Pattern Recognition 42 (2009) 2403–2420

Contents lists available at ScienceDirect

Pattern Recognition

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

Detection of unexpected multi-part objects from segmented contour maps

R. Bergevin*, J.-F. Bernier

Department of Electrical and Computer Engineering, Laval University, Quebec City, Canada G1K 7P4

ARTICLE INFO

Article history:
Received 3 July 2008
Received in revised form 18 February 2009
Accepted 29 March 2009

Keywords:
Multi-part object detection
Segmented contour map
Grouping constraints
Global shape grouping criteria

ABSTRACT

A novel method is proposed to detect multi-part objects of unknown specific shape and appearance in natural images. It consists in first extracting a strictly over-segmented map of circular arcs and straight-line segments from an edge map. Each obtained constant-curvature contour primitive has an unknown origin which may be the external boundary of an interesting object, the textured or marked region enclosed by that boundary, or the external background region. The following processing steps identify, in a systematic yet efficient way, which groups of ordered contour primitives form a complete boundary of proper multi-part shape. Multiple detections are ranked with the top boundaries best satisfying a combination of global shape grouping criteria. Experimental results confirm the unique potential of the method to identify, in images of variable complexity, actual boundaries of multi-part objects as diverse as an airplane, a stool, a bicycle, a fish, and a toy truck.

1. Introduction

Whether you are sitting looking at a picture or walking in a new place, your attention is often captured by some part of what is in your field of view. Attention is considered a key factor in cognitive science given the limited processing capabilities of humans [1], the quantity of information in the observed scenes, and the variety of perceptual and cognitive tasks and contexts. Attention is sometimes considered a purely reactive behavior where, for instance, a generic big, moving, red thing leads you to move your eyes in its direction. However, low-level, reactive behaviors are not sufficient to attain the sophistication made possible by higher-level cognitive capabilities. For instance, recognizing a known object would elicit a higher-level attention capture and a specific intelligent reaction.

Apart from these two extreme situations, other types of attention capture may also exist involving various forms and combinations of detection, localization, and recognition of objects at different levels of specificity. For instance, a widely discussed question about human vision is whether detection, localization, and recognition are sequential or parallel processes and, in the former case, what their temporal relationships are. The same questions arise in computer vision. Clearly, it is impossible to recognize a horse and still not know whether a horse is present in the field of view. However, as exemplified by recent research in computer vision, it could be possible to detect that a horse is present in the field of view without being able to localize it precisely. In that case, a specific model of a horse, e.g. a set of, possibly structured, local appearance features, is required a priori. In a different scenario, a four legged animal could also be detected and localized before its specific recognition as a horse in a given pose and posture. That scenario makes more sense from a complexity point of view since the number of known objects is huge and their possible appearances are also numerous. One way toward efficient high-level attention capture is to detect whether an unexpected but interesting object is present in the field of view and if yes, to recognize it more specifically with its pose and posture.

In this paper, a computer vision method for high-level attention capture is proposed on that basis. A major hypothesis made is that no explicit or specific model is available a priori for the shape and appearance of potential objects of interest. Hence, the addressed problem is the detection of unexpected but interesting objects. This differs from problems referred to as object (category) detection, localization, or recognition in the literature, where an explicit (possibly learned) object model of limited genericity is available a priori. One may argue that interest is often highly context-dependent. Humans may be assumed to be generally aware of context but not computer vision systems. Besides, if context is often an important factor, out-of-context vision is not uncommon even for humans. For instance, consider situations such as opening an image book, a photo album or a web image search result window, looking at the first shots of a movie in a theater, zapping television channels, or even awakening in the morning. Out-of-context detection and localization of unexpected or unknown objects by computer vision systems is to become more common as applications get more sophisticated and challenging. If specific context is not known, it means one either has to do

* Corresponding author. Tel.: +1 418 656 2131x5173; fax: +1 418 656 3159.
E-mail address: bergevin@gel.ulaval.ca (R. Bergevin).

0031-3203/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2009.03.028

without it or recognize it first. The latteris, however, a problem at least as difficult as object detection and localization itself. In this paper, specific context is assumed to be unknown and detection and localization of unexpected but interesting objects is sought for.

2. Related work

Model-based hypothesis-verification methods for object recognition match local image features to a known model in order to find the pose of a specific object [2]. Recent visual object categorization methods apply supervised statistical learning methods in order to build a more generic model for each of a number of known object categories [3–6]. Though local features such as interest points or contour fragments clearly bring useful information about image contents [7], their use in object categorization still consists in identifying features specific to a given class of object appearances. Despite rapid progress, state-of-the-art methods still have important limitations, namely their lack of a precise localization of the detected objects, their limited invariance to viewpoint, their possible confusion between different detectors responding positively to a test image, and their choice of categories which is rarely discussed and appears to be ad hoc. Scaling of the methods to a large number of categories is another important problem yet to be fully addressed.

