A survey on matching algorithms for boundary image comparison and evaluation

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

[a] *Dpto. Estadistica, Informatica y Matematicas, Universidad Publica de Navarra, 31006 Pamplona, Spain*
[b] *KERMIT, Dept. of Data Analysis and Mathematical Modelling, Ghent University, Belgium*

Abstract

Most of the strategies for boundary image evaluation involve the comparison of computer-generated images to ground truth solutions. While this can be done in different manners, recent years have seen a dominance of techniques based on the use of confusion matrices. That is, techniques that, at the evaluation stage, understand boundary detection as a classification problem. These techniques require a correspondence between the boundary pixels in the candidate image and those in the ground-truth; that correspondence is further used to create the confusion matrix, from which evaluation statistics can be computed. The correspondence between boundary images faces different challenges, mainly related to the matching of potentially displaced boundaries. Interestingly, it relates to many other fields of study in literature, from object tracking to biometrical identification. In this work, we survey all existing strategies for boundary matching, we propose a taxonomy to embrace them all, and perform a usability-driven quantitative analysis of their behaviour.

Keywords: Boundary image, Displaced boundary, Linear feature matching, Image comparison

1. Introduction

In the context of boundary detection, quality evaluation has long been studied, the first references dating back to the 1970s [1, 2]. The reasons are manifold, including the need to rank different boundary detection methods or the development of training-based methods, which demand reliable objective functions. It is generally accepted that the best way to evaluate boundary detection methods is by comparing their results to those by humans. However, there is no agreement on which is the most reliable way to (quantitatively) perform such comparison, regardless of whether it is carried out in terms of similarity (how close both images are) or dissimilarity (how different they are). As of today, several different techniques and configurations are used; this has led to a rather disorganized situation in which very few standards are kept.

In the past, most of the measures for boundary image comparison relied on distance transformations [3, 4] (see [5] for a historical presentation on the topic), which are convenient to overcome counting dilemmas due to variable boundary position or boundary cardinality [6]. However, in recent years, a new family of proposals approached the problem with a classification-based inspiration. The reason is simple: at a very basic level, boundary detection is nothing else but binary classification, since every pixel in the image must be labelled as either boundary or not. The comparison between two images is then phrased in the usual terms of binary classification: the matching boundary pixels becoming True Positives (TPs), etc. Once a confusion matrix has been created, well-known statistics and quantifiers yield final evaluations [7, 8].

The classification-based inspiration is very convenient and attractive, since it relates to machine learning and classification, two fields in which quality evaluation is deeply studied [9, 10, 11]. However, it also

Email addresses: carlos.lopez@unavarra.es (C. Lopez-Molina), cedric.marco@unavarra.es (C. Marco-Detchart), bustince@unavarra.es (H. Bustince), bernard.debaets@ugent.be (B. De Baets)

comes coupled with a critical handicap that is unresolved in the literature: if the boundary of an object appears at slightly displaced (non-overlapping) positions in two images, the comparison method should be able to recognize the circumstance and count its pixels as correctly classified. That is, the counting of the correctly/wrongly classified pixels cannot be done based on mere a pixel-to-pixel comparison. A more elaborated *matching* is needed to map the boundaries in one image to those in the other, up to some spatial tolerance. This matching would ideally be able to tolerate small spatial deviations, yet not pairing the boundaries of objects to those due to different objects, texture or noise.

Literature contains a list of different strategies for boundary matching, either presented in specific papers or simply used within another research or dataset, what yields the question on which one to choose, or how comparable are results by any pair of them. The present work analyzes the literature on boundary matching, proposing the first specific taxonomy categorizing all relevant proposals. Also, the work takes the example of boundary quality evaluation as context to present a benchmark in which four significant strategies are tested. Such benchmark does not intend to rank the performance of such strategies, but to question whether they yield significantly different results.

It is relevant to note that the idea of boundary matching is in fact related to many other open problems in literature. The boundary matching problem is, at a broad level, that of linear feature matching[1], which has been regularly addressed in computer vision literature. An evident solution is to treat the problem as bipartite graph matching, *i.e.* one-to-one mapping of the boundary pixels in one image to those in the other, typically minimizing the distance between paired pixels. However, the fit of this solution is not perfect in the context of boundary matching. Firstly, the computational cost of deterministic optimal algorithms (such as the Hungarian/Munkres algorithm [12]) is exorbitant. Secondly, there is also a theoretical problem, since bipartite graph matching requires one-to-one correspondence. This is problematic in the context of boundary matching strategies because non-overlapping boundaries might be composed of a different (yet similar) number of pixels. Several authors have proposed alternatives to this algorithm, either focusing on the computational cost (as Martin [13] using the CSA algorithm by Goldberg and Kennedy [14]) or presenting alternatives able to cope with the one-to-many correspondences (as Estrada and Jepson [15]). Additionally to graph-based solutions, alternatives based on area overlapping or mathematical morphology have also been employed. Currently, there is no clear way of knowing which is the best alternative for boundary matching. The most accepted one is the CSA algorithm, but this might due to the fact that it is used in the most popular boundary detection benchmark (the BSDS [16]).

In this work we review the most relevant boundary matching techniques with application to boundary quality evaluation. Moreover, we investigate and compare their performance. Since there is no clear way of knowing *which one works better*, we pose a different question: do different alternatives have a real impact on the resulting confusion matrices? That is, should we expect significantly different results when using different matching strategies? Otherwise said, are their results (or the conclusions extracted thereafter) compatible? Although this does not answer the question on the best possible option, it clearly reduces the controversy about which strategy to choose and the consequences of such choice.

The remainder of the work is organized as follows. Section 2 describes techniques for boundary matching and proposes a taxonomy that covers the literature. Section 3 includes some experimental comparisons of the most relevant matching techniques in the context of boundary evaluation, and Section 4 recaps a brief discussion.

2. Boundary matching techniques

While understanding the human visual system remains a challenge, it seems clear that humans make intensive use of features to compose objects and scenes [17, 18]. So is the underlying idea of Marr's Primal Sketch [19], arguably the best computational model for human vision. While there exists a diversity of

[1] In this work, we refer as *linear features* to any visible artefact whose representation is a line. This holds regardless of their origin (edges, countours, ridges) or characteristics (lines, curves, shapes). Their *matching* is the search for correspondence between such artifacts.

features, the ones represented as lines or curves play a key role in the human visual system. In fact, some famous experiments in the context of cognitive sciences are based on the recognition of contours and shape-based objects, such as the Shepard-Metzler study [20] and subsequent evolutions into the Mental Rotation Test [21, 22].

Not surprisingly, linear features became one of the most relevant low-level features in computer vision [23], either representing visually salient linear structures or any other artefact. In this work we concentrate on boundary images, which use linear features to represent the silhouette of relevant objects in a scene. Nevertheless, examples of use of linear features can also be found in biometrics [24], computational biology [25, 26] or photogrammetry [27, 28].

The ubiquity of linear features in computer vision makes it necessary to design comparison methods, either at individual or whole-scene level. The generic problem of linear feature comparison and matching has been recurrently materialized in specific challenges. The tools and techniques used to resolve it are, nevertheless, heavily dependent upon contextual matters. Such matters include, e.g., the characteristics of the linear features, the semantics of the problem to be solved, or the expected output of the matching process. As a result we find a diversity of linear feature matching techniques, most of them being applicable exclusively to the context for which they were designed, and generally not portable to boundary matching or comparison.

In Section 2.1 we review different tasks that are partially or totally based on linear feature matching. This analysis is driven by their interestingness for our final goal: boundary comparison and matching. Then, a mathematical framework for such goal is provided in Section 2.2. Finally, Section 2.3 introduces a novel taxonomy for boundary matching techniques.

2.1. Linear feature matching for image processing

Linear feature matching is the process of comparing, in quantitative terms, the linear features in two scenes. There are two factors with major influence in linear feature matching: (a) the nature and constraints of the potential displacements of the same feature in different images and (b) the expected output of the matching. The first factor relates to the nature of the linear features and the circumstances under which they were gathered. The second factor is bounded to the utility of the matching process, and the application in which it is intended to be useful. A third potential factor is the presence of multiple (normally, different) shapes in each image, which severely increases the complexity of the task. Normally, the presence of one single shape is key to object recognition and retrieval, but it is a prior which cannot be applied to context of boundary image comparison. The study of linear feature matching techniques in the literature must therefore be performed under a dual, if not ternary, perspective.

The most evident application of linear feature matching algorithms is line pattern recognition, with applications ranging from biometrics to aerial imagery registration. The complexity of linear patterns has often led to the use of derived information in the matching process, instead of the linear features themselves [29]. However, some authors propose an explicit matching of linear features (e.g. for palmprint matching [30] or for fingerprint matching [31]).

A task slightly different from line pattern matching is that of silhouette (also, shape) matching, often with the goal of object recognition or classification. Generally, silhouettes contain most of the information for object recognition, while avoiding problems related to image texture, shading or colour. In some scenarios, *texture or colour cannot be used as a cue for recognition* [32], while in others silhouettes contain enough information to complete the recognition task [33, 34]. Literature contains biometric systems based on silhouette matching and recognition, e.g. for person/gait recognition [35], forensic reconstruction [36], or hand shape-based identification [37]. Also within the context of silhouette recognition we can list linear feature matching tasks, in applications such as stereo matching or image correspondence [27]. Although stereo matching is often carried out using disparity maps [38], early proposals were based on silhouette recognition and matching [39, 40, 41]. In recent works, line matching is still used for other delicate calibration tasks, e.g. pose estimation in bifocal camera setups [42].

Recognition tasks, either based on linear patterns or shapes, can be roughly divided according to their output: verification (matching of two elements or not) [36], identification (which element, out of a pool of candidates, is a good matching) [43] or offset computation (e.g. modelling the physical 3D setting of two or more scenes)[44]. Also, they can be discriminated on the basis of the existence of prior shape models, although such models are mostly used for pre-trained segmentation tasks [45]. None of them fits the semantics of boundary image matching, mostly because recognition is grounded on the idea that a perfect (or fair enough) version of the instance is stored in the system. This often turns these tasks into the computation of an optimal deformation model.

