Supervised Scale-Invariant Segmentation (and Detection)

Yan Li, David M.J. Tax, and Marco Loog[*]

Pattern Recognition Laboratory
Delft University of Technology
The Netherlands
{yan.li,d.m.j.tax,m.loog}@tudelft.nl
http://prlab.tudelft.nl

Abstract. The scale-invariant detection of image structure has been a topic of study within computer vision and image analysis since long. To date, Lindeberg's scale selection method has probably been the most fruitful and successful approach to this problem. It provides a general technique to cope with the detection of structures over scale that can be successfully expressed in terms of Gaussian differential operators. Any detection or segmentation task would potentially benefit from a similar approach to deal with scale. For many of the real-world image structures of interest, however, it will often be impossible to explicitly design or handcraft an operator that is capable of detecting them in a sensitive and specific way. In this paper, we present an approach to the scale-selection problem in which the construction of the detector is driven by supervised learning techniques. The resulting classification method is designed so as to achieve scale-invariance and may be thought of as a supervised version of Lindeberg's classical scheme.

Keywords: Scale selection, scale-invariance, image segmentation, detection, learning and classification.

1 Introduction

Image structures, such as blobs, edges or corners, may appear in images at different scales. To detect them, it is often desired for a detector to select the locally appropriate scales. A well-known scale selection scheme was proposed by Lindeberg [18,17] for image structures which can be detected by differential operators, such as the Laplacian, the Hessian, etc. [27]. The operator under consideration is multiplied with a scale-dependent normalization factor, i.e., it is scale normalized, and applied to an image to get a response at all scales and locations. Subsequently, the scale where the normalized detector attains the maximum response over scales is selected as the local scale of the structure.

[*] Partly supported by the Innovational Research Incentives Scheme of the Netherlands Research Organization [NWO, VENI Grant 639.021.611]

A.M. Bruckstein et al. (Eds.): SSVM 2011, LNCS 6667, pp. 350–361, 2012.

Scale selection schemes have been at the basis of many successful computer vision and image precessing applications [10,21,22,1,24]. A potential problem, however, is that the schemes are merely applicable to the detection of relatively simple structures. For more complicated or very specific structures, it will often be impossible to explicitly design or handcraft an operator that is capable to detect these. Examples of such structures range from blobs that are textured or have a particular shape to faces, bikes, cars, potted plants, or other image objects.

Next to scale selection, scale-invariance is a desired property in many computer vision and image analysis tasks because an input image can have an arbitrary and unknown inner scale. Informally, the ‘inner scale’ of a pixel is proportional to the area in the real world that the pixel represents [8]. Employing the proper scale normalization, differential operators in combination with Lindeberg’s scale selection are indeed scale-invariant [18]. As with scale selection, many more advanced computer vision techniques rely, at a lower level, on some form of Lindeberg’s approach to make the overall scheme scale-invariant as well [21,22,1,24]. Other, more committed, attempts to achieve scale-invariance are to offset scaling with the log-polar and Fourier transform [16,25,15] or to incorporate features from various scales and estimate the local scales of the image under consideration [13,14].

1.1 Work’s Novelty and Related Methods

This paper develops a supervised learning approach [4,2] that allows one to construct nontrivial, scale-invariant detection, classification, or segmentation approaches based on available training data in combination with general machine learning and pattern recognition methods. All in all, the approach proposed can be seen as a supervised variation to Lindeberg’s classical scheme [18,17].

One critical advantage of our proposal is that learning techniques enable one to develop methods that can potentially handle the more complex structures encountered in real-world segmentation or detection tasks. Like for any supervised learning scheme, in order to apply the technique one needs examples of the task to solve, i.e., a training set. That is, we need to have a collection of raw images, e.g. X-rays, and the desired corresponding output one would like to obtain from them, e.g. an expert segmentation, in order for the learner, e.g. a classifier, to be able to capture the desired relationship. Now, a second advantage is in fact that our approach allows the user to pick the classifier and features of his or her liking. A third critical advantage is that scale selection is made task-dependent by integrating supervision into the selection process. The reason for doing so is that, even when the image data remains the same, different tasks may require different scales to solve them at. Current selection schemes, which are all unsupervised, obviously cannot accommodate this.

