Support Blob Machines
The Sparsification of Linear Scale Space

Marco Loog

Image Group, IT University of Copenhagen
Copenhagen, Denmark
marco@itu.dk
Image Sciences Institute, University Medical Center Utrecht
Utrecht, The Netherlands
marco@isi.uu.nl

Abstract. A novel generalization of linear scale space is presented. The generalization allows for a sparse approximation of the function at a certain scale.
To start with, we first consider the Tikhonov regularization viewpoint on scale space theory [15]. The sparsification is then obtained using ideas from support vector machines [22] and based on the link between sparse approximation and support vector regression as described in [4] and [19].
In regularization theory, an ill-posed problem is solved by searching for a solution having a certain differentiability while in some precise sense the final solution is close to the initial signal. To obtain scale space, a quadratic loss function is used to measure the closeness of the initial function to its scale σ image.
We propose to alter this loss function thus obtaining our generalization of linear scale space. Comparable to the linear ϵ-insensitive loss function introduced in support vector regression [22], we use a quadratic ϵ-insensitive loss function instead of the original quadratic measure. The ϵ-insensitivity loss allows errors in the approximating function without actual increase in loss. It penalizes errors only when they become larger than the a priory specified constant ϵ. The quadratic form is mainly maintained for consistency with linear scale space.
Although the main concern of the article is the theoretical connection between the foregoing theories, the proposed approach is tested and exemplified in a small experiment on a single image.

1 Introduction

There are many extensions, variations, perturbations, and generalizations of linear scale space [10,24]. E.g. anisotropic, curvature, morphological, α, pseudo, Poisson, i, torsion, geometry driven and edge preserving scale spaces [1,2,3,7,9, 13,14,16,17,23]. All these approaches serve one or more purposes in image processing, general signal processing or computer vision, as a whole covering a large area of applications. This paper presents another interesting possibility for generalization of linear scale space that has not been explored up to now. I.e., those

T. Pajdla and J. Matas (Eds.): ECCV 2004, LNCS 3024, pp. 14–24, 2004.

that allow for a sparse scale space representation of the function under consideration. Being sparse means that the approximation can be obtained using only a small collection of building block or basis functions. This sparseness is being enforced by setting certain constraints to the solution.

The proposed technique exploits links between scale space and support vector machines. The sparsified representation thus obtained is called the support blob machine (SBM), as blobs are the main building blocks used in the representation. The principal observation that leads to the notion of SBMs is that both support vector regression (SVR) [22,19] and scale space can be related to Tikhonov's regularization theory [21] (see [4], [15], and [20]). For more on the connection between scale spaces and regularization see [18].

Regularization is a technique typically used for solving ill-posed problems in a principled way. While scale space offers a solution to the problem of defining derivatives of a multidimensional (digital) signal in a well-posed way, the goal in regression is mainly to recover a function from a finite, possibly noisy, sampling of this function, which is also clearly ill-posed. Support vector regression (SVR) not only offers a well-posed solution to the regression problem but has the additional advantage that it can be used to give a sparse solution to the problem at hand. The sparse representation is acquired by allowing the regressed function to deviate from the initial data without directly penalizing such deviations. This sparseness behavior is accommodated for through the so-called ϵ-insensitive loss. The resulting behavior is clearly different from, for example, standard linear regression in which, using a quadratic loss function, only the slightest deviation from the given data is penalized immediately.

This article focuses on the relationships mentioned above, deriving the SBMs, providing a computational scheme to determine these scale spaces, and finally exemplifying the scale spaces obtained. However, before doing so, we mention several reasons why the kind of scale spaces proposed are of interest.

First of all, the approach may lead to improved image feature detectors. SVR, like support vector classifiers, has proven to be very successful in many applications which is partly due to the relation with robust statistics [8]. But not only may certain forms of sparsified scale space lead to more robust detection of edges, blobs and other visual cues, another advantage is that one could build on the structural risk minimization framework as proposed by Vapnik [22] and consequently one may be able to theoretically underpin the practical performance of these detectors in real-world applications.