Silhouette matching and recognition is so relevant that it led to prominent mathematical developments. A good example is the Curvature Scale-Space (CSS), which has as final goal the comparison and matching of curves and shapes. Introduced by Mokhtarian and Mackworth [46], the CSS uses a parametric representation of non-overlapping curves based on paired functions. Using this representation, the similarity between two curves is quantified as the comparison of such functions, more specifically of the zero-crossings in their second derivatives when projected into the Gaussian Scale-Space (GSS). The CSS is aimed at scale- and rotation-invariant silhouette recognition, as many other proposals in the context of object recognition [47]. These aims are possibly critical to explain human shape recognition (whose behaviour in this regard is still under analysis, see recent works by Chen et al. [48] or Han et al. [49]), but play no positive role in boundary comparison or matching, in which rotation or scale invariance is undesired. Subsequent evolutions of the CSS (see, e.g. [50]) extended the original proposal, but do not substantially enhanced its utility for boundary matching.

An interesting variation of the problem of shape matching is that of shape-based object tracking, examples being [51, 52, 53]. The difference between shape matching and shape-based object tracking is twofold. First, tracking is performed across several images, not only two of them. This enables, among other possibilities, simultaneous multi-image evaluation, often for data correction or interpolation. Second, there is contextual information that takes relevance in modelling the movement of the objects. That information involves, for example, the appearance of other (potentially occluding, or occluded) objects in the scene [54], as well as inertia influencing the object movement. Moreover, restrictions in the movement of objects can apply, e.g. in [55, 56], rotating motion matching relies on an expected geometrical transformation. Silhouette-based object tracking is usually solved by applying non-rigid models based on active contours or snakes [57, 58]. None of these solutions are available for boundary image matching, where the variation in the displacement of the boundaries can greatly differ for different objects in the same image, or even different segments of the same object boundary.

Object tracking is the most prominent example of shape matching involving multi-image analysis, but not the only one. Another very relevant example is multiscale image analysis. Within multiscale image analysis we often find inter-scale linear feature matching, especially when it involves tracking. As proposed by Bergholm [59], evolving the ideas by Witkin [60], tracking-based linear feature detection methods intend to discriminate the relevant boundaries in an image at a coarse scale; then, those boundaries are tracked down to the position they occupy at a finer scale. In this way, a feature detector can combine the good discrimination properties of large scales with the accuracy of the fine ones. The tracking process starts out by matching the boundaries at the coarsest scale to those in the immediately finer scale, then repeats the process until the image corresponding to the finest one. Normally, cross-scale feature displacement is limited by a maximum scalewise displacement, which dramatically eases the problem. Although tracking has received considerably less attention than other components of multiscale image analysis (e.g. the theoretical properties of the scale-spaces), some alternatives have appeared in the literature [61, 6]. Note that multiscale edge and ridge analysis often demand linear feature matching, even if it does not involve tracking. For example, the practical implementation of Lindeberg's ideas in the GSS [62], which involve the location of the optimal scale for each boundary, demand the explicit construction of the so-called *edge surfaces in scale-space*. Such surfaces are typically given by the correspondence of the positions of each boundary at each scale.

Although linear feature tracking (or edge surface construction) and boundary matching for quality evaluation seem to be close tasks, relevant differences arise in the detailed analysis. Linear feature tracking involves a sense of inertia and, despite being carried out as consecutive matchings between pairs of images, it aims at the construction of multi-image, cross-scale structure. Also, the scale-spaces under which the image is projected can incorporate relevant constraints to the matching problem. For example, the causality principle in the GSS imposes that any feature at a coarse scale corresponds to a *not-necessarily unique* [63] feature at a lower scale. This allows, and in a sense encourages, that a single feature at a given scale is matched to spatially diverging features at a finer scale. While some similarities between boundary matching and edge tracking are evident, in particular the fact that no explicit matching is required, solutions from linear feature tracking can hardly be ported to boundary matching for comparison.

A completely different approach to line-based object recognition is that mimicking human recognition abilities through the implementation of CNNs. This research, which roots back to early Mental Rotation Tests [22], attempts to model human abilities in early stages of the Human Visual System (HVS), specially regarding rotation, scale and eccentricity invariance [64, 48]. Since some of these experiments are based on line based draws (e.g., in [49], Korean characters), it is to be expected that eventually CNNs might mimic human line pattern recognition abilities. Such a solution would, however, raise evident questioning when used for boundary matching in the context of quantitative evaluation. Firstly, some of the features (as scale invariance) are hardly desirable when applied to comparison, even if it seems to be part of the HVS [64, 49]. Secondly, if many differently trained CNNs yield dissimilar results, how to produce a canonical comparer out of them? Alternatively, if using *any* of such CNNs for the task, how to audit their results [65]? Thirdly, as reported by Nguyen, Yosinski and Clune [66], well-performing CNNs can lead to aberrant results which are not only misleading, but also yielded in almost-full ($> 99\%$) confidence. As the authors state, their findings raise *questions about the true generalization capabilities of DNNs* [66], which are of paramount importance for the present task.

As a recap, we find very different tasks based on (or supported by) linear feature comparison or matching, but exporting such solutions to the context of boundary quality comparison and matching is troublesome. Recognition or classification systems usually look for a closest-possible match, assuming that the same object appears in two different images. Tracking systems extend that assumption to a large number of images, even if the object(s) to be tracked is (are) potentially occluded at some of them. On their side, bio-inspired CNNs seem to be more focused on object recognition, which makes them unfit for boundary comparison. Moreover, the semantics of the output of such tasks do not properly match that of boundary image evaluation.

2.2. A model for boundary evaluation and matching

Boundary matching for quality evaluation is oriented to the recognition (and quantification) of the amount of coincidental information in two boundary images. Although some controversies hold on the interpretation of boundaries [67], most authors accept that boundaries are the silhouette of the relevant objects in an image. Often, the position of such silhouettes is hard to determine even for humans, either due to limited resolution in the image or to the very configuration of the scene in terms of lightning, shading, occlusions, etc. Matching for boundary quality evaluation needs to tolerate a certain displacement, which can take place, *a priori*, in any possible direction. However, it is desirable to have some control on the displacement of nearby boundary pixels. For example, it does not seem natural that pixels that are contiguous in one image get matched to pixels that are very distant in the other one. With respect to the expected output, boundary matching for quality evaluation imposes very few restrictions. In fact, the matching does not need to be in terms of an explicit item-by-item matching, but in terms of the amount of common information instead. That is, since the final question to be cleared out is how similar the images are, the correspondence of the boundary pixels at each image does not need to be explicitly enunciated. Any other output is acceptable, as long as it serves as support for computing a subsequent confusion matrix.

The problem of boundary matching, as well as that of quality evaluation, can be put to mathematical terms. In this work, we consider all the images to have generic dimensions $M \times N$, so that the set of positions is $\Omega = \{1, \ldots, M\} \times \{1, \ldots, N\}$. A binary image can be seen both as a mapping $\Omega \to \{0, 1\}$ and as a subset of Ω.

Let E_{cd} and E_{gt} represent a candidate and a ground-truth boundary image, respectively. Classification-based approaches to boundary quality evaluation aim at generating a confusion matrix from the comparison of E_{cd} (probably due to an automated method) with E_{gt} (probably due to a human). In this context, a False Positive (FP) is a boundary pixel in E_{cd} with no correspondence in E_{gt}, a false negative (FN) is a pixel in E_{gt} that is not represented in E_{cd}, etc. Although some authors have proposed to use confusion matrices to compute χ^2 or ROC-derived measures [68, 8], most authors choose the F_α-measure to evaluate the closeness of E_{cd} to E_{gt}. The F_α-measure is given by:

$$F_\alpha = \frac{\textsc{Prec} \cdot \textsc{Rec}}{\alpha\, \textsc{Prec} + (1-\alpha)\, \textsc{Rec}} \tag{1}$$

with $\textsc{Prec} = \frac{\text{TP}}{\text{TP}+\text{FP}}$ and $\textsc{Rec} = \frac{\text{TP}}{\text{TP}+\text{FN}}$, where α is a parameter modulating the relevance of Prec and Rec (typically set to 0.5 [6, 69]), and TP, FP and FN are the usual quantities in a confusion matrix.

The F_α-measure is preferred over other quantities for a list of reasons, including the fact that it avoids using the true negatives (TN), which is typically (much) greater than the other quantities in a confusion matrix for boundary quality evaluation.

Given E_{cd} and E_{gt}, the generation of a confusion matrix is far from trivial. The reason is that the same boundary might appear at nearby, yet not overlapping, positions at each image. When comparing the images in a pixel-to-pixel manner, the non-overlapping boundary pixels at each image would be accounted for as both a FP (those in E_{cd}) and a FN (those in E_{gt}). This is undesirable, since the information contained in both images is similar and, as long as the distance between both appearances of the boundary is not excessive, the boundary should be taken as a correct detection. While some previous studies [7, 70, 71] analyze different quantitative measures to be extracted from a confusion matrix, this work focuses on the strategies for the generation of such matrix; that is, for the correspondence between E_{cd} and E_{gt}.

2.3. A taxonomy for boundary matching techniques

We consider that boundary matching techniques should be classified according to the inspiration they take, not according to the mathematical tools they make use of. Consequently, next, we present a fourfold taxonomy in which the categories are not completely disjoint. At the end of this section, we review such overlappings.

Note that this taxonomy can be seen as a specialization of a part of the taxonomy for error measures in [70, 72], since it appears as a subdivision of the statistical error measures. Also, it relates to the so-called *Confusion matrix-based error assessments* by Adbulrahman et al. [73], and to the statistics derived from confusion matrices in the survey by Magnier [71].