Also closely related to our work are face detection schemes that, at test phase, take care of scale and location variations simply by applying the detector to all scales and locations and afterwards finding its maximum responses [12,28]. In a sense, these are supervised approaches that follow Lindeberg’s scheme as well. A crucial difference with our approach, however, is in the training of this detector. The face detection techniques need a set of scale and location aligned faces at training phase, which basically takes care of the problem of scale. In many segmentation and detection setting, however, it is difficult to properly align different training instances. Take for instance any medical image segmentation task, how could one identify, even within a single image, the appropriate scale from location to location? Our approach solves the scale selection problem implicitly and does not rely on any a priori knowledge about inner scales in the training or test phase.

The problem covered in the current work has also been discussed in [20], where a supervised method was proposed by viewing classifiers as special types of scale-dependent structure detectors or filters based on which some sort of scale selection could be performed. One of the main shortcomings of this approach, however, is that it is not scale-invariant. A key contributions of this work is to remove this restriction.

1.2 Remark and Outline

To avoid confusion, we use the word structure to mean an image feature, e.g. blobs, edges, or more complicated structures, and the word feature refers to the supervised learning setting where it can mean any kind of measurement that can be made in an image, e.g., Gaussian derivatives, N-jets, differential invariants, texture features, etc.

The remainder of the paper is organized as follows. The next section sets the stage more specifically, it provides some notations used in the paper, and sketches the basics of supervised pixel-based segmentation techniques. Section 3 describes our proposed method. Some illustrative experiments can be found in Section 4. Section 5 concludes the paper.

2 Scale Space Theory and Pixel-Based Segmentation

2.1 Scale Space and Gaussian Derivatives

We will employ linear, or Gaussian, scale space [7,17,27] and limit ourselves to images on $\mathbb{R}^2$, though this limitation is not essential. Given an image $\ell : \mathbb{R}^2 \rightarrow \mathbb{R}$, the multi-scale image representation $L : \mathbb{R}^2 \times \mathbb{R}^+ \rightarrow \mathbb{R}$ is obtained as a convolution with a Gaussian kernel g_σ for varying scale σ. That is, the scale space representation of ℓ is given by

$$L(x, y; \sigma) = (\ell * g_\sigma)(x, y) \,. \tag{1}$$

The linear scale space representation is mainly used for its Gaussian image derivatives and especially the so-called N-jet [6], which we denote by $J_\sigma^N[\ell]$. The latter is the collection of all Gaussian image derivatives up to order N at a particular scale σ [8,7,27]. N-jets are basic features that are often employed in supervised image analysis techniques to capture the local image structure of interest (see, for instance, [9,19,11]). Also in our experiments, we will use N-jets. The basic theory we present, however, can be used in combination with other features and multi-scale image representations as well as long as scale can properly be dealt with.

2.2 Supervised Pixel Classification

In the test phase, the trained classifier is applied to a new and previously unseen image ℓ_j from which the same feature vectors are extracted. In this way, for every location in ℓ_j, an estimate $\hat{c}_j(x, y)$ of the true class label at (x, y) is obtained by $C[F_j(x, y)]$. Most classifiers can also output an estimated posterior probability $P(c_j(x, y) = k \,|\, F_j(x, y))$ of the true class label $c_j(x, y)$ being equal to k given the feature observed feature vector $F_j(x, y))$ [5] for which we note that

$$C[F_j(x, y)] = \underset{k \in \{1,\ldots,K\}}{\operatorname{argmax}} \; P(c_j(x, y) = k \,|\, F_j(x, y)) \,. \tag{2}$$

The posteriors can be viewed as a confidence measure of the classification result and the larger the posterior is, the more confident the classifier is. In this work, we are going to extend the basic pixel-based classification scheme to incorporate scale-invariance by exploiting these posteriors, interpreting them as the output of a complex filter procedure, and apply Lindeberg's idea of maxima selection to it.

