Scale-invariant medial features based on gradient vector flow fields

David Engel and Cristóbal Curio
Max Planck Institute for Biological Cybernetics
{*david.engel,cristobal.curio*}*@tuebingen.mpg.de*

Abstract

We propose a novel set of medial feature interest points based on gradient vector flow (GVF) fields [18]. We exploit the long ranging GVF fields for symmetry estimation by calculating the flux flow on it. We propose interest points that are located on maxima of that flux flow and offer a straight forward way to estimate salient local scales. The features owe their robustness in clutter to the nature of the GVF which accomplishes two goals simultaneously - smoothing of orientation information and its preservation at salient edge boundaries. A learning framework based on them, in contrast to classical edge-based feature detectors, would unlikely be distracted by background clutter and spurious edges, as these new mid-level features are shape-centered. We evaluate our scale-invariant feature coding scheme against standard SIFT keypoints by demonstrating generalization over scale in a patch-based pedestrian detection task.

1 Introduction

Many computer vision algorithms operate on local appearances extracted at non-accidental interest points [5][6][11][12]. A widely recognized problem is scale-invariant object recognition, e.g. [3]. One approach to achieve invariance against scale changes is to include objects at many different sizes in the training data or invoke the back-end computer vision algorithm on a pyramid of image resolutions. A codebook of such appearances would contain many redundancies, since features taken at the same part of an object at different scales will have different local appearances. Consequently, computations with such a codebook would not be as efficient as with a scale-invariant shape centered one. In this paper we explicitly address this problem and propose a novel set of shape-centered interest points derived from GVF fields. These fields are based on the method of variational calculus to solve complementary vision

goals formulated as an energy minimization problem. The cost function takes into account both smoothing of orientation information across the image and its preservation at salient edges. The GVF approach, implemented by PDEs, can closely approximate the gradient of a distance function, making our framework useful for many vision tasks even in cluttered environments. Our interest points form at locations of high symmetry corresponding to strong answers of flux flow derived from this vector field. Moreover, the same framework also allows one to derive salient local scales that we validate within an object detection framework. One property of this symmetry transformation is that it offers to find keypoints that carry specific information about object shape. We show the specificity of our features for learning an object representation that can generalize to scales not included during learning in a pedestrian detection task. In Section 3 we describe how these features are computed and show their properties. Section 4 gives an evaluation in a pedestrian detection task compared to standard SIFT keypoints. Finally, Section 5 concludes with a discussion.

2 Related Work

Mid-level features of intermediate complexity have been recognized as playing an important role for both human object recognition and computer vision systems [7][8][10][15]. Recently, mid-level features have been explicitly implemented into object recognition architectures inspired by classical Gestalt-Law principles [1].

In particular, Scale-invariant coding of image features has been addressed since the invention of e.g. SIFT features and region like feature detectors [6][11]. The development of region-based image coding descriptors is given more and more attention as they show their distinctive strength in segmentation and robustness on correspondence finding tasks [16]. Lately, the visual fragment theory [17] finds its application in patch extraction for recognition based on medial geometry [13]. Initially developed for long range contour alignments,

GVF has been also recently adopted for shape characterization tasks [4]. In this paper we demonstrate for the first time its strength for scale-invariant learning of object representation for detection in clutter.

3 Medial Feature Computation

Our medial feature interest points are computed as shock loci similar to [14][15] but are based on GVF fields. They are generated at points of high symmetry and offer a powerful way to describe local appearances.

GVF-based medial feature interest points. The GVF is the vector field $V(\mathbf{P}) = [u(\mathbf{P}), v(\mathbf{P})]$ that minimizes the energy function $\mathcal{E}$ (Eqn. 1), where $\mathbf{P} = (x, y)$ is a point in the image. It is obtained by iterative minimization of the cost function $\mathcal{E}$ based on variational calculus:

$$\mathcal{E} = \int \int \underbrace{g(|\nabla f|) |V - \nabla f|^2}_{\text{data term}} + \underbrace{h(|\nabla f|) \nabla^2 V}_{\text{smoothing term}} dxdy$$

(1)