There exist four strategies for boundary pixel matching in the context of boundary evaluation:

a) *Distance-based Matching (DbM).-* This strategy roots in validating (matching) boundary pixels in E_{cd} as long as they are closer to a boundary pixel in E_{gt} than a given threshold [68]. Such pixels become true positive detections, while the remaining pixels in E_{cd} are taken as false positives. In the context of classification, the confusion matrix is given by

$$\begin{aligned} \text{TP} &= \left|\{p \in E_{\text{cd}} \mid d(p, E_{\text{gt}}) \le t\}\right|, \\ \text{FP} &= \left|\{p \in E_{\text{cd}} \mid d(p, E_{\text{gt}}) > t\}\right|, \text{ and} \\ \text{FN} &= \left|\{p \in E_{\text{gt}} \mid d(p, E_{\text{cd}}) > t\}\right|, \end{aligned} \tag{2}$$

where $d(p, E)$ represents the distance from a pixel p to the closest boundary pixel in E, $|\cdot|$ represents the cardinality of a set, and $t \in \mathbb{R}^+$ is the maximum allowed distance between matched pixels[2].

This strategy can also be put in terms of mathematical morphology, as done by Arbelaez in [74] (Ch. 7). In this case, a circular structuring element [75] represents the potential displacement of a boundary pixel, so that a pixel in E_{cd} is validated if it falls within the dilation scope of E_{gt} (and *vice versa*):

$$\begin{aligned} \text{TP} &= \left|E_{\text{cd}} \cap \text{dil}_S(E_{\text{gt}})\right|, \\ \text{FP} &= \left|E_{\text{cd}} \cap \neg\, \text{dil}_S(E_{\text{gt}})\right|, \text{ and} \\ \text{FN} &= \left|E_{\text{gt}} \cap \neg\, \text{dil}_S(E_{\text{cd}})\right|, \end{aligned} \tag{3}$$

[2] Note that $d(p, E)$ can also be seen as the value of the pixel p in the distance transform of E by means of d.

where dil_S is the dilation operation with structuring element S, which here is a circular structuring element of radius t. The results using the formulation in Eq. (2) and that in Eq. (3) are equivalent.

The DbM strategy is extremely simple and computationally cheap, especially if using implementations based on either distance transformation or mathematical morphology. However, it lacks refinement in the discrimination of boundaries and spurious responses. As an example, spurious responses due to texture or noise might be accounted for as true positives when appearing relatively close to the actual boundaries. Hence, although simple and understandable, its use is often disregarded.

b) *Area-based Matching (AbM).-* The AbM grounds on the idea that boundaries have a certain area of influence, which is simply defined as the area surrounding it. Then, it quantifies the TP as the overlapping of the areas of influence of the boundaries at each image.

This strategy renders in a formulation similar to that of the DbM, but allows for a more delicate setting of the allowed displacement. In AbM, the tolerated displacement is modelled by a structuring element [75], leading to

$$\begin{aligned} \text{TP} &= |\,\text{dil}_S(E_{\text{gt}}) \cap \text{dil}_S(E_{\text{cd}})|\,, \\ \text{FP} &= |\,\text{dil}_S(E_{\text{gt}}) \cap \neg\,\text{dil}_S(E_{\text{cd}})|\,, \text{ and} \\ \text{FN} &= |\neg\,\text{dil}_S(E_{\text{gt}}) \cap \text{dil}_S(E_{\text{cd}})|\,, \end{aligned} \tag{4}$$

where S represents the structuring element and dil_S stands, again, for morphological dilation.

Although the formulation in Eq. (4) resembles that in Eq. (3), it yields significantly different results. Firstly, the interpretation of the quantities in the confusion matrices is completely different, since those for DbM represent the number of pixels in the boundaries, but those for AbM represent area sizes. Secondly, displaced boundaries can yield perfect matching in terms of DbM, as long as the displacement is lower than t. However, in AbM, any boundary in E_{cd} slightly displaced from its position in E_{gt} will produce a certain number of FPs and FNs, since $\text{dil}_S(E_{\text{cd}})$ and $\text{dil}_S(E_{\text{gt}})$ will not overlap completely.

Note that both AbM and DbM do not perform explicit matchings between the images. Instead, they count the pixels that are matchable to the counterpart image, which leads to the generation of the confusion matrix.

c) *Correspondence-based matching (CbM).-* This strategy attempts to create an explicit one-to-one matching of the pixels in E_{cd} to those in E_{gt}. Such matching can lead to conclusions on the amount of information in E_{cd} also present in E_{gt} (and vice versa), as well as to the average displacements of the matched pixels [76].

Put to mathematical terms, CbM is presented as the problem of finding a minimal-cost assignment of the pixels in E_{cd} to those in E_{gt}. That is, finding a (largest possible) subset $Q = \{(p_1, q_1), \ldots, (p_n, q_n)\}$ so that $p_i \in E_{\text{cd}}$, $q_i \in E_{\text{gt}}$ and $\sum_{i \in \{1,\ldots,n\}} d(p_i, q_i)$ is minimal. Note that some constraints apply to the set Q. Firstly, each boundary pixel in E_{cd} (resp. E_{gt}) can only be matched to one boundary pixel in E_{gt} (resp. E_{cd}). Secondly, since the number of boundary pixels in each image might be different, the set Q is likely not to cover any of them completely. Thirdly, pairs of pixels $(p_i, q_j) \in \Omega \times \Omega$ are eligible to belong to Q iff $d(p_i, q_j) < t$, where t represents a certain threshold in terms of a metric d. This approach to boundary matching has evident links to both the assignment problem and the transportation problem [12, 77].

Liu and Haralick [76] studied the problem deeply, covering such constraints and proposing strategies to overcome them. Finally, the authors convert the problem into an unconstrained assignment problem by adding *ghost* pixels (to equalize the cardinality of both sets) and setting artificially long distances to the *unmatchable* pairs of pixels. Another detailed analysis of this problem, leading to a different proposal, can be found in [13] (Ch. 3).

Regardless of the algorithm used in the process, CbM leads to the creation of Q, containing the matched pairs of boundary pixels. Once Q is created, the confusion matrix is constructed as:

$$\begin{aligned} \mathrm{TP} &= \left|\{p \in E_{\mathrm{cd}} \mid (p,y) \in Q \text{ for some } y\}\right|, \\ \mathrm{FP} &= \left|\{p \in E_{\mathrm{cd}} \mid (p,y) \notin Q \text{ for any } y\}\right|, \text{ and} \\ \mathrm{FN} &= \left|\{p \in E_{\mathrm{gt}} \mid (x,p) \notin Q \text{ for any } x\}\right|. \end{aligned} \tag{5}$$

CbM has interesting properties and is mathematically sound, but also requires a careful revision. Some of its features are controversial, especially when it comes to the restriction of the one-to-one assignment. On the one hand, this helps avoiding situations in which one single pixel in E_{gt} is matched to (or validates) multiple boundary pixels in E_{cd}, potentially far from each other. Also, the fact that CbM creates an explicit matching allows for the direct computation of related, contextual information, e.g. the average distance between matched boundaries. On the other hand, the fact that each pixel can be matched only once in CbM produces problems in matching slightly displaced boundaries, which are usually composed of a similar, yet different, number of pixels. Although unmatched pixels can be seen as an implicit penalization for the displacement, such penalization is hard to interpret or predict.

Apart from the one-to-one restriction, finding the optimal set Q in CbM is computationally prohibitive. Some alternatives have been presented based on the Hungarian/Munkres algorithm [77], even including pre- and post-processing of E_{cd} and E_{gt} to reduce the computational load. However, most authors use pseudo-optimal algorithms to produce Q, more specifically the implementation of the CSA algorithm [14] distributed within the BSDS300 and the BSDS500 [13, 16].

d) *Pixelwise validation (Pv).-* This strategy takes individual decisions on the validity of each of the boundary pixels in E_{cd}, using information from the surrounding region in both images. The validation process can be expressed as a mapping $\psi : \mathcal{P}(\Omega) \times \mathcal{P}(\Omega) \mapsto \mathcal{P}(\Omega)$ so that $\psi(A,B) \subseteq A$ is the subset of boundary pixels in A that are validated w.r.t. B. Depending on the rules used for validation, ψ might not be symmetric (*i.e.* it might happen that $\psi(A,B) \neq \psi(B,A)$). Given ψ, a confusion matrix can be constructed as

$$\begin{aligned} \mathrm{TP} &= \left|\{p \in E_{\mathrm{cd}} \mid p \in \psi(E_{\mathrm{cd}}, E_{\mathrm{gt}})\}\right|, \\ \mathrm{FP} &= \left|\{p \in E_{\mathrm{cd}} \mid p \notin \psi(E_{\mathrm{cd}}, E_{\mathrm{gt}})\}\right|, \text{ and} \\ \mathrm{FN} &= \left|\{p \in E_{\mathrm{gt}} \mid p \notin \psi(E_{\mathrm{gt}}, E_{\mathrm{cd}})\}\right|. \end{aligned} \tag{6}$$

However, as proposed in [15], it is more natural to skip the confusion matrix and to compute Prec/Rec as

$$\textsc{Prec} = \frac{|\psi(E_{\mathrm{cd}}, E_{\mathrm{gt}})|}{|E_{\mathrm{cd}}|} \text{ and } \textsc{Rec} = \frac{|\psi(E_{\mathrm{gt}}, E_{\mathrm{cd}})|}{|E_{\mathrm{gt}}|}. \tag{7}$$

The Pv strategy has properties of great interest. First, it is able to produce a deterministic one-to-many matching of the boundary pixels in each image at an acceptable cost (among the previous strategies, only certain cases of CbM are able to do so). Secondly, it allows for the application of rather complex, yet meaningful, validation rules, including boundary orientation or interference of different boundaries. However, the fact that it is based on a local analysis makes it computationally inefficient compared to other strategies, especially to those based on mathematical morphology or distance transformations.

The four categories in this taxonomy are not completely disjoint. For example, DbM can be seen as a specialization of Pv in which no information other than the distance between pixels is used. Also, DbM with $d = 1$ would produce results equal to those by AbM with a (rather useless) circular structuring element with radius 1. Still, we consider each category to have different semantics, even if leading to equivalent instantiations.