(2.3 ... and Detection)

This work does not explicitly deal with the detection task. We do however want to point out that detection can be formulated in terms of classification (see for example [12,23,26,28]). In our setting, this would, for instance, mean that the desired corresponding outputs that should be provided for the training phase are not necessarily accurate expert segmentations. Instead, for supervised detection it may suffice to label one or a few locations within the structure to be detected with one class label, say *object*, while all other locations are labeled with the label *background*. Strong local maxima among the posterior probabilities $P(c_j(x, y) = \mathit{object} \,|\, F_j(x, y))$ would correspond to a detection of a structure from the *object* class.

3 Supervised Scale-Invariant Segmentation

Our method builds further on standard pixel-based segmentation but is extended so as to take into account scale variations. The idea is to build a classifier that can be applied to all image locations at all feature scales, i.e., instead of considering classification results $C[F_j(x, y)]$, we initially consider its extension to $C[F_j(x, y, \sigma)]$, which provides labels, or for our purpose posteriors $P(c_j(x, y, \sigma) = k \,|\, F_j(x, y, \sigma))$, for the complete scale space of an image ℓ_j.

Ultimately, we are interested in a single overall segmentation and not a segmentation for every scale. Here is where the scale selection comes in. For a particular image location (x, y) in ℓ_j we check over scale which class label receives the highest posterior and assign that label to that location (cf. [20]):

$$\hat{c}_j(x, y) = \operatorname*{argmax}_{k \in \{1,\ldots,K\}} \max_{\sigma \in \mathbb{R}^+} P(c_j(x, y, \sigma) = k \,|\, F_j(x, y, \sigma)) \,. \tag{3}$$

This approach also solves the scale selection problem in a supervised way. It draws the analogy with Lindeberg's scheme and (implicitly) selects the scale at which the classifier is most confident of its decision, i.e., where classes can be best separated from each other.

The way this classification approach takes into account scale may already be interesting in itself, but we aimed for the segmentation approach to be scale-invariant.

3.1 Additional Remarks

Scale-invariance in the current context means that if we rescale an image ℓ_j with a factor $a > 0$ to an image $\ell_{j'} := \ell_j \circ S_a$ that the corresponding classification result scales in the same way, i.e., $\hat{c}_{j'}(x, y) = \hat{c}_j \circ S_a$. Now this is achieved by relying on scale-invariant features. That is, we generally require that $F_j(x, y, \sigma)$ in the original image ℓ_j equals $F_{j'}(ax, ay, a\sigma)$ in the scaled image $\ell_{j'}$. With this choice of features, corresponding feature vectors are mapped to the same location in feature space and therefore classified in the same way, which results in the desired scale-invariance. In the case of Gaussian derivative features from an N-jet at scale σ, this means for example that every nth order derivative should be normalized by σ^n.

It is copacetic that there are no restrictions on the classification scheme to use. With the choice of scale-invariant features, any choice of classifier results in a scale-invariant segmentation approach and this allows us to employ the full arsenal of machine learning and pattern recognition techniques [4,2].

4 Illustrative Experiments

Our contribution is primarily of a conceptual nature with no need for extensive experimental validation. Nonetheless, we provide some basic, yet nontrivial, illustrations of our scale-invariant segmentation approach as defined through Equation (3). We applied the method to two different tasks. The first one is the segmentation of two simple geometric shapes from the background. The second one comprises a texture segmentation task.

4.1 Classifiers and Features

Before we can apply our segmentation scheme, we need to choose features to describe for every location the relevant local image structure. In basically all of