The data term guarantees stability of the vector field $V(\mathbf{P})$ near an edge map f whereas the second term suppresses noise and propagates orientation information across the image. The two functions g and h determine a trade off between stability of orientation information at the gradients and its smoothness across the image. With the normalized solution, $V_N(\mathbf{P})$, to Eqn. 1 we determine symmetries and thereby approximate the gradient of the L_2-norm distance function, $\nabla D(\mathbf{P})$, by $\nabla D(\mathbf{P}) \approx V_N(\mathbf{P}) = V(\mathbf{P})/\|V(\mathbf{P})\| \ \forall \mathbf{P}$. Given this, points at which the vectors of the GVF field collide are points of high symmetry and are called *shock loci*. To find such loci [14] presents a medial feature detection framework based on an flux flow operator that we apply in our framework to the normalized GVF field V_N

$$\mathcal{F}(V_N(\mathbf{P})) = \frac{\oint \langle V_N, \mathcal{N} \rangle ds}{Area}.$$

(2)

Here V_N denotes the GVF field and $\mathcal{N}$ the normals on the ring through which the flux flow is computed. To obtain the medial feature interest points, we find the local maxima by a non-maximal suppression operation. All required operations can be computed efficiently allowing algorithms using medial feature interest points to run in real-time systems. Figure 1a renders the evolution of one GVF field and the corresponding response to the flux flow operator $\mathcal{F}(V_N(\mathbf{P}))$. Note that medial feature interest points are object-centered *internal features*, meaning they are unlikely to be generated by spurious overlap of background clutter and target objects.

Local scale estimation. The properties of the underlying GVF leads to interest points which are formed by at least two equidistant salient edges. Consequently, a local scale can be estimated by detecting the preserved image edges of the flux flow field $\mathcal{F}$ that led to the interest point. Estimating local scales at edge-based interest-points would be less intuitive since edges are by definition points where two objects of usually different scales intersect. We estimate the scale by computing the mean flux flow of discs centered at the interest points. Eqn. 3 shows the mapping $\mathcal{S} : r \rightarrow s$ which is derived from the flux flow field and maps from the radius of the disc r to the mean flux flow energy s

$$\mathcal{S}(r) = \frac{\int_0^{2\pi} \int_0^r \mathcal{F}(\alpha, r) \, dr d\alpha}{\pi r^2},$$

(3)

where $\mathcal{F}(\alpha, r)$ is the flux flow rewritten in polar coordinates around the interest point. $\mathcal{S}(r)$ is low for small discs, hits a local maximum where it fully encloses the edges which generated the interest point and drops for sizes beyond that. We define the local scale at the current interest point by finding the first local maximum of $\mathcal{S}(r)$. Eqn 3 can be computed efficiently on a discrete grid using lookup tables and an iterative accumulation scheme. Examples of medial feature interest points with the estimated local scales are shown in Figure 1 b), c) and d).

Patch-based descriptor. For the application presented below we employ quadratic patches from edge energy maps corresponding to the estimated local scales as descriptor. Other descriptors like histograms of oriented gradients would be applicable as well. The descriptor is based on scale-invariant interest points combined with an estimated local scale and thus provides a scale-invariant way to describe local appearance.

Properties. Our features are internal features, meaning that if they are located on the target object, ideally they have been formed by at least two salient object-edges, or at even more informative locations such as 'end-stop'- and 'junction'-points [2][7]. This means that they are more likely to contain valuable and compact information of one shape than descriptors computed at corner-based interest points which are formed at object boundaries and usually contain more clutter. The robustness of the features is demonstrated in a sequence of a pedestrian crossing a street (Figure 1d). Features and corresponding scales are stably detected over time and thus could also be used for tracking.

4 Scale-invariant object detection

We validated the scale-invariance of our medial features within a pedestrian detection task. Since pedestrians can appear in a large variety of scales and poses
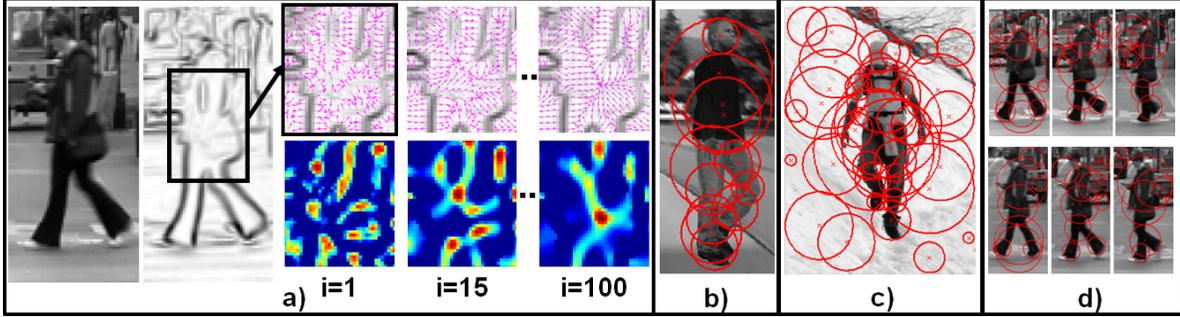
Figure 1. a) Evolving solution to GVF field, V_N, and clipped scalar flux flow, with $\mathcal{F} < 0$. Interest points on an image from our artificial dataset (b), the INRIA dataset (c), and on a recorded sequence (d).