The theoretical comparison of the four strategies can be performed from different perspectives, none of them providing clear leverage to any of the strategies. Pv matching is the strategy able to use most of the available local information, while all of the other ones exclusively consider the area/length of a potential displacement of each boundary pixel. However, the fact that CbM considers a global solution (instead of pixelwise validation) seems more adequate for the task, since boundaries are semilocal features. With respect to the robustness against small variations in the input, we find AbM to be the most adequate strategy, since it is robust against boundary rugging and progressively penalizes displaced edges; also, this displacement-derived penalty is easily interpretable. Interestingly, DbM, CbM and Pv might judge a boundary pixel to be either completely matchable or completely unmatchable on the basis of a 1-position displacement, (the displacement making the boundary pixel inside or outside the allowed displacement scope). This is certainly undesired. In terms of robustness against the inclusion of noise in one of the images, CbM appears as the best option, since each noisy pixel contributes with at least one unit to TN or FN (maybe both).

On top of the previous considerations, we find that some desirable properties are not provided by any of the strategies. For example, one might consider that neighbouring pixels in a boundary image should be matched to nearby pixels in the other image with some spatial coherence. That is, there should be some spatial coherence between the neighbouring *source* pixels matched in one image and their *destination* pixels in the other. No such proposal have appeared in the literature, and only Pv, due to its flexibility in neighbourhood analysis, seems to allow for similar goals.

Overall, we can conclude that all strategies share common properties, but their different inspirations makes them non-fully-comparable. More important, none of them seems to be, from a qualitative point of view, neither complete enough nor superior to the others.

3. Experimental results

Ideally, we could investigate which is the *best* matching strategy in boundary image comparison, or at least the one that fits better some specific goals. However, there is no clear way of doing so, and, moreover, there is no evident way of generating ground-truth for the boundary matching. As an alternative, the present work questions whether there are practical differences in using different boundary matching techniques. That is, whether there exist significant differences in the results (quality evaluations) obtained when the matching is performed based on different techniques.

The matching methods considered are the following:

- Distance-based Matching (DbM) using the Euclidean metric;
- Area-based Matching (AbM) using a circular structuring element;
- Correspondence-based Matching (CbM) using the CSA algorithm [14];
- Pixelwise validation (Pv) using the constraints by Estrada and Jepson [15].

All the matching methods above have a parameter representing the maximum allowed displacement of a boundary pixel, *i.e.* the maximum distance between two matched pixels. This parameter is embodied differently in each method: the maximum distance in the DbM, the radius of a circular structuring element in the AbM, and the maximum matching distance in both CbM and Pv. Apart from that common parameter, which we refer to as t, the only parameter required in the experiment is the maximum angular distance used by Estrada and Jepson, which we set to $\frac{\pi}{2}$ (see [15] for more details).

3.1. Experiment on the correlation of results

Our first inquiry relates to how similar it is, the evaluation of a candidate image E_{cd} w.r.t. E_{gt} using different matching strategies. Hence, we measure the Pearson correlation of the $\text{F}_{0.5}$ evaluations after performing the matching with each of the considered techniques. The set of images used in the experiment are the ground truth (human-made) images in the BSDS500 set (500 original images, 2696 ground truth images [16]). Note that this dataset contains several human-made solutions for each of the original images.

Figure 1 displays the Pearson correlation between the values of $\text{F}_{0.5}$ in the one-to-one comparison of the images. The results are displayed separately for the intra-class comparisons (comparisons of non-identical images that are ground-truth to the same image), and the inter-class ones. The results are discriminated in this way for two reasons. Firstly, because the matching problem is completely different when a large number of coincidences exists, as is the case of the intra-class comparisons, from when very few elements

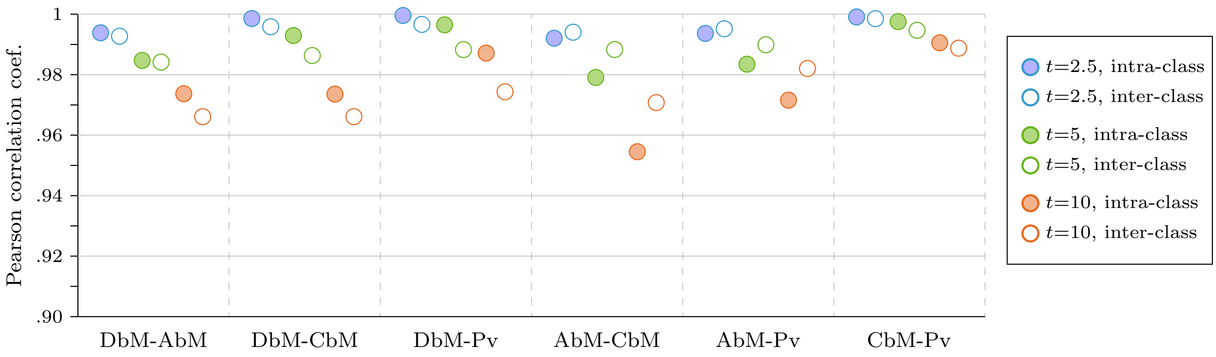


Figure 1: Pearson correlation coefficient between the $F_{0.5}$ values obtained in the comparison of the ground truth images in the BSDS500 using four different strategies for boundary matching. For each pair of strategies and maximum distance between matched pixels, we list the correlation coefficients in the intra-class (left) and inter-class (right) comparisons.

are to be matched. We intend to know how alike matchings are when comparing either similar boundary images (intra-class), or disimilar boundary images (inter-class). Secondly, since the number of inter-class comparisons is about 300 times greater than the number of intra-class ones, a joint display would obfuscate the visibility of the latter[3]. We use $t = 2.5$, $t = 5$ and $t = 10$ pixels, what corresponds to around 0.5%, 1%, 2% of the length of the image diagonal, respectively.

In Figure 1 we observe that there exists a high correlation between the values yielded by any pair of matching strategies, greater than 0.95 for any tested situation. This holds for both intra- and inter-class comparisons. Also, there is a slight decrease in the correlation coefficients as t increases. This is natural, since the opportunities for disparities in the matching of the boundaries increase as t increases. We also observe that the pair AbM-CbM produces slightly lower correlations than the other pairs. Still, the correlation is very high for any possible combination of matching strategies and maximum matching distance.

3.2. Experiment on the comparability of rankings

The results in Section 3.1 indicate that all of the matching techniques lead to highly correlated results. The linear correlation does not necessarily mean that the $F_{0.5}$ values are similar, but we can expect the results with a given matching algorithm to be a scaled version of the others. However, quality evaluation is often not about scoring, but about ranking and/or picking the best contender. Although the results by different matching algorithms are similar, would they lead to similar rankings of boundary image? With this second experiment we intend to shed light on this fact.

Let $\mathbf{R}$ be the set of triplets of unrepeated ground truth images (A, B, C) in the BSDS500 Test Set. That is, every triplet (A, B, C) so that $A \neq B \neq C$. Let q_1 and q_2 be any two comparison measures.

The Equal-Sorting Ratio (ESR) between two measures q_1 and q_2[4] in a dataset $\mathbf{R}$ is

$$\mathrm{ESR}^{\mathbf{R}}_{q_1,q_2} = \frac{|\mathbf{G}|}{|\mathbf{R}|} \ , \tag{8}$$

with $\mathbf{G} \subseteq \mathbf{R}$ defined as

$$\mathbf{G} = \{ r \in \mathbf{R} \mid q_1(A,B) \geq q_1(A,C) \text{ iff } q_2(A,B) \geq q_2(A,C) \} \ . \tag{9}$$

The ESR is hence the proportion of triplets $r = (A, B, C)$ in $\mathbf{R}$ for which the measures q_1 and q_2 agree on whether B or C is closer to A. Hence, it aims at measuring how consistent would rankings be if created with different boundary matching techniques.

[3] Ignoring the comparison of images with themselves, there is around $3 \cdot 10^4$ intra-class and 10^7 inter-class pairs of images to be compared.

[4] In fact, in this experiment, we have the same measure ($F_{0.5}$) embodied with different matching strategies.

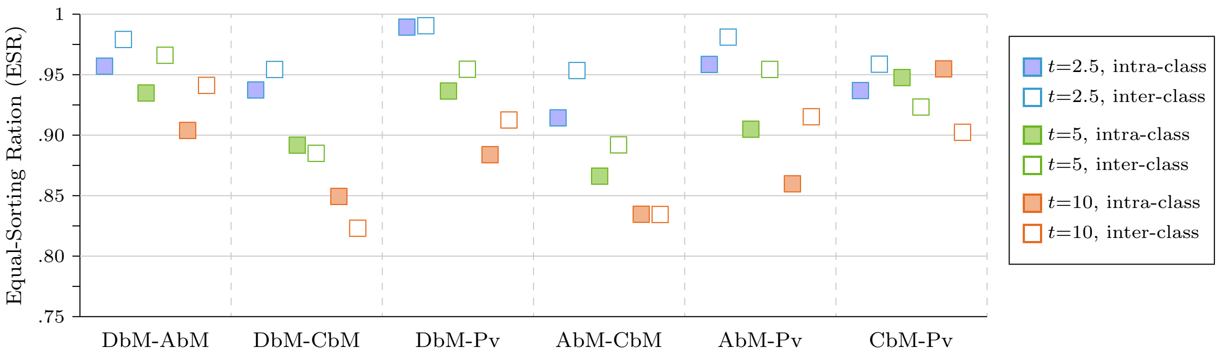


Figure 2: Equal-Sorting Ratio (ESR) obtained in the comparison of the ground truth images in the BSDS500 using four different strategies for boundary matching. For each pair of strategies and maximum distance between matched pixels, we list the ratios for intra-class (left) and inter-class (right) triplets.

Figure 2 displays the ESR for the intra- and inter-class triplets in **R** separately. The former contain all images from the same class, while the latter contain images from at least two different classes. In the BSDS500 Dataset there are over $2 \cdot 10^5$ and 10^{10} of intra- and inter-class triplets, respectively.