Another useful application is in the area of feature-based image analysis and the related research into metamery classes and image reconstruction [6,11,12]. E.g., in [12], certain greedy methods are discussed for the selection of points of interest in an image, i.e., based on a certain form of reconstruction energy (that can be calculated for every point individually), the representing points are chosen starting with the one with the highest energy and going down gradually. These points, together with their associated receptive field weighting function are then taken as image representation. A greedy approach to the point selection problem was considered appropriate by the authors, because features must be detected and represented individually in early vision. However, they also note that this approach is not necessarily optimal for image representation purposes as the reconstruction information mutually conveyed by two or more points is not taken into account, leading to only a suboptimal solution for the image reconstruction task. The SBM does take the global information into account and can therefore obtain an improved selection of points of interest, which all in all could prove useful for, e.g., image compression.

Finally, a little more speculative, the data reduction that is achieved by the sparse representation of the data may facilitate the use of otherwise prohibitively computer intensive techniques in computer vision or image analysis. As an example one could think of the registration of two large 3D data volumes based on their sparse representation instead of all of the voxels.

The remainder of the paper is organized as follows. Subsection 2.1 gives the regularization formulation of scale space after which Subsection 2.2 discusses a more general formulation of regularized regularization taken from the SVR literature. Section 2.3 links the aforementioned techniques. The quadratic ϵ-insensitive loss functions is then introduced in Subsection 2.4 on which our principal generalizations is based. Section 3 describes some experiments to exemplify the novel scale space. Section 4 contains the discussions and concludes the article.

2 The Sparsification of Scale Space

2.1 Scale Space Regularization Formulation

This subsection recapitulates one of the principle contributions of [15] in which scale space is related to a specific instance of Tikhonov regularization. The authors consider the general regularization formulation given by Tikhonov [21]: The regularized function f associated to the function g on $\mathbb{R}^n$ minimizes the functional $\mathcal{E}$ defined as

$$\mathcal{E}[h] := \tfrac{1}{2} \int (h(x) - g(x))^2 + \sum_{j=1}^{\infty} \sum_{|J|=j} \lambda_J \left(\frac{\partial^{|J|} h(x)}{\partial x^J} \right)^2 dx \,. \tag{1}$$

All λ_J are nonnegative and J is an n-index used for denoting derivatives of order $|J|$. The first term of the right hand side penalizes any deviation of the function h from the given function (the data) g. The second part does not involve g and is the regularization term on h, which, in a certain way controls the smoothness.

It can be shown in this setting, that the solution to the problem, f, can be obtained by a linear convolution of g. Moreover, the authors prove that f equals $g * G_t$, where the latter is the Gaussian kernel

$$G_t(x) := \frac{1}{(2\sqrt{\pi t})^n} e^{-\frac{\|x\|^2}{4t}} \,, \tag{2}$$

if and only if

$$\lambda_J = \frac{t^{|J|}}{|J|!} = \frac{t^j}{j!} \tag{3}$$

for all J, thus relating scale space at scale $\sqrt{2t}$ to a specific form of Tikhonov regularization.

2.2 Regression and Regularization

The regression problem is generally formulated in terms of a discrete data set of ℓ (noisy) samples $\hat{g}(x_i)$ from the function g and the goal is to recover this underlying function g merely based on these ℓ samples [4,19,22]. To this end, an underlying functional form of g based on a set of linearly independent basis functions φ_i is assumed, i.e., g can be represented as

$$g(x) := \sum_{i=1}^{\infty} c_i \varphi_i(x) + c_0 \,. \tag{4}$$

The constant term c_0 and the parameters c_i have to be estimated from the data. This is clearly an ill-posed task since the problem as such has an infinite number of solutions. Again, regularization can be used to turn it into a well-posed problem by imposing smoothness constraints on the final solution. The regularized solution f minimizes the functional $\mathcal{R}$

$$\mathcal{R}[h] := C \sum_{i=1}^{\ell} L(\hat{g}(x_i) - h(x_i)) + \tfrac{1}{2}\Lambda[h] \,, \tag{5}$$

where L is the loss function penalizing deviation of f from the measurement data $\hat{g}$, Λ is a general constraint that enforces smoothness of the optimal solution f, and C is a positive constant that controls the tradeoff between the two previous data terms.