As a result, the previous methods may hardly be seen as properly modeling high-level attention capture. In a sense, human attention capture implies that one does not have specific expectations and hence does not look or visually search for a particular object in a given pose but is instead surprised by what is observed. Hence, having a classifier for each type of possibly interesting object appearance and verifying in turn which one is in fact present in an image is unrealistic and inefficient for high-level attention capture. Detecting that an interesting but unexpected object is present in an image implies that it is also localized. Delimiting the region occupied by an interesting object in a static image is both useful and easy for humans. This is sometimes referred to as figure-ground discrimination. In computer vision, the related problem of partitionning an image into *object* and *background* regions is referred to as figure-ground segmentation. This is still a fundamental problem in computer vision with no existing general solution yet. The method proposed in this paper offers a possible solution to that problem for multi-part objects, a large and very generic class of objects.

State-of-the-art segmentation methods in computer vision rely on very generic contextual knowledge, e.g. interesting objects are compact, contrasting, and of at least a certain size, but they fail to provide a satisfying means of detecting and localizing objects of interest in challenging situations. On the other hand, state-of-the-art detection methods typically rely on specific contexts e.g. the object of interest is a horse seen from the side with the head facing left and whose formal model has specific parameters, in order to provide the best results in proper contexts that is, good combinations of detection and false alarm rates. The method proposed in this paper efficiently detects and precisely localizes, out of the huge number of possibilities, unexpected but interesting objects in a given contour primitive map obtained from a single intensity image. It relies on contextual knowledge more generic than state-of-the-art detection methods but more specific than state-of-the-art segmentation methods. The proposed approach is to use an extended set of grouping criteria and constraints in order to find objects of interesting shapes. Much evidence has been published that perceptual grouping of contour primitives is important in human vision [8,9]. However, as discussed in [10], few results are known concerning high-level grouping for object detection and localization.

In computer vision literature, generic grouping criteria for shapes are typically limited to simple ones, e.g. local continuity and smoothness [11], or global closure [12,13] and convexity [14]. Elder and Zucker [12] used a shortest-path algorithm to find maximum-likelihood closed contours. Even though no criterion was used to explicitly target multi-part objects, they presented a result where a single contour for the complete shadow of a multi-part wooden doll was extracted. However, very few if any confounding primitives seemed to be present on the object, on its shadow, or in the background (the actual primitive map was not shown). Different partial contours were also extracted on the doll and in between the doll and its shadow. No attempt was made at identifying the most interesting contours. Targeting simpler shapes and high efficiency, Jacobs [14] used a constraint (threshold) on a global salience criterion based on the relative amount of gaps in a convex closed shape. Efficiency resulted mainly from using monotonic constraints, meaning that a subset of primitives could be rejected as soon as it did not satisfy the constraint since adding more primitives could not make the larger subset acceptable. While convexity is properly monotonic, [14] also used the salience constraint as if it were monotonic, even though it is not since adding a long convexity-preserving segment with no gap to a convex group may increase its salience. In fact, relatively few grouping criteria and constraints are proposed in literature and they are mostly monotonic. Unfortunately, this choice limits the possibilities of proposed methods in terms of detecting shapes of interest. A comparative study of [12] and two competing methods was made by Wang et al. [15]. For natural images of animals, the optimal boundary has a simple near-convex shape not representative of the animal shape. The limitation to monotonic criteria and near-convex shapes is also typical of previous saliency-based methods [11]. For instance, saliency networks impose that an optimal curve must be composed of sub-curves that are themselves optimal, a property equivalent to monotonicity and referred to as *extensibility* in [16].

With complex shapes, effective grouping criteria are more likely to be non-monotonic. Non-monotonic grouping refers to the fact that the value of a given formal criterion is not changing monotonically i.e. either always increasing or always decreasing as the number of grouped (added, merged) primitives increase. For instance, a group of two primitives may get a higher score on a given criterion than another group of two primitives but the same third primitive added to the first group may give a lower score than when added to the second group. This makes it much more difficult to reach even a local optimum simply by following an ascending or descending gradient in the value of a single grouping criterion. In the method proposed in this paper, many such non-monotonic grouping criteria are combined in the objective function which makes typical deterministic and even random discrete optimization techniques most likely ineffective. The effect of grouping non-monotonicity becomes more apparent as the number of elements in each group grows from two to three and more, as is the case in the proposed method. In practice, an important consequence of the non-monotonicity of the grouping criteria is that, as the subsets of primitives considered increase in size, the constraints need to be adapted. For instance, as the shape and interest of a partial object becomes clearer, constraints may appear or disappear or more simply diminish or increase. Previous search-based methods avoided this added complexity by implicitly or explicitly assuming monotonic criteria and constraints. A recent method by Estrada and Jepson [13] extracts salient non-convex contours by developing a search tree, enforcing fixed constraints and rejecting a fixed proportion of less promising partial contours.

In this paper, the hypothesis is made that global, non-monotonic grouping criteria are needed in order to detect and localize unexpected but interesting objects directly from a noisy and cluttered primitive map. A single generic set of grouping criteria is used in the proposed detection method. Our approach is a generalization of [14] where the criteria and constraints are more numerous and non-monotonic. The main contextual knowledge to be relied on in defining the shape-based grouping criteria is that objects of interest are of