Figure 2 displays the ESR for different pairs of matching strategies, and exposes facts similar to those in Figure 1. The ESR is very high for any possible combination of matching strategies and maximum matching distance. As in the first experiment, the increase of t comes coupled to a greater divergence between the matching strategies (in this case, lower ESR). It is also noticeable that the pairs AbM-CbM and AbM-Pv again produce lower ESR values than any other pair, specially for $t = 10$. From the results in Figure 2, we conclude that the rankings obtained by different matching techniques are very similar for any two matching strategies.

3.3. Experiment on computer-generated boundary images

From the experiments in Sections 3.1 and 3.2, we can infer that the matching technique used to create the confusion matrix is nearly irrelevant, since any choice leads to similar conclusions in terms of amount of matched information (Section 3.1), and also fine grained ranking (Section 3.2). However, there is still a question to be posed, related to the validity of the dataset used in the experiments, *i.e.* to the fact that the comparisons are always carried out between human-made images. Truly, the goal of the boundary detection and segmentation communities is to produce methods whose results look like human-made images. Nevertheless, this is not always true, and current experiments on boundary detection might involve the evaluation of images whose characteristics are different from those in the ground truth of the BSDS500. Clearly, the final goal in quality evaluation is not to measure the quality of human-made images, but that of computer-generated ones. We intend to inquiry whether the results in previous sections be replicated using computer-generated boundary images.

The experiments in Sections 3.1 and 3.2 have been repeated using a set of boundary images generated with the Canny method [78]. That is, comparing the human-made images in the BSDS Test Set with those generated automatically. The Canny method has been selected because it is a good representative of boundary detection methods based on gradient magnitude, a historically significant class of methods. Even if it is not a state-of-the art competitor in the BSDS, this fits our intention of using not-very-human boundary images in the comparison. Using more advanced boundary detection methods (as gPb [16]), whose results look more like those by human labellers, would have led to the replication of the results in the previous experiments. The effect of increasing the standard deviation in the Gaussian kernels for the Canny method has a well-known effect; larger standard deviations leads to less frequent, poorly positioned boundaries, embodying the classical precision-recall trade-off in boundary detection methods. The Canny method is hence both significant for its position in the literature and appropriate for the characteristics of its results, which fit the idea of this experiment.

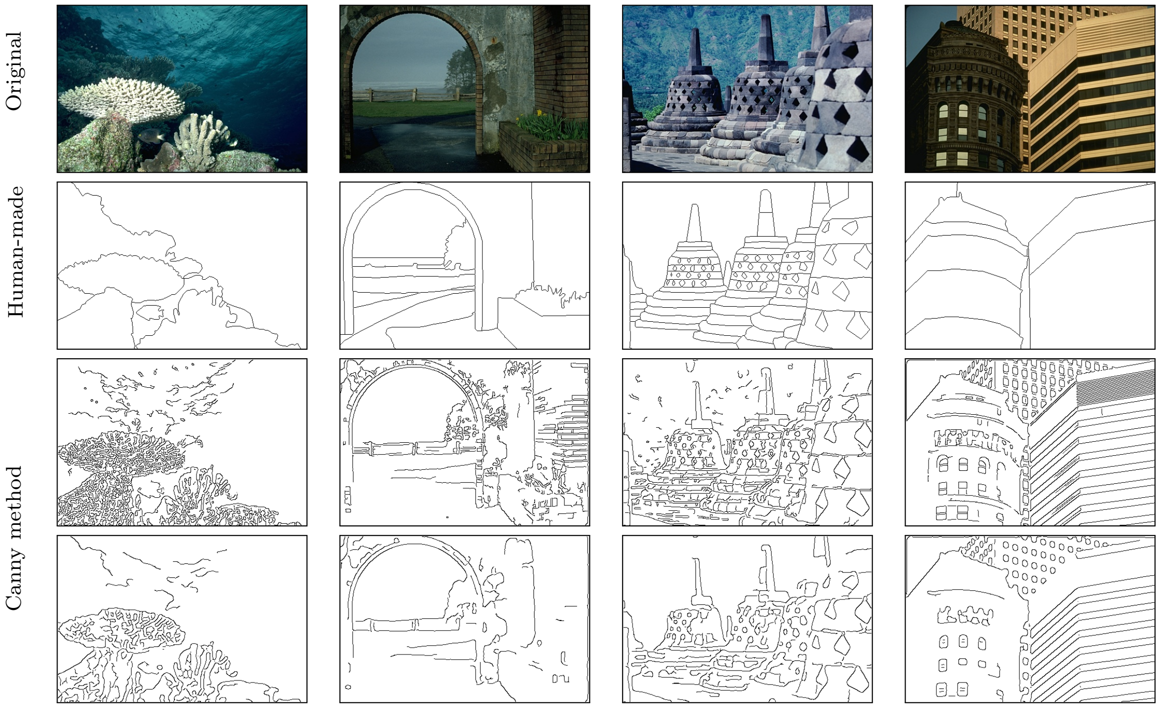
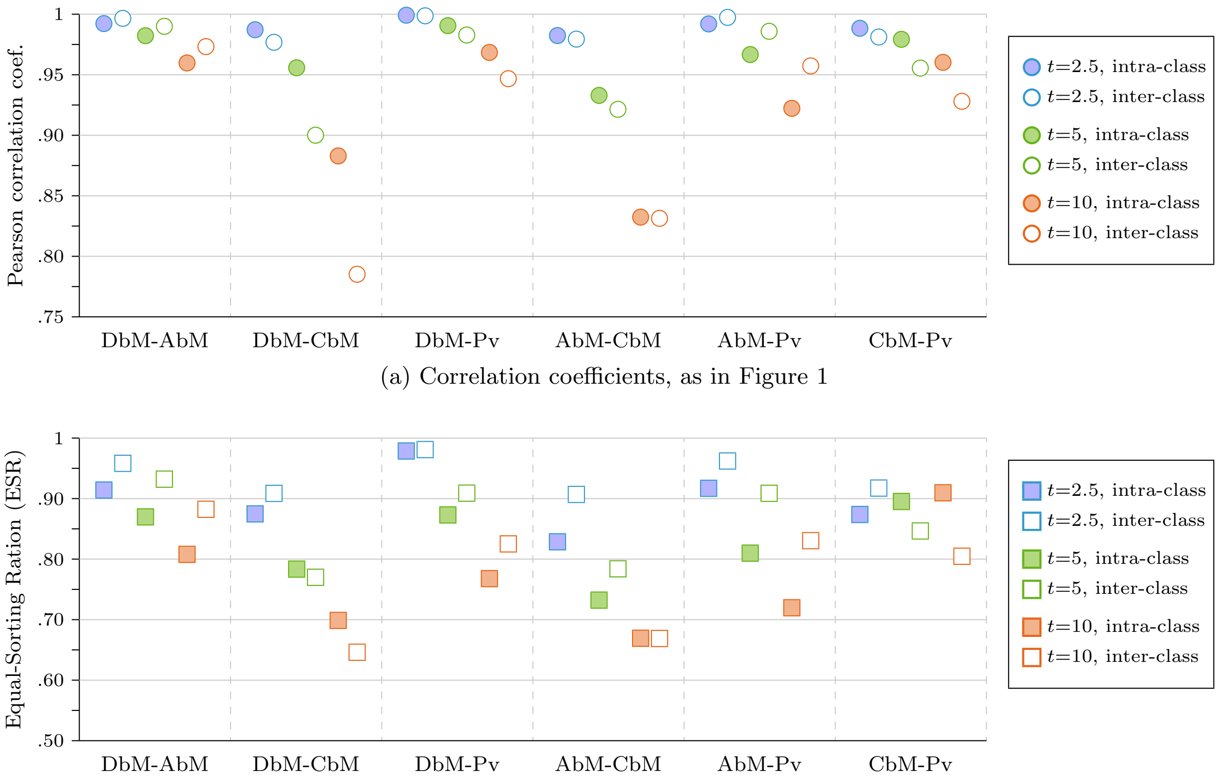


Figure 3: Comparison of boundary image when produced by hand by human labellers or automatic methods. The upmost row contains the original images in the BSDS500 [16]. The second contains human-made ground-truth images. The two lowest rows contain the result by the Canny method on two different configurations: $\sigma_1 = 1, \sigma_2 = 1$ and $\sigma_1 = 2, \sigma_2 = 2$, where $\sigma = 1$ and σ_2 refer to the standard diviation for the smoothing filter and the differentiation filter, respectively.



(a) Correlation coefficients, as in Figure 1

(b) Equal-sorting ratios, as in Figure 2

Figure 4: Repetition of the experiments in Figures 1 and 2, using different images. In this case the comparisons are run between hand-made images in the BSDS500 Test Set and automatically-generated boundary images. The second set is created running the Canny method with 6 different parameter settings on each of the (grayscale) images in the BSDS500 Test Set. For each pair of strategies and maximum distance between matched pixels, we list the ratios for intra-class (left) and inter-class (right) triplets.

In this experiment, intra-class comparisons are those between images generated (either by humans or by the Canny method) from the same original image. In the Canny method we take $\sigma_1 \in \{1, 2\}$ for the Gaussian smoothing and $\sigma_2 \in \{1, 2, 3\}$ for the differentiation kernels. The binarization is performed using NMS, hysteresis and the double-threshold determination technique by Liu [79]. In this manner, the (six) boundary images in each class are similar, but differ in the position (and potentially, in the selection) of some boundaries due to the variation in the kernel sizes[5]. Moreover, the intra-class divergences are due to the parameter setting, not to human interpretation, what leads to a more realistic scenario. In fact, as explained before the Canny method is preferred over other alternatives because of how it allows a trade-off between common errors as texture false detection, edge displacement, edge breaking, etc.

Some examples of the differences between the human-made images and those obtained with the Canny method are shown in Fig. 3. We can observe that, in textured areas, human typically label no boundaries (first, second and third column from the left), while the Canny method might do it. Also, salient, but semantically unimportant structures, might also be tagged by algorithms, yet not by humans. Examples of such are the windows on the building facades in the rightmost column of Fig. 3.