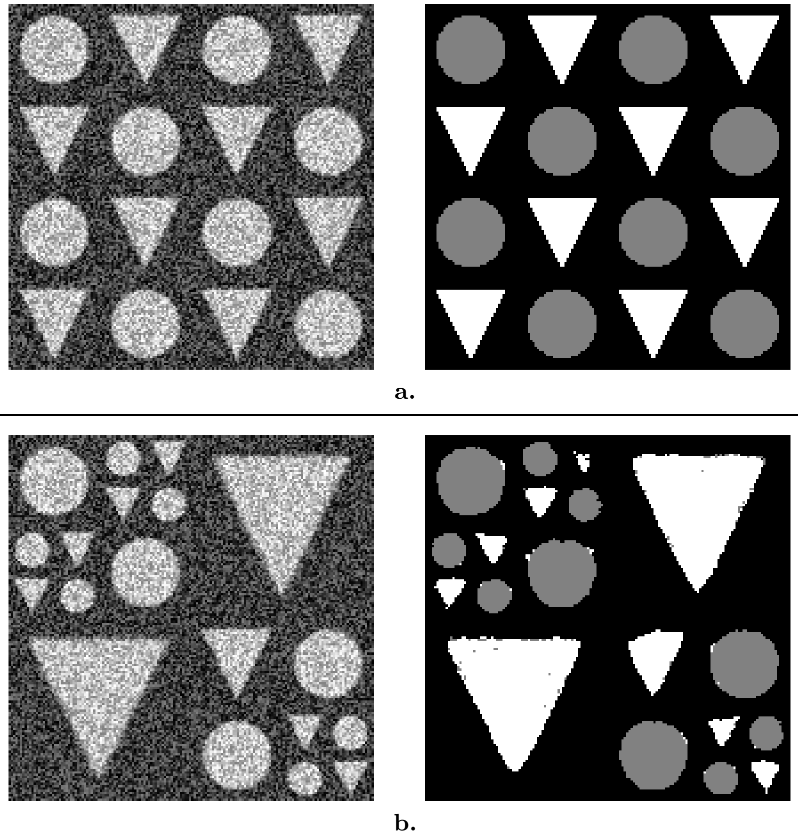


Fig. 1. a. Triangular and circular shapes and their corresponding segmentation used in the training phase. The segmentation is formulated as a three-class pixel classification problem. **b.** An example test image with triangles and circles of different size and the corresponding classification result.

the experiments, we choose the scale-normalized 6-jet, which results in a total of $D = 28$ features for every location. The normalization is depends on the order n of the derivative; every derivative is scaled by σ^n, which makes the features scale-invariant as required.

We also limit ourselves to a relatively straightforward classification technique, namely classical quadratic discriminant analysis (QDA) [4,2]. This classifier makes multivariate normality assumptions about every individual class and based on that constructs a classifier. For every class a feature mean and a class-conditional covariance matrix should be estimated from the training data. This quadratic model is accurate enough for our illustratory purposes and more advanced techniques such as support vector machines, nearest neighbor methods, and boosting approaches provide little extras in the current setting.

4.2 Shapes

Figure 1.a shows on the left examples of noisy triangles and circles, eight each, that should be segmented from the equally noisy background. The right displays the ground truth pixel labeling, which is used as training output. Obviously, a simple blob detector would probably be able to pick out the 16 objects from the input image. It would however be more challenging to design detectors that are more specific and respond merely to one of the two geometric structures. Our scheme therefore also tries to discriminate between the two different shapes and should respond differently to them, i.e., by giving different label outputs. Consequently, we model this problem as a three-class classification problem. The gray-scale in the righthand image of Figure 1.a is of no significance and only indicates that there are indeed three different classes in the image and which pixels belong to which class.

The procedure is tested on the image on the left of Figure 1.b. It also contains scaled versions of the triangular and circular shapes in order to test the scale-invariance of our approach. After extraction the 6-jets from the training data from a range of scales, a QDA is trained and applied to the test image. The resulting segmentation can be found in Figure 1.b on the righthand side. It shows that our procedure is fairly accurate and that the majority of the pixels has been labeled correctly in spite of the relatively straightforward classifier and scale space features. The most notable mistakes seem to be on the small scale triangles. Some of these segmentations are deformed and the one at the top even has been missed almost entirely. The main reasons for these glitches is that the images used are discrete and the 'size' of the added noise does not scale with the shape scale. As a result, scale-invariance will only hold approximately and over a restricted range of scales, which is reflected in somewhat deteriorated performance on the small scale structures.