An important result is the following (see [4]). If the functional Λ has the form $\Lambda[h] = \sum_{i=1}^{\infty} \frac{c_i^2}{\lambda_i}$, where all λ_i are positive and $\{\lambda_i\}_{i=1}^{\infty}$ is a decreasing sequence, then the solution f to the regularization problem (5) takes on the form

$$f(x) = \sum_{i=1}^{\ell} a_i K(x, x_i) + c_0 \,, \tag{6}$$

with the kernel function K being defined as

$$K(x, x_i) = \sum_{i=1}^{\infty} \lambda_i \varphi_i(x) \varphi_i(x_i) \,. \tag{7}$$

A large class of regularizations can be defined based on the foregoing class of smoothing functionals and in the remainder of the article, only these kind of smoothness constraints are considered.

Note that if the cost function L is quadratic, the unknown parameters in Equation (6) can be determined by solving a linear system comparable to standard regression. When the cost function is not quadratic, the a_i can not be readily obtained by solving a linear system and one has to resort to different optimization methods. For particular loss functions, as the main one considered in this article, $\mathcal{R}[h]$ can be optimized using quadratic programming (QP), allowing the optimization to be done in a fairly straightforward way (see Subsection 2.4).

2.3 Kernel Formulation of Scale Space

Before formulating the sparsified version of scale space, first the regularizations in (1) and (5) are related to each other. That is, in the functional $\mathcal{R}$, the 'free parameters' are chosen such that a solution f that minimizes this functional is equal to the minimizing solution of $\mathcal{E}$ with the λ_J as given in Equation (3). In this case

$$\mathcal{E}[h] = \tfrac{1}{2} \int (h(x) - g(x))^2 + \sum_{j=1}^{\infty} \frac{t^j}{j!} \sum_{|J|=j} \left(\frac{\partial^{|J|} h(x)}{\partial x^J} \right)^2 dx \,. \tag{8}$$

The fact that one formulation is continuous and the other uses discrete observations is disregarded. Doing so, it is of course immediately clear that in (5), the loss function L has to be the quadratic loss, i.e.,

$$L(\cdot) = (\cdot)^2 \,. \tag{9}$$

So now our main concern is the smoothing term in both functionals.

Starting with the smoothing term from (1)—with the λ_J as defined in Equation (3), based on induction and partial integration, and in addition properly rearranging terms, it can be shown that the following equivalence holds (cf. [15])

$$\sum_{|J|=j} \left(\frac{\partial^{|J|} h(x)}{\partial x^J} \right)^2 = \begin{cases} \left(\nabla \Delta^{\frac{j-1}{2}} h(x) \right)^2 & \text{if } j \text{ is odd} \\ \left(\Delta^{\frac{j}{2}} h(x) \right)^2 & \text{if } j \text{ is even} \end{cases} \,. \tag{10}$$

Where Δ is the Laplacean and ∇ is the gradient operator.

Setting $2t = \sigma^2$ in Equation (8) and substituting the results from Equation (10), the expression obtained can be related to the result discussed in [20,19] (cf. [5]) in which Gaussian functions are shown to be the kernel functions associated to this specific form of regularization. That is, K in (7) should be defined as

$$K(x, x_i) := e^{\frac{-\|x-x_i\|^2}{2\sigma^2}} = e^{\frac{-\|x-x_i\|^2}{4t}} \,, \tag{11}$$

for the regularized regression (5) to be equivalent to the Tikhonov regularization resulting in linear scale space of g.

Finally, with the constant C in (5) set to $\frac{1}{2}$, the regularization functionals $\mathcal{E}$ and $\mathcal{R}$ become completely equivalent.