In the replication of the experiment in Section 3.2, triplets are created so that (A, B, C), with A a human-made ground truth and B, C automatically-generated images by the Canny method. In this way we

[5] This set of boundary images is, as well as the code of the comparisons, available at [80].

simulate a realistic setup in which two computer-generated images need to be ranked taken a human-made one as reference.

The results gathered in the comparisons are listed in Figure 4. These results illustrate a rather mild decrease of correlation and, more sharply, of ESR. Again, this divergences increase along with t. This indicates that, even if results are highly correlated, the small perturbations in the $F_{0.5}$ values might lead to different rankings. A direct hypothesis which would explain all results is that missortings in Fig. 4(b) are decided by very small margins. Hence, we attempt to verify it.

Let $r = (A, B, C)$ be any triplet of images. According to the definition of ESR, r is missorted by two comparison measures q_1 and q_2 iff

$$(q_1(A, B) - q_1(A, C)) \cdot (q_2(A, B) - q_2(A, C)) < 0. \qquad (10)$$

Capitalizing on this idea, given two error measures q_1 and q_2, we define the Sorting Margin (SM) of a triplet $t = (A, B, C)$ as:

$$\mathrm{SM}_{q_1,q_2}(r) = \mathrm{sign}(\alpha) \cdot \sqrt{|\alpha|}\ , \qquad (11)$$

with

$$\alpha = (q_1(A, B) - q_1(A, C)) \cdot (q_2(A, B) - q_2(A, C)). \qquad (12)$$

The SM is positive for well-sorted triplets, and becomes negative in case of missorted ones. At the same time, negative SMs of small absolute value indicate marginal missortings, while SMs of large absolute value are due to severe differences between q_1 and q_2. For every pair of comparison measures and matching distance t we have computed the distribution of SM over the triplets in $\mathbf{R}$. Figure 5 displays the distribution of triplets with negative SM for each combination of q_1, q_2 and matching distance t. In this figure we can see how most of the missorted triplets are so by a margin which rarely exceeds 0.03. In the cases in which the ESR is negative, it is normally close to zero. For example, in the case of CbM-DbM, with $t = 5$, the 2.5 percentile stays in -0.049, being the lowest 2.5 percentile of the distributions represented in the figure. Overall, it can be concluded that the SM is always relatively low.

Three factors can, still, raise the SM. First, increasing the matching distance t increases the SM, as seen in Fig. 5. Second, interclass triplets (which are less representative for real comparison) normally yield lower SM, meaning that more realistic (intraclass) comparisons are missorted by smaller margins. Third, some pairs of matching strategies clearly produce greater ESR and SM than others. A paradigmatic case is that of CbM-DbM, mostly due to the fact that CbM enforces a 1-to-1 correspondence of matched pixels, while DbM allows any number of nearby pixels to be matched. This, in images containing strong textures near actual edges, can lead to significant variability in the matching and, hence, in the measured F. Such cases are, however, rare over a standard dataset as the BSDS, and mostly produced by high-frequency gradient characterization filters (in this experiment, the Canny method with $\sigma_1 = \sigma_2 = 1$).

Figure 6 contains one of the triplets having a greatest SM in the overall comparison. The triplet composed by those three images produces, for example, a SM of -0.094 when using the strategies DbM-CbM with matching distance $t = 5$. It can be observed that the image recaps the characteristics that can create a perfect storm in terms of discrepancies between different matching strategies. There is a large number of objects which human labellers ignore due to contextual facts. For example, the public leaning on the fence, which is relatively salient, yet contextually unimportant. Also, it combines two facts that create divergent results by different matching strategies: (a) strong textures being detected near edge structures, so that strategies as DbM or AbM will count as TPs boundary pixels which will become FPs in CbM; (b) multiple lines that are uniquely tagged by users as a single line, yet being multiply labelled by automatic methods. Examples of the former are the people in the background, or the elements in the cyclist bodies, while the latter manifests at the limits of the cycling track.

Figure 7 contains a visual representation of the matching by each strategy in the triplet featured in Fig. 6, setting the matching distance to $t = 5$. Specifically, for each image and matching method, two areas are highlighted for better illustration of the relevant characteristics of the images and the output of the matching strategies. In the two upper rows we observe how strong textures near human-labelled boundaries

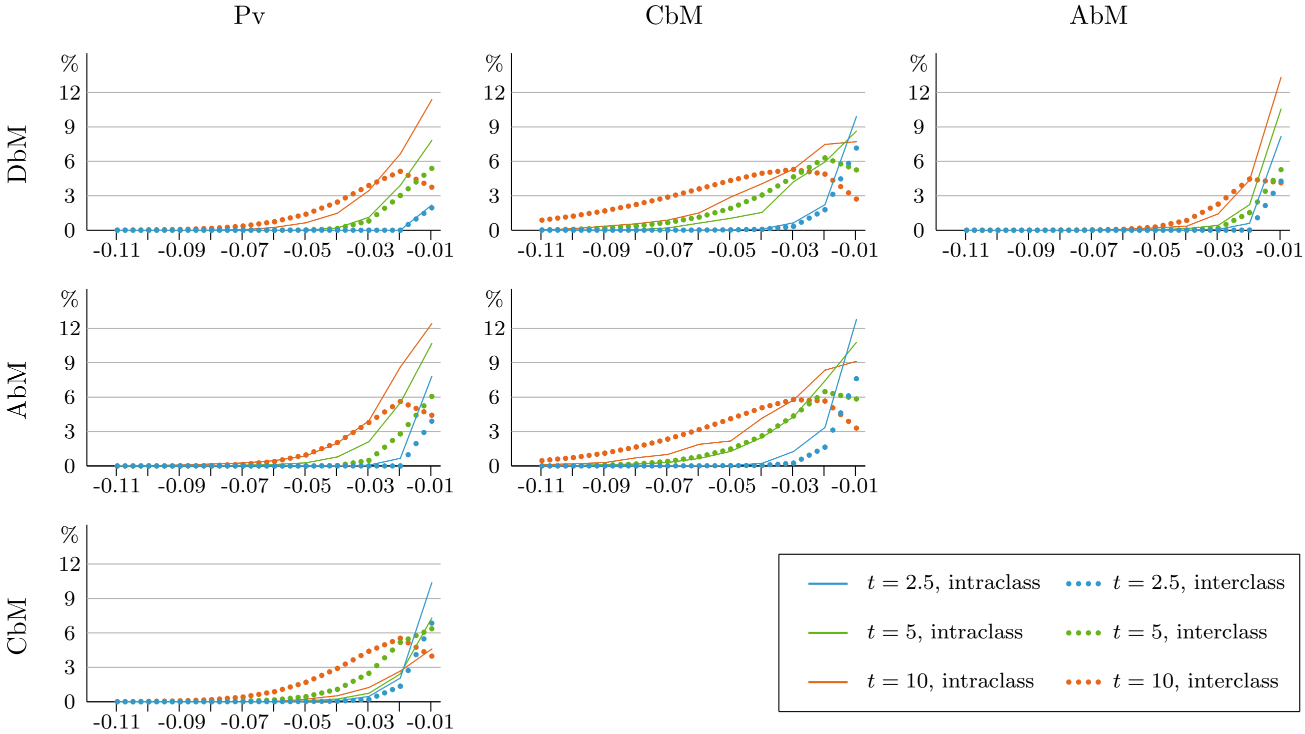
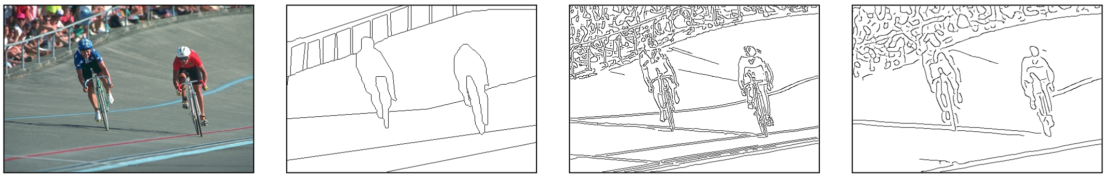


Figure 5: Percentual distribution of triplets with negative ESM when combining different matching strategies and matching distances. The images used for the experiment are those in the BSDS500 Test Set, compared to automated Canny method-generated solutions, as in Figure 4. The distribution is binned with 0.01 granularity. It is clearly seen that, even in case of a large ESR (as is the case for CbM-DbM and AbM-CbM), the ESM is mostly in the range $[0, 0.03]$. Triplets for which the ESM is greater than 0.05 are rather unusual.

.



Original Image Human-labelled GT Canny, $\sigma_1 = \sigma_2 = 1$ Canny, $\sigma_1 = 2$, $\sigma_2 = 3$

Figure 6: Ground truth image 226022_05 extracted from the BSDS500 Test Set, together with two automatically-generated images using the Canny method. The parameter setting of the Canny method are specified for each image, with σ_1 representing the standard deviation of the regularization filter and σ_2 represneting the standard deviation of the differentiation filter.

