Signal Processing: Image Communication 58 (2017) 240–257

Contents lists available at ScienceDirect

Signal Processing: Image Communication

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

Severe-occluded 3D object identification via region-based descriptions☆

Marcos Escudero-Viñolo *, Jesus Bescos

Video Processing and Understanding Lab. Escuela Politécnica Superior, Universidad Autónoma de Madrid, Spain

ARTICLE INFO

Keywords:
3D-object identification
Severe-occlusion
Region-based PoI descriptions
Identification in clutter
Self-organized maps
Distributed encoding

ABSTRACT

This paper describes a region-based strategy for part-based object identification with independence of the external factors that affect its captured image: light variations, capture point-of-view or occlusions. Starting from color images and depth estimations, i.e. not requiring 3-dimensional models, we focus on the identification of learned objects in severe-occlusion scenarios. To face this problem, we assume that objects have been preliminarily segregated from the scene. Strong changes of appearance—due to one or several of the aforementioned factors or to the object nature, e.g. deformable objects—substantially increase the problem complexity. The proposed algorithm operates by splitting segregated objects in successively coarser region-partitions, with each region representing a part of the object from which it was extracted. For the characterization of these parts, two region-driven descriptors are proposed: R-DAISY and R-SHOT. Their novelty relies on the use of a size-and-shape-variable description support which is automatically defined by the object part itself. Descriptions obtained in this way are self-organized in a single neural structure by an unsupervised learning process. Experimental results are promising in the identification of severe-occluded objects using a small set of training instances—1-to-8 short-varied views per object.

1. Introduction

1.1. A review of existing approaches with a connection to human perception

Humans are able to identify an object in a wide range of sizes and points of view. In order to achieve invariance with respect to these factors, the visual information is supposed to be projected from the retinotopic organization on a cerebral area so that projections from various retinotopic locations converge to the same invariant area. Although humans are aware of an object's relative size, position and rotation during its identification process, the process is highly robust to changes in its aspect. There are two main information-codification theories that attempt to explain this fact [1,2]: object centered and viewer centered.

The **object centered** perspective emerges from a psychophysical point of view. Object information is assumed to undergo a set of transformations until it matches a single stored 3D template. Object-centered studies adopt a common identification path. In a first stage some sort of features are extracted (depending on the theory, these range from simple shapes, lines, edges or salient areas to volumetric primitives); then, these features are sorted and combined. In the final stage, the object is identified by formulating a query to the stored knowledge [3–5]. In machine vision applications inspired by this theory, the connection is particularly noticeable in the object characterization stage. The use of different types of gradient-based descriptions – as described in the first layers of the Marr computational theory [3] – has been widely reported in the literature, either looking for salient and invariant-to-appearance points or regions ([6]), or arranged in local histograms – e.g. via HoG ([7]) –, or part-wise combined as proposed in the well know deformable part models [8]. Alternatively, object descriptions via 2D and 3D primitives were proposed in [9], whereas descriptions via shape and Textons were the core of [10,11] and [12], respectively. These essence-based descriptions are closely related to the elementary features described in the Feature Integration Theory [4] in the case of schemes based on primitives, and closely related to the Biederman's axon-based studies [5] in the case of Texton-based schemes.

The **viewer centered** perspective is based on the assumption that knowledge increases with experience. Having observed a *representative* number of views of an object, new observations may be easily matched with the previous observed set [2]. Almost every artificial training system is coherent with this assumption. However, how many views are required to reasonably succeed in the identification of objects from new points of view? State-of-the-art methods require either the tedious

☆ Partially supported by the Spanish Government through its TEC2014-53176-R HAVideo project.
* Corresponding author.
E-mail addresses: marcos.escudero@uam.es (M. Escudero-Viñolo), j.bescos@uam.es (J. Bescos).

http://dx.doi.org/10.1016/j.image.2017.07.007
Received 27 December 2016; Received in revised form 25 July 2017; Accepted 25 July 2017
Available online 3 August 2017
0923-5965/

collection and annotation of large data corpora to learn object models, or the use of both detailed and parametrizable 3D object models, such as the well known CAD models. The use of features derived from these models has been proven to be a successful strategy for the identification of objects in complex scenarios [13]. Nevertheless, these designs require the existence or pre-construction of the CAD models, a requirement that hinders their scalability as well as their use in user-oriented applications. Furthermore, their operation is usually limited by the sampled model's orientations used for training, and, when used to develop holistic models, by the presence of occlusion. In this sense, region-based approaches – already well established in 2D scenarios [8,14–16] – are supposed to operate well in the eventuality of partial occlusions and, if applied adequately, may be more robust than holistic approaches to unobserved variations – hence untrained – of the object's appearance.

