Knowledge-Based Systems 98 (2016) 162–171

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Twofold consensus for boundary detection ground truth

C. Lopez-Molina[a,b,*], B. De Baets[b], H. Bustince[a]

[a] Departamento de Automatica y Computacion, Universidad Publica de Navarra, Pamplona 31006, Spain
[b] KERMIT, Department of Mathematical Modelling, Statistics and Bioinformatics, Ghent University, Coupure links 653, Ghent 9000, Belgium

ARTICLE INFO

Article history:
Received 15 June 2015
Revised 19 January 2016
Accepted 21 January 2016
Available online 2 February 2016

Keywords:
Boundary detection
Quality evaluation
Binary image
Consensus image
Ground truth

ABSTRACT

In the evaluation of boundary detection methods it is common to use as ground truth a set of boundary images that are hand-made by human experts. This work proposes a novel representation of this ground truth. More specifically, we propose to combine the hand-made boundary images into a set-based consensus, which is constructed from the concordances and discordances among the images. We study the theoretical and visual properties of this consensus and present an application to boundary image quality evaluation.

1. Introduction

An important aspect of technical and scientific developments is the ability to quantitatively measure the quality of an automatically generated solution to a given problem, ideally in a comparable and unambiguous way. Image processing is not an exception to this. Each task in image processing has different characteristics and requirements, which lead to the use of task-specific quality evaluation techniques (see [1] for segmentation or [2] for image compression, for instance). If the task has a direct application to a real-life scenario, the evaluation can be based on the utility of the computer-generated solution for that application [3]. However, most of the works on image processing do not intend to solve specific problems, but rather propose general-purpose techniques instead, so that they have to be evaluated considering their fitness in general scenarios.

In the case of boundary detection, there exist plenty of options for evaluating the quality of an automatic method [4,5]. However, none of the proposals in the literature has achieved wide acceptance. Moreover, we demonstrated in [5] that they all have theoretical and/or practical flaws that might lead to misinterpretations of the evaluation of the quality of a given image.

One of the obstacles we identify in the process of evaluating the quality of a boundary image is the inherent difficulty in obtaining and representing ground truth solutions. Differently from what happens in other image processing tasks, it is unclear how to obtain perfect solutions the automatically generated boundary images can be compared to. In fact, some authors use humans as evaluators instead of using ground truth [6,7], but this option is frequently rejected due to its low reproducibility. The problems in obtaining ground truth are derived from the fact that there is no mathematical, unambiguous definition of what boundaries are, so that most of the proposals for boundary detection rely on loose, intuitive definitions. Examples of such definitions are *contour at the center of the slope between two adjacent regions with a considerable difference in gray level* [8] and *sharp change in intensity* [9]. Therefore, in practical terms, the boundary detection task is redefined as marking up the boundaries a human would consider to be worth tagging in an image. This enforces the ground truth to be human-made boundary images. Although this is not a problem itself, it leads to multiple situations for which no answer has been provided in the literature, the most relevant being that in which different humans produce very divergent solutions. The discrepancies can result from marking up (or not) certain objects whose importance in the image is *debatable*, but also from locating their boundaries at different positions.

In this paper we tackle the management of diverse ground truth boundary images generated by humans, and show how to represent their consensus to make the most of the combined information. In Section 2 we describe the problem and list the solutions provided in the literature so far. In Section 3 we unfold the

* Corresponding author at: Departamento de Automatica y Computacion, Universidad Publica de Navarra, 31006 Pamplona, Spain.
E-mail addresses: carlos.lopez@unavarra.es (C. Lopez-Molina), bernard.debaets@ugent.be (B. De Baets), bustince@unavarra.es (H. Bustince).

http://dx.doi.org/10.1016/j.knosys.2016.01.033
0950-7051/

motivations and properties of our proposal for representing the ground truth. The applicability of our construction to quality evaluation, is studied in Section 4, while Section 5 recaps the benefits and drawbacks of our proposal.

2. Handling diversity in ground truth for boundary detection

2.1. Notation

In the remainder of this work we consider the images to have some fixed dimensions $\mathcal{M} \times \mathcal{N}$. The set of all binary images is denoted $\mathbb{E}$, and can be seen as the power set $\wp(\Omega)$, where $\Omega = \{1, \ldots, \mathcal{M}\} \times \{1, \ldots, \mathcal{N}\}$ represents the set of positions in an image. We refer to individual boundary images with upper case (e.g. E, I), while bold-faced upper case is reserved for sets of images (e.g. $\mathbf{A} = \{A_1, \ldots, A_n\}$).

An error measure for boundary detection is a function $q : \mathbb{E} \times \mathbb{E} \to \mathbb{R}$, where the first argument is the candidate image and the second is the perfect solution (when single ground truth images are provided for each image). If the ground truth for each image is provided as a set of images, then an error measure is defined as $q^* : \mathbb{E} \times \wp(\mathbb{E}) \to \mathbb{R}$. For obvious reasons, the functions q and q^* need to be monotone w.r.t. the quality of the candidate image, *i.e.* monotone w.r.t. the perceived closeness between the candidate boundary image and the ground truth. Without loss of generality, we consider the error measures to be decreasing w.r.t. the quality of the solution, so that their range becomes $\mathbb{R}^+$ and 0 stands for the minimum error (maximum quality).