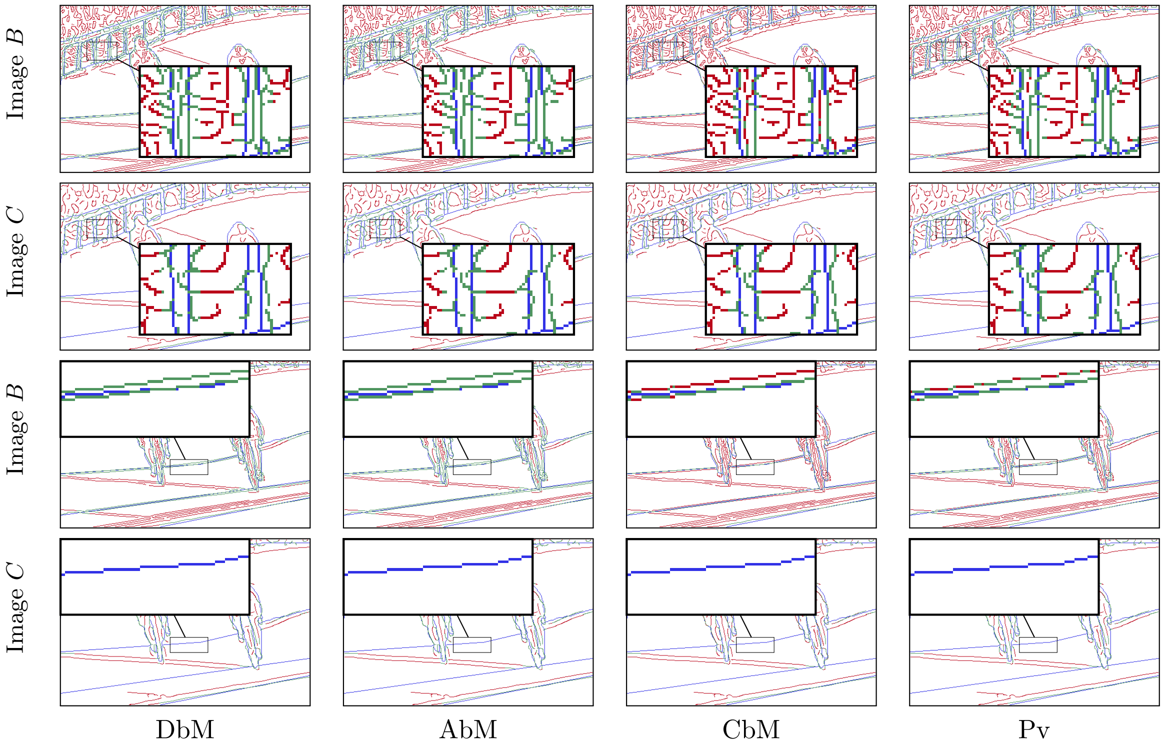


Figure 7: Visual comparison of the matching made on the triplet of images in Fig. 6. Each image displays the result of the matching, so that (a) green pixels are validated boundaries, (b) pink pixels are unmatched ones and (c) blue pixels are the boundaries in the ground truth.

produce a variable response depending on the matching strategies. This is mostly due to the fact that CbM restricts the matching to be a 1-to-1 correspondence, which in this case results in judicious responses. In the two lower rows we observe the behaviour of the matching strategies when multiple boundaries appear. While DbM or AbM validate all responses near *actual* boundaries, CbM restricts the matching (again) to a 1-to-1 correspondence. The behaviour by DbM or AbM seems, in this case, more appropriate. Anyhow, the combination of both types of situations lead to divergent interpretations by different matching strategies.

In general, even if some triplets can show the problematics seen in the triplet in Fig. 6, such cases are exceptional. In order for a triplet to generate a SM under -0.03, a list of circumstances must co-occur, including poor performance by the automatic method, and adversary situations in the image (as those seen in Fig. 7) leading to variable interpretation by the matching algorithms. In general, most of the missorted triplets produce SMs in the range [-0.03,0], what proves that such co-ocurrences are rather unusual. This, additionally, explains the results in Figure 4.

4. Discussion

In this work, we have reviewed the problem of boundary matching for boundary image quality evaluation. This problem can be seen as an instantiation of a frequent problems in image processing or graph theory, but the majority of the practical solutions found in literature are inapplicable to the present task. We have proposed a novel taxonomy for the different strategies for boundary matching present in the literature. Finally, we have questioned whether the different strategies might lead to significantly different (quantitative) results in boundary quality evaluation.

We have found that most of the matching strategies lead to similar results in terms of evaluation, despite their fundamental differences in inspiration and realization. Overall, almost all strategies hold very high correlation in terms of the $F_{0.5}$ measure (Figures 1 and 4(a)). Also, they produce very similar rankings (Figures 2 and 4(b)) and, when rankings by different strategies are heterogeneous, such difference is due to very small rankings (Figure 5). Apart from exceptional cases[6], any comparison produces a Pearson correlation coefficient ≥ 0.9 and ESR ≥ 0.75. Two addendums might be made to such conclusion: (1) The theoretical properties by different strategies might differ greatly. If, for example, a researcher is interested in keeping the pixel-to-pixel matching, CbM appears as a much better option. (2) Divergences between matching strategies are enlarged when either (a) the boundary images to be matched/evaluated contain a large number of spurious responses or (b) the maximum matching distance between boundary pixels is long, or both.

Given these findings, our recommendation to select a boundary matching strategy in a real problem consists of three steps. Firstly, it is necessary to analyze the characteristics of the images to be matched. With such analysis, and considering whether one or more of the theoretical properties might make a strategy preferable to any other (see Figure 7). Secondly, it is important to weight in the use of the output by each matching. For example, CbM or Pv produce a correspondence between matched pixels, which might be of interest for some applications, while DbM or AbM do not. Finally, one might consider the computational efficiency of each strategy. Most of the algorithms upon which these strategies rely (distance transformation, mathematical morphology operators, constrained optimization,...) are pre-built in most computing packages, hardware-accelerated for computing clusters, and extensively dependant on memory use/access operations. Hence, a study on the computational efficiency is hardly portable to the reality of current state-of-the-art. Still, as of today, DbM and AbM tend to be faster than CbM and Pv.

Acknowledgements

The authors gratefully acknowledge the financial support of the Spanish Ministry of Science (project TIN2016-77356-P, AEI/FEDER, UE), ass well as that of the Research Foundation Flanders (FWO project 3G.0838.12.N).

References

[1] J. Fram, E. S. Deutsch, Quantitative evaluation of edge detection algorithms and their comparison with human performance, IEEE Trans. on Computers 24 (6) (1975) 616–628.

[2] A. Herskovits, T. O. Binford, On boundary detection, Tech. rep., Massachusetts Institute of Technology, project MAC Memo no. 183 (1970).

[3] I. Abdou, W. Pratt, Quantitative design and evaluation of enhancement/thresholding edge detectors, Proceedings of the IEEE 67 (5) (1979) 753–763.

[4] A. J. Baddeley, An error metric for binary images, in: W. Förstner, S. Ruwiedel (Eds.), Robust Computer Vision: Quality of Vision Algorithms, Wichmann Verlag, Karlsruhe, 1992, pp. 59–78.

[5] G. Borgefors, Distance transformations in digital images, Computer Vision, Graphics, and Image Processing 34 (3) (1986) 344–371.

[6] C. Lopez-Molina, B. De Baets, H. Bustince, Quantitative error measures for edge detection, Pattern Recognition 46 (4) (2013) 1125–1139.

[7] R. Koren, Y. Yitzhaky, Automatic selection of edge detector parameters based on spatial and statistical measures, Computer Vision and Image Understanding 102 (2) (2006) 204–213.

[8] Y. Yitzhaky, E. Peli, A method for objective edge detection evaluation and detector parameter selection, IEEE Trans. on Pattern Analysis and Machine Intelligence 25 (8) (2003) 1027–1033.

[9] T. Fawcett, An introduction to ROC analysis, Pattern Recognition Letters 27 (8) (2006) 861–874.

[10] F. J. Provost, T. Fawcett, Analysis and visualization of classifier performance: Comparison under imprecise class and cost distributions., in: Proc. of the International Conf. on Knowledge Discovery & Data Mining, Vol. 97, 1997, pp. 43–48.

[11] W. Waegeman, B. De Baets, L. Boullart, ROC analysis in ordinal regression learning, Pattern Recognition Letters 29 (1) (2008) 1–9.

[12] H. W. Kuhn, The Hungarian method for the assignment problem, Naval Research Logistics Quarterly 2 (1-2) (1955) 83–97.

[6] Specifically, the comparison of the most divering strategies, CbM and AbM/DbM, which take totally different strategies in near-boundary noise validation, combined with the setting $t = 10$.

[13] D. R. Martin, An empirical approach to grouping and segmentation, Ph.D. thesis, University of California, Berkeley (2003).

[14] A. V. Goldberg, R. Kennedy, An efficient cost scaling algorithm for the assignment problem, Mathematical Programming 71 (1995) 153–177.

[15] F. J. Estrada, A. D. Jepson, Benchmarking image segmentation algorithms, International Journal of Computer Vision 85 (2) (2009) 167–181.

[16] P. Arbelaez, M. Maire, C. Fowlkes, J. Malik, Contour detection and hierarchical image segmentation, IEEE Trans. on Pattern Analysis and Machine Intelligence 33 (2011) 898–916.

[17] T. Serre, L. Wolf, T. Poggio, Object recognition with features inspired by visual cortex, Tech. rep., Dept. of Brain and Cognitive Sciences, Massachusets Inst. of Technology (2006).

[18] M. Riesenhuber, T. Poggio, Hierarchical models of object recognition in cortex, Nature Neuroscience 2 (11) (1999) 1019.

[19] D. Marr, Vision, MIT Press, 1982.

[20] R. N. Shepard, J. Metzler, Mental rotation of three-dimensional objects, Science 171 (3972) (1971) 701–703.

[21] J. Metzler, R. N. Shepard, Theories in cognitive psychology: The Loyola Symposium, Lawrence Erlbaum, 1974, Ch. Transformational studies of the internal representation of three-dimensional objects.

[22] S. G. Vandenberg, A. R. Kuse, Mental rotations, a group test of three-dimensional spatial visualization, Perceptual and Motor Skills 47 (2) (1978) 599–604.

[23] D. Marr, E. Hildreth, Theory of edge detection, Proceedings of the Royal Society of London 207 (1167) (1980) 187–217.

[24] F. Zana, J.-C. Klein, A multimodal registration algorithm of eye fundus images using vessels detection and Hough transform, IEEE Trans. on Medical Imaging 18 (5) (1999) 419–428.

[25] G. P. Boswell, F. A. Davidson, Modelling hyphal networks, Fungal Biology Reviews 26 (1) (2012) 30–38.

[26] G. Vidal-Diez de Ulzurrun, J. Baetens, J. Van den Bulcke, B. De Baets, Modelling three-dimensional fungal growth in response to environmental stimuli, Journal of Theoretical Biology 414 (2017) 35–49.