it is a very challenging problem for algorithms based on template-matching. We want to stress that we use pedestrian detection only to demonstrate and compare the scale-invariance of our features. We do not aim at providing a state-of-the-art pedestrian detection algorithm.

A medial feature ISM. As an algorithmic back-end, we modified the implicit shape model (ISM) [9] to accommodate our features. The ISM has proven to be a powerful tool for object-detection and segmentation. Several extensions to cope with scale estimation and multi-category object detection have been shown. The ISM is based on a codebook of local appearances for which a set of votes is learned for possible object positions $\mathbf{x}$, relative to the interest point where the entry matched. Given a new local appearance $\mathbf{e}_k$ extracted around an interest point k in a test image, each matching codebook entry votes for possible object positions via the learned probability function $p(\mathbf{x} \mid \mathbf{e}_k)$. By summing over all N patches $\mathbf{e}_k$ which are extracted from the new image the final probability density in the voting space is obtained:

$$p(\mathbf{x}) = \sum_{k=1}^{N} p(\mathbf{x} \mid \mathbf{e}_k) \qquad (4)$$

Using the local scale estimation of our features as a pre-processing step for the template matching has two advantages: Firstly, it allows to capture a greater variety of unique local appearances in a codebook of the same size, since patches from the same object part at different sizes can be clustered together. Secondly, it allows the template matching scheme to generalize to a wide range of yet unseen scales.

Evaluation. To directly evaluate the stability of our features against scale changes we train the ISM on images containing pedestrians of a fixed size and evaluate

its detection performance on images containing persons of a different size. This allows us to measure the ability of the algorithm to transfer knowledge gained at one scale to a different one. The broader the range of scales to which the algorithm is able to generalize to, the more information can be shared between even more distant scales. This way we can measure how well the scale estimation step facilitates the integration of information from different scales.

We used two datasets for evaluation. The first one consists of 3D avatars driven by motion capture data which were inserted into clutter using alpha-blending (Figure 1b). It includes 56 images for each, training and testing. In the second one, we took images from the INRIA pedestrian database containing only one single upright person, allowing us to fix the size of the pedestrians (Figure 1c). This dataset is comprised of 100 images for training and 87 for testing. We compare three approaches with respect to their scale-invariance: Our medial features against Lowe's SIFT keypoint detector which also yields a local scale estimation. As a baseline we evaluate our medial features with a fixed local scale of 50 pixel, which is approximately the mean of the estimated local scales. Since the two features produce very different codebooks we have to limit the codebook sizes to keep them comparable. We limit the size of the codebook to 400 entries by drawing randomly from the full codebook. The number of votes in the ISM voting space is also limited the same way. As descriptor we always applied our patch based descriptor. The detection rates at different scales on both databases are reported in Figure 2. It can be seen that our features provide a useful shape descriptor in artificial and real world images with difficult background clutter even with a small codebook of only 400 entries. Furthermore, it is shown that the algorithm benefits greatly from the scale estimation the more the scale factor differs from 1. Without scale es-
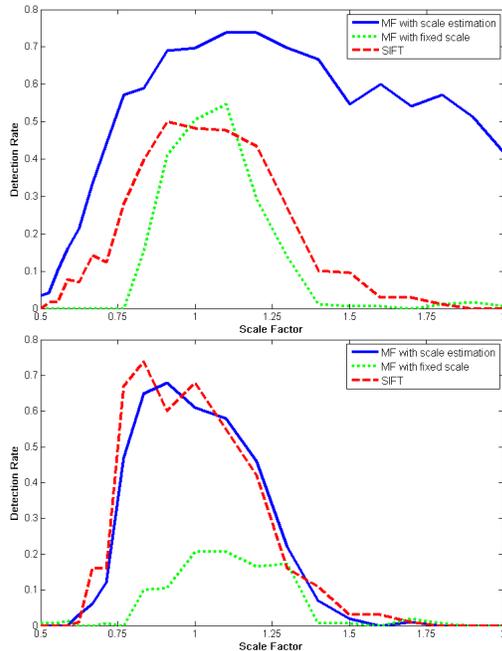
Figure 2. Detection rates on the computer generated (top) and on the INRIA database (bottom) over size scale ratio training vs test images.