2.4 Quadratic ϵ-Insensitive Loss and SBMs

Based on the foregoing equivalence, this subsection introduces to the generalization of linear scale space within the SVR framework via a quadratic ϵ-insensitive loss: The support blob machines (SBMs). The main idea behind using this quadratic ϵ-insensitive loss is that the generalization should possess a similar kind of ability to obtain sparse representations as the (linear) ϵ-insensitive loss function exhibits.

It is formulated for pixel-based discrete images and related to SVR based on the (linear) ϵ-insensitive loss. Subsequently, it is demonstrated how to obtain the minimizing solution under this loss function based on a quadratic programming (QP) formulation. This is similar to the optimization procedure used in standard ϵ-insensitive loss support vector machines [4,22,19].

The loss function originally proposed in the context of SVR [22] is the ϵ-insensitive loss $|\cdot|_\epsilon$. It allows for minimization of (5) using QP and is defined as

$$|x|_\epsilon := \begin{cases} 0 & \text{if } |x| \leq \epsilon \\ |x| - \epsilon & \text{otherwise} \end{cases} . \tag{12}$$

This loss function bears a resemblance to some loss functions used in statistics which provide robustness against outliers [8]. In addition to this important property, the loss has another distinctive feature: It assigns zero cost to deviations of h from $\hat{g}$ that are smaller than ϵ and therefore, every function h that comes closer than ϵ to the ℓ data points $\hat{g}(x_i)$ is considered to be a perfect approximation.

The similar quadratic ϵ-insensitive loss function, more closely related to the well-known quadratic loss in Equation (9), can be defined as follows

$$(x)^2_\epsilon := \begin{cases} 0 & \text{if } |x| \leq \epsilon \\ (|x| - \epsilon)^2 & \text{otherwise} \end{cases} . \tag{13}$$

Using this loss, deviations form the underlying data are essentially quadratically penalized. However, the ϵ allows zero cost deviations from the data points $\hat{g}(x_i)$, which, for the minimizing solution f, leads to several a_i being zero in Equation (6). The number of a_i being equal to zero is dependent on the parameter ϵ. The larger ϵ is, the more a_i are zero. If $\epsilon = 0$, then $a_i \propto \hat{g}(x_i)$ (note that the Gaussian kernel in (11) is not normalized) and so in general a_i will be nonzero. Taking ϵ larger than zero, a sparse solution to the problem can be obtained (see [4] and [22] for the actual underlying mechanisms leading to sparseness).

Taking all of the foregoing into consideration, the regularization functional $\mathcal{E}_\epsilon$ for sparse scale space is now readily defined in its discrete form as

$$\mathcal{E}_\epsilon[h] := C \sum_{i=1}^{\ell} (h(x_i) - g(x_i))^2_\epsilon + \sum_{i=1}^{\ell} \sum_{j=1}^{\infty} \frac{t^j}{j!} \sum_{|J|=j} \left(\frac{\partial^{|J|} h(x_i)}{\partial x^J} \right)^2 , \tag{14}$$

with, in general, C equal to $\frac{1}{2}$. Taking ϵ equal to zero, results in ordinary linear scale space.

Exploiting the link with SVR, a dual QP formulation that solves $\mathrm{argmin}_h \mathcal{E}[h]_\epsilon$ can be stated (cf. [4,19]):

$$\begin{aligned} \underset{\alpha_i^-, \alpha_i^+}{\mathrm{argmin}} \ \sum_{i=1}^{\ell} \Bigg[& \epsilon(\alpha_i^+ + \alpha_i^-) - \hat{g}(x_i)(\alpha_i^+ - \alpha_i^-) \\ & + \tfrac{1}{2} \sum_{j=1}^{\ell} (\alpha_i^+ - \alpha_i^-)(\alpha_j^+ - \alpha_j^-)(K(x_i, x_j) + \frac{1}{C} + \delta_{ij}) \Bigg] \end{aligned} \tag{15}$$

