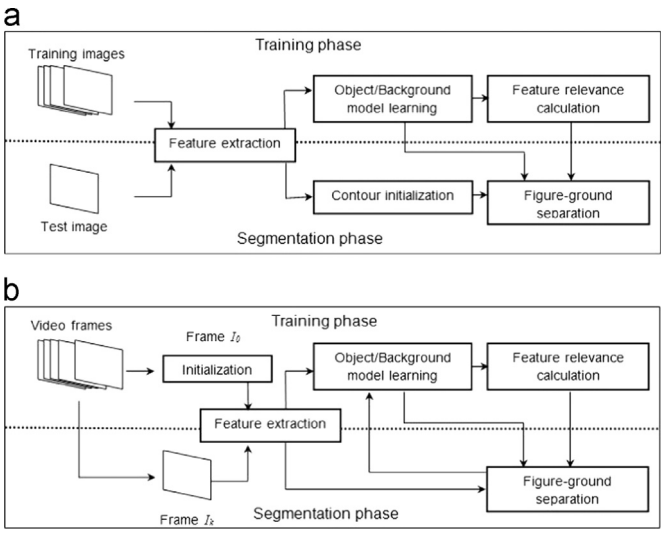

Fig. 1. Outline of our learning-based FR computation and its application to figure-ground segmentation in (a) images and (b) videos.

2. Feature relevance learning for figure-ground segmentation

2.1. Figure-ground distribution models in images

The essence of our approach is to achieve figure-ground segmentation using visual features that best discriminate between objects and their contextual backgrounds. For this goal, we exploit appearance patterns shared between instances of the same object category using a set of training images. These images are chosen to reflect the variety of contextual backgrounds the object may lie against (e.g., tigers in savanna, cars on roads, swans on water, etc.). To build a learning set for FR computation, we manually label locations as object/background in each training image. We call a "positive example" any location chosen in the object; otherwise it is referred to as a "negative example". For each location, we extract color, texture and gradient orientation features from patches centered around the location (see Fig. 2 for illustration). Suppose that we have D features $\{f_1,\ldots,f_D\}$ extracted in each location. Let $\mathcal{C}=\{\mathbf{e}_1,\ldots,\mathbf{e}_{n_1}\}$ and $\overline{\mathcal{C}}=\{\overline{\mathbf{e}}_1,\ldots,\overline{\mathbf{e}}_{n_2}\}$ be two sets of D-dimensional feature vectors extracted from positive and negative examples, respectively, and n_1 and n_2 are their cardinalities. Given that the elements in $\mathcal{C}$ and $\overline{\mathcal{C}}$ can be multi-modal (see for instance Fig. 4), we model feature distributions in $\mathcal{C}$ and $\overline{\mathcal{C}}$ using finite Gaussian mixture models (GMMs) [16].

In a similar way to the naive Bayes classifier [15], we suppose that the features are mutually independent in each class $\mathcal{C}$ and $\overline{\mathcal{C}}$. This is a reasonable assumption since it allows for assessing the discrimination power of each feature individually and reducing complexity and computation time of parameter estimation. Note that to enforce the independence assumption, one could perform *independent component analysis* (ICA) [24] on the features in a pre-processing step. We estimate the GMMs parameters using the *Expectation–Maximization* (EM) algorithm [16], where the number of clusters in each model is automatically determined using the *minimum message length principle* (MML) [39]. We recall that the MML is an information-theoretic principle that gives a good compromise between model complexity and goodness of fit to data [39]. It allows to obtain less complex models which have good fitting to object/background data. In what follows, we denote by $\vec{\theta}_d$ and $\vec{\omega}_d$ the GMM parameters computed for the dth feature using the data in $\mathcal{C}$ and $\overline{\mathcal{C}}$, respectively.

2.2. Feature relevance learning for figure-ground separation

Since we want our segmentation to be driven by the most discriminative features for each object category, we must determine in advance each feature discrimination power. Feature selection methods have been proposed in the past to enhance classification performance [22]. To determine feature subsets ensuring higher discrimination between classes of data, quantitative criteria can be used [28,29]. These criteria can be categorized by whether the evaluation process is data-intrinsic (filters) or classifier-dependent (wrappers). Since exhaustive search of subsets and their evaluation is time-consuming [20], we are constrained to consider simplified and non-exhaustive evaluation strategies to assess about features discrimination. For example, augmented variance ratio (AVR) has been shown to be effective for feature ranking [12]. Similar to Fisher discriminant analysis (FDA) [15], AVR maximizes the ratio between inter-class variance and within-class variance to estimate feature