4.3 Textures

Experiments similar to those in the previous subsection have been performed on two times two Brodatz textures [3]. The two pairs of textures can be found in Figures 2 and 4. For both pairs, we use the two images as the training set and assume them to be from different classes. The corresponding test images, which include both textures from the training set, are displayed in Figures 3.a and 5.a, respectively. All four images are constructed from scaled versions of the original training textures, one of which is in a circular area in the center while the other fills the remainder of the image. The aim is to segment the one texture from the other.

All texture intensities have been normalized to mean zero and unit standard deviation. As a result, a generic blob detector is unable to localize the texturized blobs in the middle of the test images. We really need to employ more rich features that are capable of capturing the relevant higher-order structure and combine these in order to perform the detection or segmentation successfully. This is what QDA, the classifier, does. Figures 3.b and 5.b give the segmentations

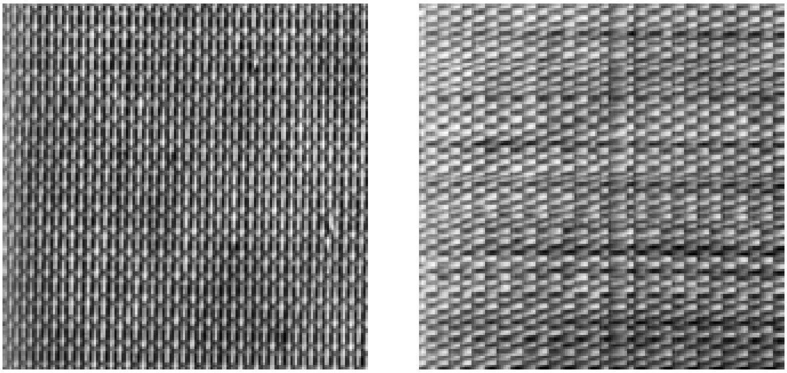


Fig. 2. Two training images taken from the Brodatz collection of textures [3]. On the left is D53, the right shows D55.

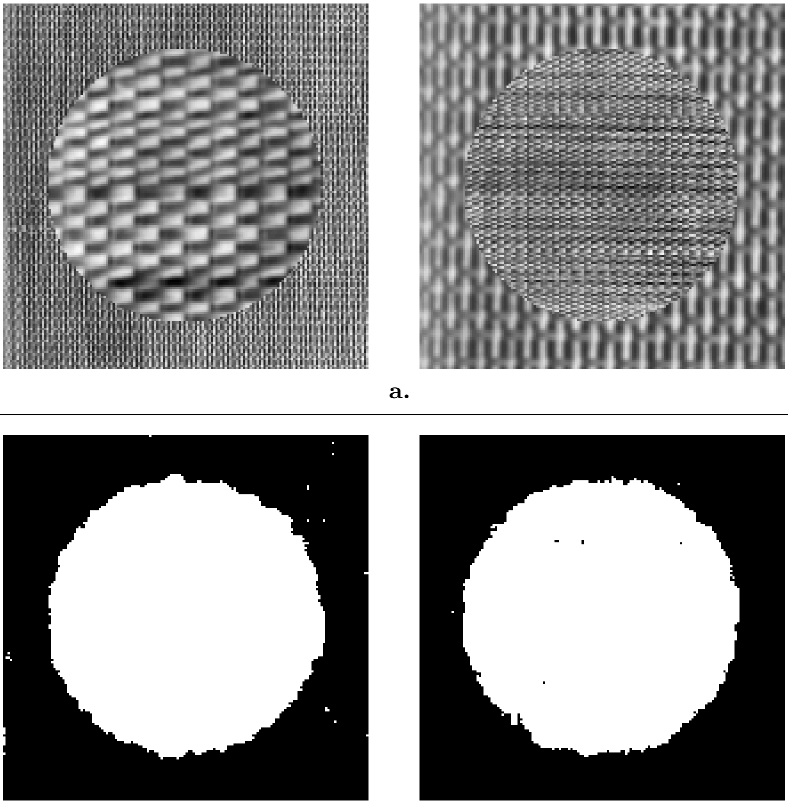


Fig. 3. a. Two example test images in which the texture scales are varied and set differently from those in the training set in Figure 2. Both images contain both textures. **b.** Segmentation results obtained with 6-jets in combination with QDA.