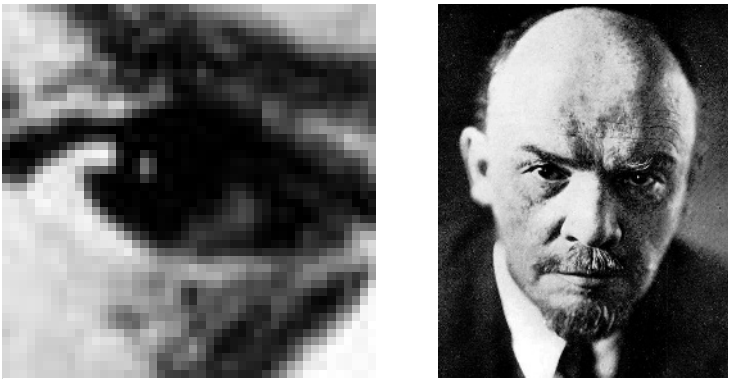

Fig. 1. On the right, the original image of Lenin. On the left is his right eye, the 50 by 50 sub-image used to exemplify the SBMs.

subject to $\sum_{j=1}^{\ell}(\alpha_i^+ - \alpha_i^-) = 0$ and $\alpha_i^-, \alpha_i^+ \geq 0$. After which the $\alpha_i^{-\star}$ and $\alpha_i^{+\star}$ that optimize (15) determine the optimal solution to the minimization of the functional $\mathcal{E}_\epsilon$, i.e., (cf. Equation (6))

$$f(x) = \sum_{i=1}^{\ell}(\alpha_i^{+\star} - \alpha_i^{-\star})K(x, x_i) + c_0^\star. \tag{16}$$

The optimal offset $c_0^\star$ can be determined by exploiting the Karush-Kuhn-Tucker condition, after the solution to the foregoing problem has been obtained. The condition basically states that at the optimal solution of QP (15), the product of the dual variables and their constraints should vanish (see [19,22]).

3 An Illustrative Experiment

The SBMs are exemplified on a single, small gray value image. The image is a 50 by 50 sub-image taken from a larger image of Lenin (see Figure 1). The sub-image is Lenin's right eye. The gray values of this image are scaled between 0 and 255 for this experiment. Note that it is important to know the range the gray values are in, because the function minimizing $\mathcal{E}_\epsilon$ with $\epsilon > 0$ is not invariant under (linear) intensity scalings, which is due to the ϵ-insensitivity.

For this image, for several settings of the parameters σ $(= \sqrt{2t})$ and ϵ, the support vectors are determined using the functional (14). Simultaneously, the optimal values for the parameters α_i^+ and α_i^- are obtained. These values, together with the Gaussian kernels, define the regularized image via Equation (16).

Figure 2 plots the values $\alpha_i^{+\star} - \alpha_i^{-\star}$ in the position of the blob it supports in the SBM for varying σ and ϵ. Figure 3 gives the regularized images, which are actually blurred versions (as is clear from Equation (16)) of the images in Figure 2. Put differently, the images in Figure 2 can be considered de-blurred images.