timation we see worst performance, especially on the challenging real world imagery. Both graphs demonstrate that our medial features perform comparably to SIFT-based keypoints and depending on the dataset may even outperform them.

5 Discussion

We have provided a new feature set based on GVF-based medial features for computer vision tasks, with at least comparable performance to SIFT key point detectors. We want to point out that in most situations medial features and SIFT are not necessarily in competition to one another. Since edge-based and interior features may convey different information they can complement each other in an ISM that combines both. The medial features could be used as a binding force for low level corner features, allowing features generated by the same object to act in unison. Our approach would be equally well applicable to object tracking where smooth scale changes are frequently encountered and our features may prove useful. The code containing the parameters and implementation details is freely available at *http://www.kyb.mpg.de/~engel/medial_features.html*.

References

[1] S. Bileschi and L. Wolf. Image representations beyond histograms of gradients: The role of gestalt descriptors. In *CVPR*, pages 1–8, 2007.

[2] H. Blum. A transformation for extracting new descriptors of shape. In *Models for the Perception of Speech and Visual Form*, pages 363–380, 1967.

[3] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *CVPR*, volume 2, pages 264–271, 2003.

[4] W.-B. Goh and K.-Y. Chan. Shape description using gradient vector field histograms. In *Scale Space Methods in Computer Vision*, pages 1611–3349, 2003.

[5] C. Harris and M. Stephens. A combined corner and edge detector. In *4th Alvey Vision Conference*, pages 147–152. University of Sheffield Printing Unit, 1988.

[6] T. Kadir and M. Brady. Saliency, scale and image description. *Int. J. Comput. Vision*, 45(2):83–105, 2001.

[7] B. Kimia and A. Tamrakar. The role of propagation and medial geometry in human vision. *Journal of Physiology*, 97:155–190, 2003.

[8] T. Lee, D. Mumford, R. Romero, and V. Lamme. The role of the primary visual cortex in higher level vision. *Vision Research*, 38:2429–2454, 1998.

[9] B. Leibe, A. Leonardis, and B. Schiele. Robust object detection with interleaved categorization and segmentation. *Int. J. Comput. Vision*, 77(1-3):259–289, 2008.

[10] A. Levinshtein, C. Sminchisescu, and S. Dickinson. Learning hierarchical shape models from examples. In *Energy Minimization Methods in Computer Vision and Pattern Recognition*, pages 251–267, 2005.

[11] T. Lindeberg. Automatic scale selection as a pre-processing stage to interpreting real-world data. In *ICTAI*, pages 490–490, 1996.

[12] D. G. Lowe. Object recognition from local scale-invariant features. In *ICCV*, pages 1150–1157, 1999.

[13] O. C. Ozcanli and B. B. Kimia. Generic object recognition via shock patch fragments. In *British Machine Vision Conference*, pages 1030–1039, 2007.

[14] S. Pizer, K. Siddiqi, G. Szekely, and S. Zucker. Multiscale medial loci and their properties. *Int. Journal of Computer Vision*, 55(2-3):155–179, 2003.

[15] K. Siddiqi, S. Bouix, A. Tannenbaum, and S. Zucker. The hamilton-jacobi skeleton. In *ICCV*, pages 828–834, 1999.

[16] A. Toshev, J. Shi, and K. Daniilidis. Image matching via saliency region correspondences. In *CVPR*, 2007.

[17] S. Ullman. Object recognition and segmentation by a fragment-based hierarchy. *Trends in Cognitive Sciences*, 11(2):58–64, February 2007.

[18] C. Xu and J. Prince. Gradient vector flow: A new external force for snakes. In *CVPR*, volume 1, pages 66–71. IEEE, 1997.