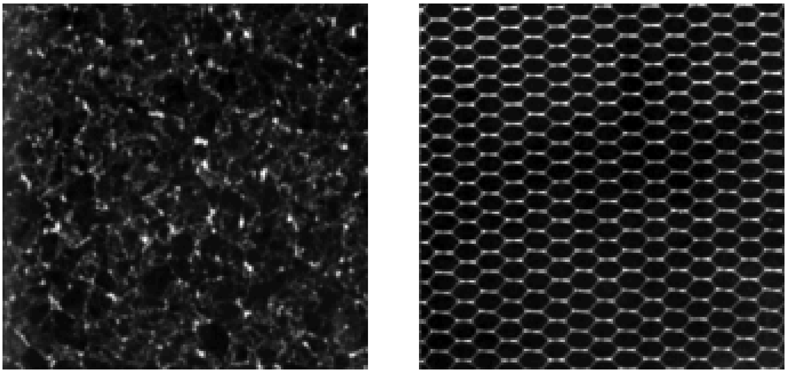
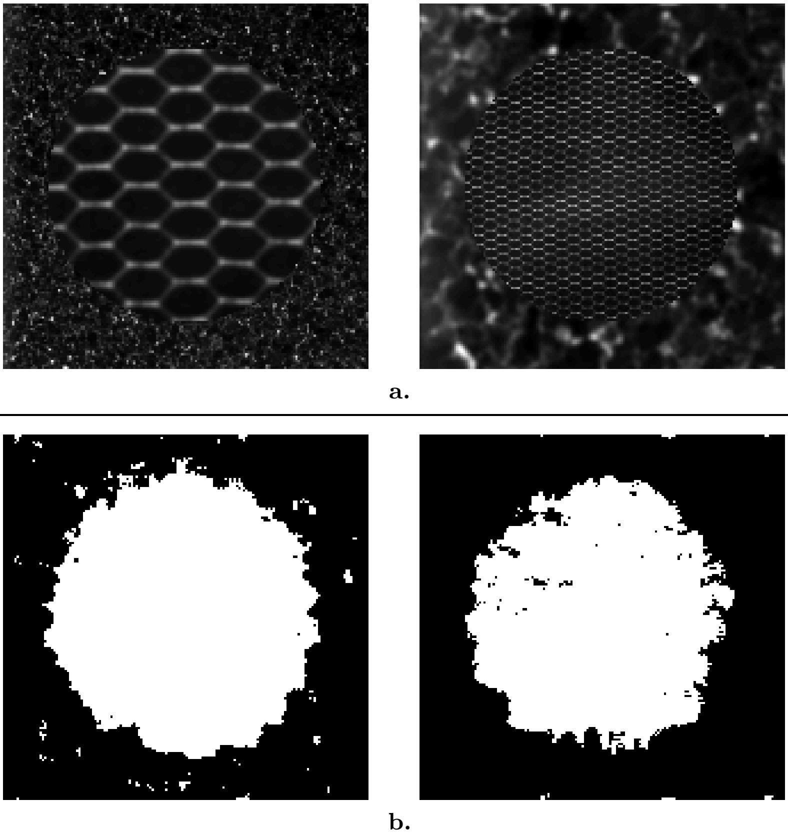


Fig. 4. Compare to Figure 2. Two training images taken from the Brodatz textures collection [3]. On the left is D33, the right displays D34.



a.

b.

Fig. 5. Compare to Figure 3. **a.** Two example test images in which the texture scales are varied and set differently from those in the training set in Figure 4. Both images contain both training textures. **b.** Segmentation results obtained with 6-jets in combination with QDA.