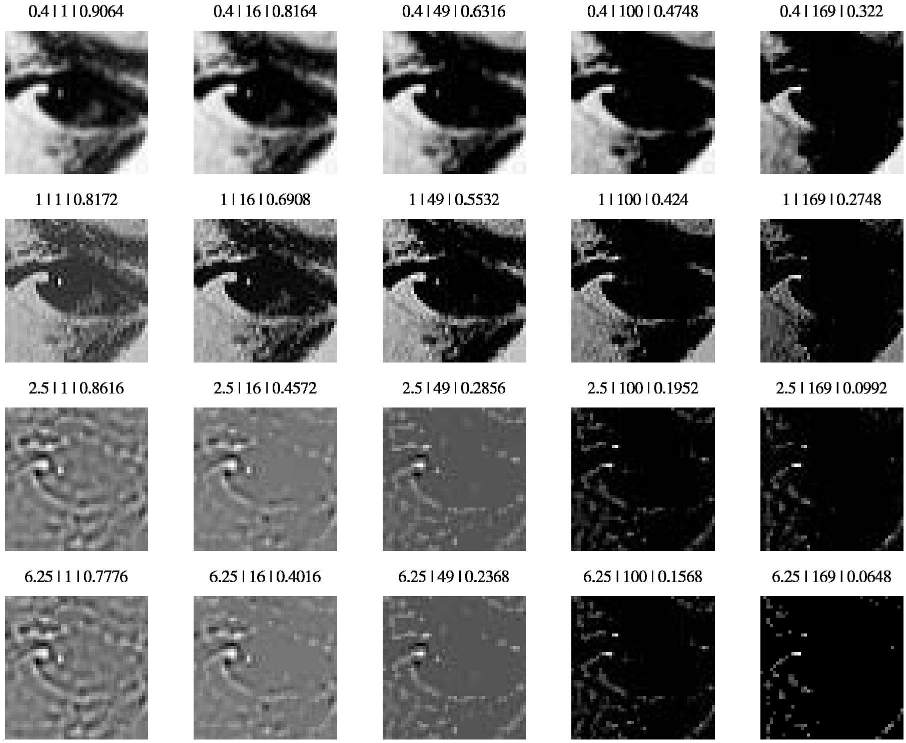

Fig. 2. Plots of the values $\alpha_i^{+\star} - \alpha_i^{-\star}$ given by the SBMs. On the top of every image the scale, the value of ϵ, and the relative number of support vectors is given.

Because it is not immediately clear from Figure 2 when there is actually a support vector present in a certain position, i.e., when $\alpha_i^{+\star} - \alpha_i^{-\star}$ is not equal to zero, Figure 4 indicates in black the positions that contain a support vector.

The additional text added at the top of every images in Figures 2 to 4 gives information on the scale σ, the ϵ, and the relative amount of support vectors (as $\cdot|\cdot|\cdot$). This last number is simply calculated by dividing the number of support vectors by $2500 = 50^2$.

4 Discussion and Conclusion

We introduced support blob machines (SBMs) based on a link between scale space theory and support vector regression, which are connected to each other via regularization theory. The SBMs give a sparsification of linear scale space by employing a quadratic ϵ-insensitive loss function in its regularization functional. Through the sparseness obtained, the regularized function can be represented using only a small collection of building block or basis functions.