[27] C. Baillard, C. Schmid, A. Zisserman, A. Fitzgibbon, Automatic line matching and 3D reconstruction of buildings from multiple views, in: ISPRS Conf. on Automatic Extraction of GIS Objects from Digital Imagery, Vol. 32, 1999, pp. 69–80.

[28] F. Tupin, H. Maitre, J.-F. Mangin, J.-M. Nicolas, E. Pechersky, Detection of linear features in SAR images: Application to road network extraction, IEEE Trans. on Geoscience and Remote Sensing 36 (2) (1998) 434–453.

[29] A. Kong, D. Zhang, M. Kamel, A survey of palmprint recognition, Pattern Recognition 42 (7) (2009) 1408–1418.

[30] J. Dai, J. Feng, J. Zhou, Robust and efficient ridge-based palmprint matching, IEEE Trans. on Pattern Analysis and Machine Intelligence 34 (8) (2012) 1618–1632.

[31] A. N. Marana, A. K. Jain, Ridge-based fingerprint matching using hough transform, in: Proc. of the Brazilian Symposium on Computer Graphics and Image Processing, 2005, pp. 112–119.

[32] K. Mikolajczyk, A. Zisserman, C. Schmid, Shape recognition with edge-based features, in: Proc. of the British Machine Vision Conference, Vol. 2, 2003, pp. 779–788.

[33] S. Jaggi, W. C. Karl, S. G. Mallat, A. S. Willsky, Silhouette recognition using high-resolution pursuit, Pattern Recognition 32 (5) (1999) 753–771.

[34] Z. Wang, Z. Chi, D. Feng, Shape based leaf image retrieval, in: IEE Proc.- Vision, Image and Signal Processing, Vol. 150, IET, 2003, pp. 34–43.

[35] Z. Liu, S. Sarkar, Effect of silhouette quality on hard problems in gait recognition, IEEE Trans. on Systems, Man, and Cybernetics, Part B: Cybernetics 35 (2) (2005) 170–183.

[36] O. Gómez, O. Ibáñez, A. Valsecchi, O. Cordón, T. Kahana, 3D-2D silhouette-based image registration for comparative radiography-based forensic identification, Pattern Recognition 83 (2018) 469–480.

[37] A. Kumar, D. Zhang, Personal recognition using hand shape and texture, IEEE Trans. on Image Processing 15 (8) (2006) 2454–2461.

[38] D. Scharstein, R. Szeliski, A taxonomy and evaluation of dense two-frame stereo correspondence algorithms, International Journal of Computer Vision 47 (1-3) (2002) 7–42.

[39] D. Marr, T. Poggio, A theory of human stereo vision., Tech. Rep. No. AI-M-451, Massachusetts Institute of Technology (1977).

[40] G. Medioni, R. Nevatia, Segment-based stereo matching, Computer Vision, Graphics, and Image Processing 31 (1) (1985) 2–18.

[41] N. M. Nasrabadi, A stereo vision technique using curve-segments and relaxation matching, IEEE Trans. on Pattern Analysis and Machine Intelligence (5) (1992) 566–572.

[42] C. Xu, L. Zhang, L. Cheng, R. Koch, Pose estimation from line correspondences: A complete analysis and a series of solutions, IEEE Trans. on Pattern Analysis and Machine Intelligence 39 (6) (2017) 1209–1222.

[43] D. Maltoni, D. Maio, A. K. Jain, S. Prabhakar, Handbook of fingerprint recognition, Springer-Verlag, 2009.

[44] Y. Liu, T. S. Huang, O. D. Faugeras, Determination of camera location from 2-D to 3-D line and point correspondences, IEEE Trans. on Pattern Analysis and Machine Intelligence 12 (1) (1990) 28–37.

[45] D. Cremers, T. Kohlberger, C. Schnörr, Shape statistics in kernel space for variational image segmentation, Pattern Recognition 36 (9) (2003) 1929–1943.

[46] F. Mokhtarian, A. Mackworth, Scale-based description and recognition of planar curves and two-dimensional shapes, IEEE Trans. on Pattern Analysis and Machine Intelligence (1) (1986) 34–43.

[47] L. J. Latecki, R. Lakämper, Application of planar shape comparison to object retrieval in image databases, Pattern Recognition 35 (1) (2002) 15–29.

[48] F. X. Chen, G. Roig, L. Isik, X. Boix, T. Poggio, Eccentricity dependent deep neural networks: Modeling invariance in human vision, in: AAAI Spring Symposium Series, 2017.

[49] Y. Han, G. Roig, G. Geiger, T. Poggio, Is the human visual system invariant to translation and scale?, in: AAAI Spring Symposium Series, 2017.

[50] F. Mokhtarian, A. K. Mackworth, A theory of multiscale, curvature-based shape representation for planar curves, IEEE Trans. on Pattern Analysis and Machine Intelligence (8) (1992) 789–805.

[51] R. Deriche, O. Faugeras, Tracking line segments, in: Proc. of the European Conference on Computer Vision, 1990, pp. 259–268.

[52] J. C. McEachen, J. S. Duncan, et al., Shape-based tracking of left ventricular wall motion, IEEE Trans. on Medical Imaging 16 (3) (1997) 270–283.

[53] M. Yokoyama, T. Poggio, A contour-based moving object detection and tracking, in: Proc. of the International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005, pp. 271–276.

[54] A. Yilmaz, X. Li, M. Shah, Contour-based object tracking with occlusion handling in video acquired using mobile cameras, IEEE Trans. on Pattern Analysis and Machine Intelligence 26 (11) (2004) 1531–1536.

[55] C. Schmid, A. Zisserman, Automatic line matching across views, in: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, 1997, pp. 666–671.

[56] J. H. Han, J. S. Park, Contour matching using epipolar geometry, IEEE Trans. on Pattern Analysis and Machine Intelligence 22 (4) (2000) 358–370.

[57] J. Ha, Real-time visual tracking using image processing and filtering methods, Ph.D. thesis, Georgia Institute of Technology (2008).

[58] Y. Wang, E. K. Teoh, D. Shen, Lane detection and tracking using B-Snake, Image and Vision Computing 22 (4) (2004) 269–280.

[59] F. Bergholm, Edge focusing, IEEE Trans. on Pattern Analysis and Machine Intelligence 9 (6) (1987) 726–741.

[60] A. P. Witkin, Scale-space filtering, in: Proc. of the International Joint Conf. on Artificial Intelligence, Vol. 2, 1983, pp. 1019–1022.

[61] G. Papari, P. Campisi, N. Petkov, A. Neri, A biologically motivated multiresolution approach to contour detection, EURASIP Journal on Advances in Signal Processing 2007, article ID 71828.

[62] T. Lindeberg, Edge detection and ridge detection with automatic scale selection, International Journal of Computer Vision 30 (2) (1998) 117–156.

[63] P. Perona, J. Malik, Scale-space and edge detection using anisotropic diffusion, IEEE Trans. on Pattern Analysis and Machine Intelligence 12 (7) (1990) 629–639.

[64] T. Poggio, J. Mutch, L. Isik, Computational role of eccentricity dependent cortical magnification, Tech. rep., Center for Brains, Minds and Machines, Massachusets Inst. of Technology (2014).

[65] H. Kuwajima, M. Tanaka, M. Okutomi, Improving transparency of deep neural inference process, Progress in Artificial Intelligence 8 (2) (2019) 273–285.

[66] A. Nguyen, J. Yosinski, J. Clune, Deep neural networks are easily fooled: High confidence predictions for unrecognizable images, in: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, 2015, pp. 427–436.

[67] G. Papari, N. Petkov, Edge and line oriented contour detection: State of the art, Image and Vision Computing 29 (2-3) (2011) 79–103.

[68] K. Bowyer, C. Kranenburg, S. Dougherty, Edge detector evaluation using empirical ROC curves, Computer Vision and Image Understanding 84 (1) (2001) 77–103.

[69] D. Martin, C. Fowlkes, J. Malik, Learning to detect natural image boundaries using local brightness, color, and texture cues, IEEE Trans. on Pattern Analysis and Machine Intelligence 26 (5) (2004) 530–549.

[70] C. Lopez-Molina, B. De Baets, H. Bustince, J. Sanz, E. Barrenechea, Multiscale edge detection based on Gaussian smoothing and edge tracking, Knowledge-Based Systems 44 (2013) 101–111.

[71] B. Magnier, Edge detection: a review of dissimilarity evaluations and a proposed normalized measure, Multimedia Tools and Applications 77 (8) (2018) 9489–9533.

[72] C. Lopez-Molina, B. De Baets, H. Bustince, Twofold consensus for boundary detection ground truth, Knowledge-Based Systems 98 (2016) 162–171.

[73] H. Abdulrahman, B. Magnier, P. Montesinos, From contours to ground truth: How to evaluate edge detectors by filtering, Journal of the World Society for Computer Graphics 25 (2) (2017) 133–142.

[74] P. A. Arbelaez, Une approche métrique pour la segmentation d'images, Ph.D. thesis, Université Paris-Dauphine (2005).

[75] R. M. Haralick, S. R. Sternberg, X. Zhuang, Image analysis using mathematical morphology, IEEE Trans. on Pattern Analysis and Machine Intelligence 9 (4) (1987) 532–550.

[76] G. Liu, R. M. Haralick, Optimal matching problem in detection and recognition performance evaluation, Pattern Recognition 35 (10) (2002) 2125–2139.

[77] J. Munkres, Algorithms for the assignment and transportation problems, Journal of the Society for Industrial and Applied Mathematics 5 (1) (1957) 32–38.

[78] J. Canny, A computational approach to edge detection, IEEE Trans. on Pattern Analysis and Machine Intelligence 8 (6) (1986) 679–698.

[79] X. Liu, Y. Yu, B. Liu, Z. Li, Bowstring-based dual-threshold computation method for adaptive Canny edge detector, in: Proc. of the International Conf. of Image and Vision Computing New Zealand, 2013, pp. 13–18.

[80] KERMIT Research Unit (Ghent University), B. De Baets, C. Lopez-Molina (eds.) The Kermit Image Toolkit (KITT). URL www.kermitimagetoolkit.com