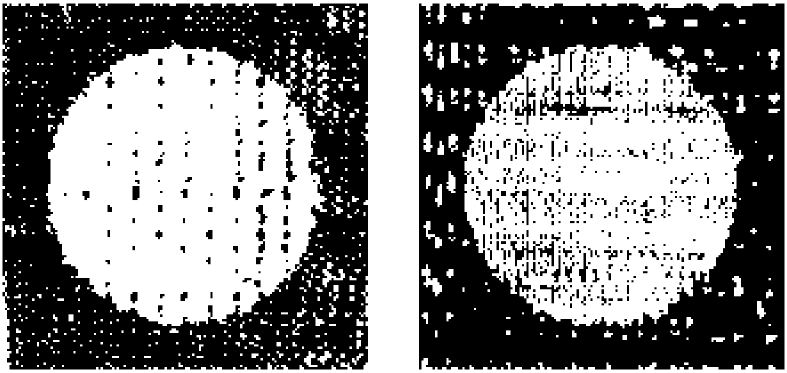


Fig. 6. Segmentation results obtained with 2-jets in combination with QDA. Compare to Figure 3.b.

for the test images in Figures 3.a and 5.a, respectively. As for the results in the previous subsection, similar comments can be made about the reasons for misclassification in these experiments. In the case of these textures, however, there may be two additional reasons at play. First of all, textures are generally more difficult to segment than a shape consisting of a homogenous intensities even though the latter may be noisy. Secondly, the training set does not contain any examples of the two textures bordering, which causes unreliable classification results at such boundaries in the test images. It is indeed at these locations where the segmentation seems most inaccurate.

The first texture segmentation task is probably simpler than the second one. There is a strong difference in orientation between the two textures, which basically sets them apart and one might suspect that a descriptor based on simple second-order, or even first-order, derivatives should be able to capture this difference. Figure 6 shows what happens to the segmentations corresponding to the test images in Figure 3.a if we replace the 6-jets with 2-jets in our procedure. Indeed, to quite a large extent the segmentation is still successful, but the results cannot match the accuracy from those in Figure 3.b, which shows the importance of including higher-order derivatives. Results using the 1-jet are worse even.

5 Discussion and Conclusion

A scale-invariant supervised approach to image segmentation has been presented that draws inspiration from Lindeberg's classical scale selection approach. There are two major advantages compared to other supervised scale-invariant segmentation techniques. Firstly, we are not necessarily committed to specific features that have been designed to achieve invariance in a rather intricate way, as for example in [15]. Our scheme allows the inclusion of any scale-invariant feature set, allowing for very problem specific choices. More important might be the second point, which is the fact that we stay close to the general pixel classification framework and can exploit the full arsenal of powerful pattern recognition and machine learning techniques. In our experiments, we only scratched the surface of possible techniques. They nonetheless show the potential of the approach.

Possibly the most restricting feature of our method is that it is supervised, so we do need training data in order for our approach to work. As a general rule, we may expect to need more complex features, more complex classifiers, and a larger number of examples, with an increasingly complex segmentation problem that we want to tackle. The interplay of these aspects of learning are at the core of general pattern recognition and machine learning research. It is however interesting to study these aspects within the more confined context of image segmentation and detection as this may lead to stronger, more generally applicable guidelines to come to the selection of the right classifier, the right features, etc.

One specific topic for further research we want to mention here concerns Equation (3) and in particular the maximum operator over all scales, which basically picks out a single scale and allows for a close link with Lindeberg's scale-selection scheme. The question remains however if we can do better. An answer might be found in the analysis of the deep structure of the probabilistic posterior scale space (cf. [7,17,27]).