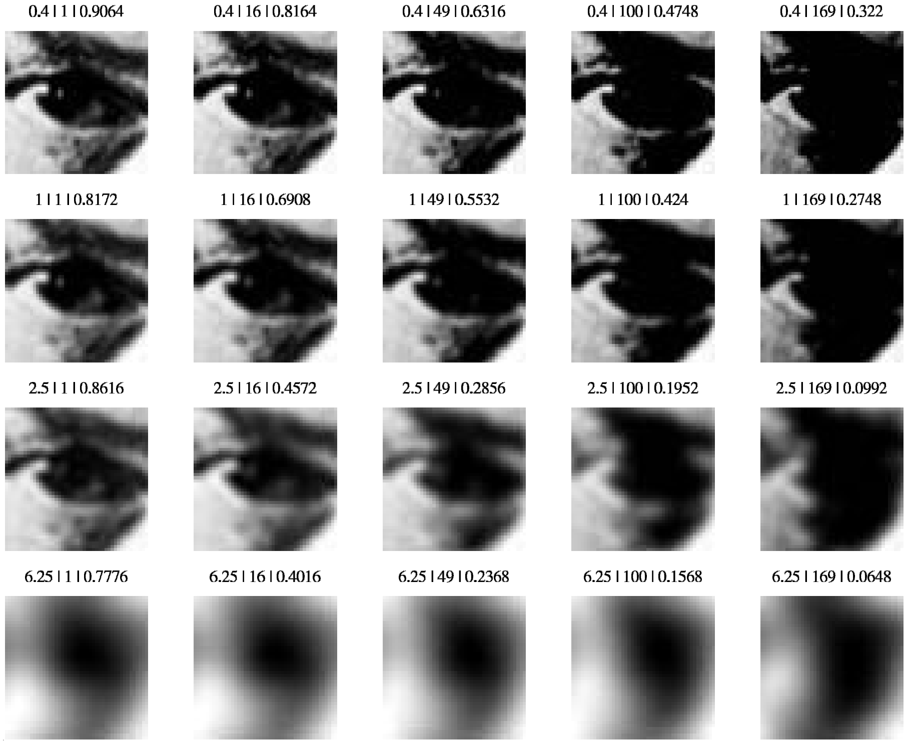

Fig. 3. The regularized images associated to a certain scale σ and value of ϵ. (The values are given on the top of every image. The relative number of support vectors is given here also as the last number.)

Some SBM instances were exemplified on a small 50 by 50 image, and it was shown that the technique indeed obtains sparse representations of images at a certain scale. However, in our tests, the reduction of information that was attained is certainly not overwhelming and further research should be conducted before a definite conclusion about the performance of SBMs can be stated. A simple suggestion, which could lead to improved sparseness performance is to increase the parameter C in the functional. In the tests this was set to $\frac{1}{2}$ to keep a close link with standard scale space. A larger value for C leads automatically to a sparser representation of the underlying signal.

The principal contribution of this article is, however, the formal link between two interesting techniques: scale space and support vector machines. This link could now be further exploited and more advanced regularization approaches may be considered.

Our future research will focus on developing a formulation that gives a sparse representation while taking all scales into account simultaneously and not merely one scale at a time. This may, in combination with different types of loss functions, lead to a robust form of scale selection in combination with blob detection

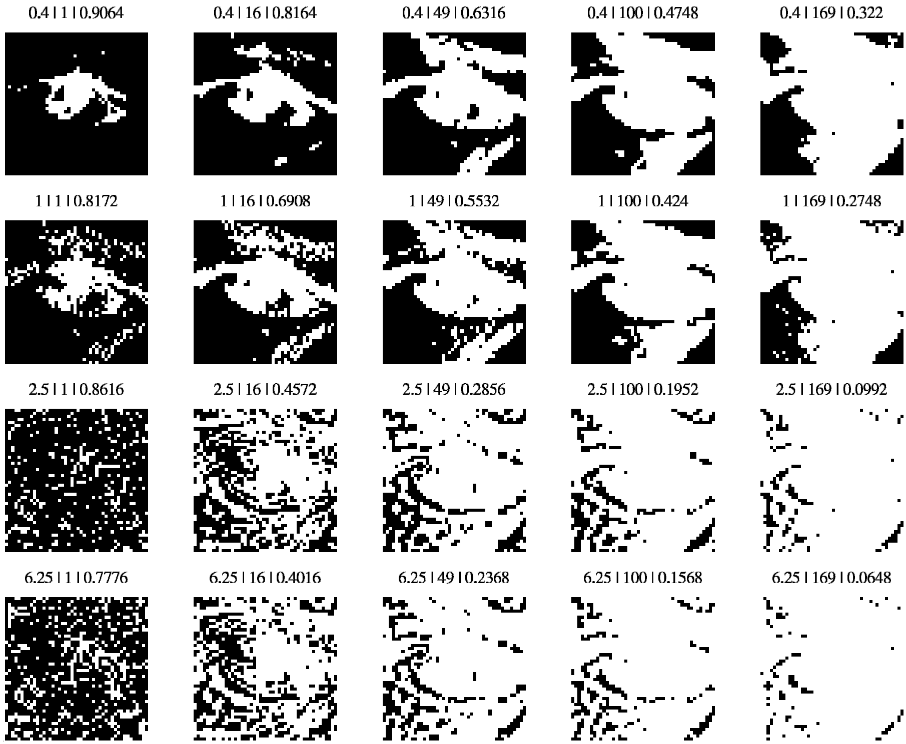

Fig. 4. In black are the positions that contain a support vector of the SBMs. The last number on top of every image gives the relative area of the image that is black, i.e., it gives the relative number of support vectors.