Since binary images can be represented as subsets of Ω, we consider the classical set-theoretic operations on binary images, namely intersection ($\wedge$), union ($\vee$), inclusion ($\subseteq$, $\subset$). The symbols $\cap$ and $\cup$ are reserved for the intersection and union of sets of images, respectively.

As defined by Serra [10,11], the dilation of a binary image A by a structuring element K, denoted $\mathcal{D}_K(A)$, is given by $\mathcal{D}_K(A) = \{c \in \Omega \mid c = a + b \text{ for some } a \in A \text{ and } b \in K\}$.

2.2. Combining multiple ground truth images for boundary detection

The most evident option to handle multiple ground truth images is to take all the images generated by humans as perfect solutions, without further processing. In this way, a candidate boundary image $E_c \in \mathbb{E}$ is evaluated against each of the ground truth images, $\mathbf{S} = \{S_1, \ldots, S_n\}$, and some sort of aggregation of the one-to-one comparisons is used to generate a final (combined) evaluation. This option can be formulated as

$$q^*(E_c, \mathbf{S}) = g(q(E_c, S_1), \ldots, q(E_c, S_n)), \tag{1}$$

where g represents a symmetric n-ary aggregation function [12].

Generally, the minimum operator is chosen as aggregation function g, so that the final result is given by the one-to-one comparison producing the lowest error. This choice leads to quantifying how close the candidate image is to a human-made one. One of its advantages is that the human-made boundary images will always provide perfect scores in the quality evaluation (*i.e.* if $E_c \in \mathbf{S}$ then $q^*(E_c, \mathbf{S}) = q(E_c, E_c) = 0$). However, it also has some deficiencies. First, the information is not processed in any way, so that there is no derived knowledge generated from the human-made images. As a consequence, the quality is always dependent upon a single human-made image. Moreover, the human-made images correspond to isolated error minima in the space of solutions, hindering the potential use of optimization methods.

If some sort of mean is used as aggregation function g (for example, the arithmetic mean of the one-to-one comparisons), we avoid some of the deficiencies associated with the minimum operator, but we force the human-made images to be rated as non-perfect. That is, as soon as we have some discordances in $\mathbf{S}$, we are likely to assign a non-zero error to every boundary image, including those generated by humans. In fact, human-made images could eventually be rated worse than automatically generated ones. Although some authors have stated that not even the solutions in the ground truth are *perfect* [13], we will we will treat them as such.

An alternative to the individual comparison against each image in the ground truth set is producing a consensus image out of them. Ideally, the consensus would be some sort of boundary image on which the humans would agree, so that the evaluation of an image E_c given a set of ground truth boundary images $\mathbf{S}$ is reformulated as

$$q^*(E_c, \mathbf{S}) = q(E_c, \text{cons}(\mathbf{S})). \tag{2}$$

The role of the consensus operator in Eq. (2) is to perform image fusion or aggregation, which is far from trivial. When the images represent a standard scene (in whichever tonal or spectral representation), the problem mostly reduces to the aggregation of the values at each pixel or region [14]. Methods based on band decomposition have been proposed, typically using wavelets [15] or similar mathematical constructs [16]. Although particular conditions might affect the interpretation of each of the images to be fused (e.g. in the case of multifocus or multiresolution image fusion), the problem amounts to numerical aggregation. When the images represent certain features, such as boundaries or ridges, the problem is more intricate. In these situations the fused image is not meant to be representative of the initial ones in visual terms. Instead, it is their contents, and their interpretation, which needs to be preserved in the fused image. In the case of boundary images, image fusion techniques at the pixel or region level [17] are not advisable, since boundary images might contain boundaries corresponding to the same silhouette at fairly displaced positions. In the specific case of boundary images generated by humans, hardly ever will two humans produce exactly the same boundary image, nor will they mark up a complete silhouette at the same exact positions. In general, given the special characteristics of boundary images (they are binary, mostly black, and boundaries must be represented as 1 pixel-wide lines), standard image fusion techniques cannot be applied to this task.

As far as we know, the only proposal in the literature to generate a consensus image out of a set of boundary images is due to Fernandez-Garcia et al. [18], and consists of selecting a combination of the human-made images. More specifically, the authors stack up the n boundary images and select the level cut having the lowest average distance to the ground truth boundary images, in terms of Baddeley's Delta Metric [4,19]. This work is similar to previous techniques used to avoid the problem of parameter setting by selecting a best candidate from a pool of boundary images [20,21]. The basic difference between the proposal in [18] and those in [20,21] is that in the former a new image is created from the initial pool of candidates, and might not correspond to any of the original images, while in the latter the candidate boundary image minimizing the average distance to the others is kept. Despite the difficulties, creating a consensus image provides some theoretical advantages compared to the formulation in Eq. (1). For example, it allows for the generation of new knowledge from the human-made images, in the sense that the initial information (original images) is combined to produce more elaborated knowledge. Moreover, the evaluation process gains robustness, since boundaries marked up in a minority of the boundary images could be relegated in favor of those selected in a majority. However, we still identify problems, such as the fact that some human-made boundary images are likely to be taken as non-perfect.

Note that different authors have analyzed how to handle multiple hand-made ground truth for tasks other than boundary detection. A good example is medical image segmentation, for which works such as STAPLE [22] or shape-based averaging [23]