References

1. Bay, H., Tuytelaars, T., Van Gool, L.: SURF: Speeded up robust features. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3951, pp. 404–417. Springer, Heidelberg (2006)
2. Bishop, C.: Pattern recognition and machine learning. Springer, Heidelberg (2006)
3. Brodatz, P.: Textures: A Photographic Album for Artists & Designers. Dover, New York (1966)
4. Duda, R., Hart, P., Stork, D.: Pattern classification, vol. 2. Wiley, Chichester (2001)
5. Duin, R., Tax, D.: Classifier conditional posterior probabilities. In: Advances in Pattern Recognition, pp. 611–619 (1998)
6. Florack, L., Ter Haar Romeny, B., Viergever, M., Koenderink, J.: The Gaussian scale-space paradigm and the multiscale local jet. International Journal of Computer Vision 18(1), 61–75 (1996)
7. Florack, L.: Image Structure. Kluwer Academic Publishers, Dordrecht (1997)
8. Florack, L., ter Haar Romeny, B., Koenderink, J., Viergever, M.: Scale and the differential structure of images. Image and Vision Computing 10(6), 376–388 (1992)
9. Folkesson, J., et al.: Segmenting articular cartilage automatically using a voxel classification approach. IEEE Trans. on Medical Imaging 26(1), 106–115 (2007)
10. Frangi, A.F., Niessen, W.J., Vincken, K.L., Viergever, M.A.: Multiscale vessel enhancement filtering. In: Wells, W.M., Colchester, A.C.F., Delp, S.L. (eds.) MICCAI 1998. LNCS, vol. 1496, pp. 130–137. Springer, Heidelberg (1998)
11. van Ginneken, B., Stegmann, M., Loog, M.: Segmentation of anatomical structures in chest radiographs using supervised methods: a comparative study on a public database. Medical Image Analysis 10(1), 19–40 (2006)
12. Hjelmås, E., Low, B.: Face detection: A survey. Computer Vision and Image Understanding 83(3), 236–274 (2001)

13. Janssen, J., et al.: Scale-invariant segmentation of dynamic contrast-enhanced perfusion MR images with inherent scale selection. J. Visualization and Computer Animation 13(1), 1–19 (2002)
14. Kang, Y., Morooka, K., Nagahashi, H.: Scale invariant texture analysis using multiscale local autocorrelation features. In: Kimmel, R., Sochen, N.A., Weickert, J. (eds.) Scale-Space 2005. LNCS, vol. 3459, pp. 363–373. Springer, Heidelberg (2005)
15. Kokkinos, I., Yuille, A.: Scale invariance without scale selection. In: IEEE Conf. on Computer Vision and Pattern Recognition, pp. 1–8. IEEE, Los Alamitos (2008)
16. Leung, M., Peterson, A.: Scale and rotation invariant texture classification. In: The 26th Asilomar Conference on Signals, Systems and Computers, pp. 461–465 (1992)
17. Lindeberg, T.: Scale-Space Theory in Computer Vision. Kluwer Academic, Dordrecht (1994)
18. Lindeberg, T.: Feature detection with automatic scale selection. Int. J. of Computer Vision 30(2), 79–116 (1998)
19. Loog, M., Ginneken, B.: Segmentation of the posterior ribs in chest radiographs using iterated contextual pixel classification. IEEE Trans. on Medical Imaging 25(5), 602–611 (2006)
20. Loog, M., Li, Y., Tax, D.M.J.: Maximum Membership Scale Selection. In: Benediktsson, J.A., Kittler, J., Roli, F. (eds.) MCS 2009. LNCS, vol. 5519, pp. 468–477. Springer, Heidelberg (2009)
21. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vision 60, 91–110 (2004)
22. Mikolajczyk, K., Schmid, C.: Scale & affine invariant interest point detectors. Int. J. Computer Vision 60(1), 63–86 (2004)
23. Papageorgiou, C., Oren, M., Poggio, T.: A general framework for object detection. In: Sixth International Conference on Computer Vision, pp. 555–562. IEEE, Los Alamitos (2002)
24. Platel, B., Kanters, F., Florack, L., Balmachnova, E.: Using multiscale top points in image matching. In: International Conference on Image Processing, ICIP 2004, vol. 1, pp. 389–392. IEEE, Los Alamitos (2005)
25. Pun, C., Lee, M.: Log-polar wavelet energy signatures for rotation and scale invariant texture classification. IEEE Trans. PAMI, 590–603 (2003)
26. Shotton, J., Blake, A., Cipolla, R.: Contour-based learning for object detection. In: Tenth IEEE International Conference on Computer Vision, ICCV 2005, vol. 1, pp. 503–510. IEEE, Los Alamitos (2005)
27. Ter Haar Romeny, B.: Front-End Vision and Multi-Scale Image Analysis. Kluwer Academic, Dordrecht (2002)
28. Yang, M., Kriegman, D., Ahuja, N.: Detecting faces in images: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence 24(1), 34–58 (2002)