[13]. In addition to this, representations of higher order features may be incorporated, i.e., not only blobs, which makes a more closer connection to the work in, for example, [12] (see also Section 1) in which several receptive field weighting function are to be chosen in such a manner to represent the image in an optimal way. In this, also anisotropic forms of SBMs may be of interest.

References

1. R. van den Boomgaard and A. W. M. Smeulders. The morphological structure of images, the differential equations of morphological scale-space. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(11):1101–1113, 1994.
2. R. Duits, M. Felsberg, L. M. J. Florack, and B. Platel. α Scale spaces on a bounded domain. In *Scale Space Methods in Computer Vision, 4th International Conference, Scale-Space 2003*, pages 494–510, Isle of Skye, 2003.
3. L. M. J. Florack, R. Maas, and W. J. Niessen. Pseudo-linear scale-space theory. *International Journal of Computer Vision*, 31(2/3):247–259, April 1999.
4. F. Girosi. An equivalence between sparse approximation and support vector machines. *Neural Computation*, 10(6):1455–1480, 1998.

5. F. Girosi, M. Jones, and T. Poggio. Prior stabilizers and basis functions: From regularization to radial, tensor and additive splines. Technical Report AI Memo 1430, CBCL Paper 75, MIT, Cambridge, MA, 1993.
6. L. D. Griffin. Local image structure, metamerism, norms, and natural image statistics. *Perception*, 31(3), 2002.
7. B. M. ter Haar Romeny (ed.). *Geometry-Driven Diffusion*. Kluwer, Dordrecht, 1996.
8. P. J. Huber. *Robust Statistics*. John Wiley *&* Sons, New York, 1981.
9. P. R. Jackway. Morphological scale-space. In *11th IAPR International Conference on Pattern Recognition*, pages 252–255, The Hague, The Netherlands, 1992.
10. J. J. Koenderink. The structure of images. *Biological Cybernetics*, 50:363–370, 1984.
11. J. J. Koenderink and A. J. van Doorn. Metamerism in complete sets of image operators. In *Advances in Image Understanding '96*, pages 113–129, 1996.
12. M. Lillholm, M. Nielsen, and L. D. Griffin. Feature-based image analysis. *International Journal of Computer Vision*, 52:73–95, 2003.
13. T. Lindeberg. *Scale-Space Theory in Computer Vision*. Kluwer Academic Press, Boston, 1994.
14. M. Loog, M. Lillholm, M. Nielsen, and M. A. Viergever. Gaussian scale space from insufficient image information. In *Scale Space Methods in Computer Vision, 4th International Conference, Scale-Space 2003*, pages 757–769, Isle of Skye, 2003.
15. M. Nielsen, L. M. J. Florack, and R. Deriche. Regularization, scale-space, and edge detection filters. *Journal of Mathematical Imaging and Vision*, 7(4):291–307, 1997.
16. K.-R. Park and C.-N. Lee. Scale-space using mathematical morphology. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(11):1121–1126, 1996.
17. P. Perona and J. Malik. Scale space and edge detection using anisotropic diffusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12:429–439, 1990.
18. O. Scherzer and J. Weickert. Relations between regularization and diffusion filtering. *Journal of Mathematical Imaging and Vision*, 12:43–63, 2000.
19. A. J. Smola and B. Schölkopf. A tutorial on support vector regression. Technical Report NC-TR-98-030, Royal Holloway College, University of London, UK, 1998.
20. A. J. Smola, B. Schölkopf, and K.-R. Müller. The connection between regularization operators and support vector kernels. *Neural Networks*, 11:637–650, 1998.
21. A. N. Tikhonov and V. Y. Arseninn. *Solution of Ill-Posed Problems*. W. H. Winston, Washington D.C., 1977.
22. V. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, New York, 1995.
23. J. Weickert. *Anisotropic Diffusion in Image Processing*. B. G. Teubner, Stuttgart, 1998.
24. A. P. Witkin. Scale-space filtering. In *Proceedings of IJCAI*, Germany, 1983.