This study faces the problem of identifying arbitrarily textured 3D objects captured with a Kinect camera, in the absence of a 3D model or of a huge training data-set for such objects. Object models that were trained using a very small number of instances – in our experiments no more than 8 instances per object were used – are then used to identify new instances of these objects when they are captured from new points of view and/or under different illumination conditions. Our aim is to identify objects when they are severely occluded by other objects, which prevents the use of templates and holistic models. Most of the existing works facing this problem select or design a set of ideally robust and discriminative features to determine correspondences between an object instance and the modeled objects. Examples of this scheme are the studies dealing with descriptions based on: point signatures [17], spin images [18], spherical spin images [19] and local surface patches ([20]). Alternatively, closer in time studies explicitly devoted to object identification with the Kinect device have also been proposed [21–25]. Some of them export classical two dimensional object descriptions to *RGB-D* situations (e.g. singular-points in [24] or signatures of histograms in [26]). Moreover, in order to provide robustness to occlusions, techniques based on Hough voting for both non-deformable and deformable objects have been proposed as well [21]. Regarding the learning strategies, linear and non-linear support vector machines [27] are usually preferred; however, Hierarchical Kernel descriptions [23] and tree-based approaches [28] have also been used with relative success. Nowadays, powerful deep learning schemes have improved every other approach ([29,30]).

1.2. Main idea and motivation

It should be made clear that this study in no way intends to partially replicate how the human visual system works; instead, it aims to *mimic* some operation mechanisms that are supposed to take place on it, as suggested by research results in the field of visual perception in highly-developed visual systems (in particular, this study is highly inspired by [1]). In our opinion, the existing computer vision approaches to object identification in *complex* scenarios are constrained not only by the features or the metrics used to model and compare object instances, but also by their operational paths and strategies. This study is not focused on learning enormous bunch of data or on designing deep learning schemes to cope with every possible object appearance, but rather on providing alternative description schemes and operation paths to pave the road for potential region-driven schemes that bypass occlusions by part-based identification. In essence, this study is mainly motivated by three premises or targets:

1. Define an object model that can be trained with a very small number of samples.
2. Confront object identification under severe occlusion situations.
3. Provide a distributed modeling approach to self-arrange training evidences.

Training with a few samples. A critical component of vision is the creation of visual entities, that is, representations of surfaces and objects that do not change the perceived scene but change those parts we see as belonging to other objects and how they are arranged in depth. Humans learn the appearance of objects by combining examples with their knowledge of the behavior of the visible world – i.e. their expertise [31] –. In general, a person would not require many examples to re-identify an object when it is rotated or when it appears in a scenario different to that in the *learned* examples (few-shot learning [32]). This is agreed to be achieved by the feature management mechanisms used to perceive objects by the ventral path [33]. Back to the computer vision world, template-matching approaches train the object model with thousands of templates usually extracted from a CAD-model [13]. In these cases, robustness to object rotation and scale change is obtained by quantifying and sampling the potential appearances of an object observed from a set of plausible points of view. One of our targets is to model objects with a low number of training samples. To cope with object rotation, we propose to provide robustness in the description itself, via the use of a local reference frame [34], i.e. locally aligning descriptions with the object – or with a particular part of it – so that these are independent of the point of view from which the object is captured. This is not a novel approach, rather the authors of [26] have recently compiled reported studies in the design of such sort of descriptions. Nonetheless, in order to recover – at least putative – correspondences between local referenced descriptions, the reference for the local alignment should be stable for different views of the same object; that is, the reference should be stable to point-of-view changes. To cope with scale changes, the scale-space theory [35] establishes a well-founded mathematical framework to discover singular points of an object view that are recoverable to some degree from a moderate affine-transformed view of the object. In fact, this theory establishes the basis of classical point-of-interest detectors including the well-known SIFT points [36].

Handling severe occlusion situations. When an object is occluded, only part of it – potentially as little as 10% – is visible. A system aiming to identify the object in these scenarios should adapt its trained knowledge – observed samples, which in order to maximize the amount of training data are generally extracted from holistic examples – to an unpredictable occluding situation which always results in incomplete instances of the target object. The extent of these incomplete instances is visually defined by both the real contours of the object and the occluding contours of the interfering objects; hence, contours and holistic templates might not be a reliable cue for characterization. From our perspective, there are two main ways of facing this situation. The first one consists of fitting a holistic model to the visible and non-visible parts of the occluded instance, assuming that the not-visible part of the object remains unaltered. The likelihood of the instance being the test object can be obtained by measuring the similarity between the instance's visible part and its corresponding part in the model [13]. This top-down approach, strongly linked with the Gestalt's principle of continuity [37], may fail if the initial holistic fitting is inaccurate or if the visible parts are insufficient to establish a reliable correspondence to a specific part of the model. The second one, a bottom-up alternative, consists of considering the occlusion in the learning process, i.e. dividing the object in its *semantic* parts and training each part independently. The instance may be then identified, at least partially, by integrating the likelihood of each identified part. A system driven by this philosophy should operate better in situations where only a small part of the object is visible, which is our objective. Advantages of using this part-based modeling philosophy for the identification of cars captured from different 3D view-points were illustrated in [38].

Distributed model encoding. Humans have limited resources available for storing knowledge ([39]), which disregards the idea of encoding specificity, that is, the existence of devoted neurons that are activated by a particular complex stimulus (e.g. a particular object or face). However, many of the existing studies on artificial object identification follow precisely this path: models are trained for specific objects and only the