Interactive Motion Segmentation

Claudia Nieuwenhuis[1], Benjamin Berkels[2], Martin Rumpf[2], Daniel Cremers[1⋆]

[1] Technical University of Munich, Germany
{nieuwenhuis,cremers}@in.tum.de
[2] University of Bonn, Germany
{berkels,rumpf}@ins.uni-bonn.de

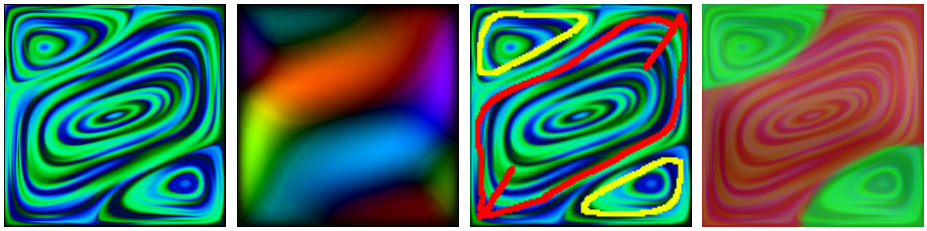
Abstract. Interactive motion segmentation is an important task for scene understanding and analysis. Despite recent progress state-of-the-art approaches still have difficulties in adapting to the diversity of spatially varying motion fields. Due to strong, spatial variations of the motion field, objects are often divided into several parts. At the same time, different objects exhibiting similar motion fields often cannot be distinguished correctly. In this paper, we propose to use spatially varying affine motion model parameter distributions combined with minimal guidance via user drawn scribbles. Hence, adaptation to motion pattern variations and capturing subtle differences between similar regions is feasible. The idea is embedded in a variational minimization problem, which is solved by means of recently proposed convex relaxation techniques. For two regions (i.e. object and background) we obtain globally optimal results for this formulation. For more than two regions the results deviate within very small bounds of about 2 to 4 % from the optimal solution in our experiments. To demonstrate the benefit of using both model parameters and spatially variant distributions, we show results for challenging synthetic and real-world motion fields.

1 Introduction

Motion segmentation refers to grouping together pixels undergoing a common motion. It aims at segmenting an image into moving objects and is a powerful cue for image understanding and scene analysis. For a semantic interpretation of a sequence motion is an important feature just like color or texture. For tracking and video indexing it is often desirable to divide the scene into foreground and background objects and to perform independent motion analysis for both classes. Another perspective application is video compression, where several encoding standards such as MPEG represent a sequence as objects on a series of layers and, hence, require the objects to be identified before encoding.
Most motion segmentation methods identify objects by grouping pixels with approximately constant motion vectors. This approach leads to several problems.

1. Object motion is often characterized by complex motion patterns such as vortices or curls, which are impossible to be segmented based on constant

⋆ This work was supported by the DFG SPP 1335 and grant CR250/6-1



a) Input image b) HSV Motion Field c) User Scribbles d) Interact. Segment.

Fig. 1. Segmentation of a motion field where traditional motion segmentation approaches fail. The proposed algorithm allows to compute interactive image segmentations based on a spatially varying partitioning of the velocity field (motion field input data courtesy of Eberhard Bänsch, University of Erlangen-Nuremberg).

motion vectors (Figure 1 shows a water vortex with motion vectors pointing in all directions along the current). The computation of affine model parameters instead in combination with a spatially varying distribution allows for the grouping of vectors belonging to the same motion pattern.

2. Objects sometimes undergo similar motion as other objects or the background. In such cases, segmentation approaches not taking into account the spatial variance of the motion will fail to separate between similarly moving objects.
3. Moving objects, which are not planar, exhibit different motion in different parts of the object due to changing depth. Also in such cases segmentation based on a constant motion field will fail to recognize all parts belonging to the moving object, e.g. a human with fast moving arms but head and legs almost at rest.
4. Frequently, there are several similar foreground objects which follow different motion patterns, e.g. cars on a road. Similarity based segmentation approaches will not assign all of them to the same class.

To summarize, automatic motion segmentation is often problematic and even for humans it is not clear, where a motion pattern begins and where it ends. Take for example neighboring vortices in water or heat flows (Figure 1) or sequences with several distinct foreground objects belonging to the same object class (Figure 4). By means of minimal user interaction such semantic problems can be overcome.

1.1 Related Work

Non-interactive motion segmentation has been studied in the literature in particular within the scope of optical flow estimation. In [1], Cremers and Soatto introduce the concept of motion competition, i.e. a variational model that, given two consecutive frames of a video, estimates the motion between the given frames and jointly segments the image domain based on the estimated motion. Therein, a parametric motion model is used and particularly the case of piecewise affine motion is considered. Their target functional can be understood as an extension

of the Mumford–Shah functional [2] and the applied minimization techniques include a multiphase level set formulation based on the Vese–Chan model [3]. Brox et al. [4] propose a variational approach that combines optic flow estimation with the segmentation of the image domain into regions of similar optical flow, extending the motion competition concept to non-parametric motion and a more elaborate data term for the motion estimation while still using a multiphase level set formulation.

Independently, interactivity has proven itself as a feasible method to facilitate difficult image segmentation tasks. For instance, Bai and Sapiro [5] presented an interactive framework for image and video segmentation. Their technique calculates weighted geodesic distances to scribbles interactively provided by the user, where the weighting is based on spatial or temporal gradients. The segmentation then is obtained automatically from these distances. More related to our approach is the TVSeg algorithm by Unger et al. [6] that also incorporates user interaction in the segmentation of images into foreground and background. The actual segmentation uses an geodesic active contour model [7] that integrates the user input as local constraint and is minimized numerically in a globally optimal manner using a total variation based reformulation.

1.2 Contribution

The contribution of this paper is the introduction of locally adaptive model parameter distributions into a variational motion segmentation approach. Instead of learning a global motion vector distribution for each object, we make two important modifications. First, we do not estimate the probability of the motion vectors directly but of their motion model parameters instead. In this way, the similarity of vectors belonging to the same moving object is preserved and issue 1 solved. Second, as different objects and object parts can still exhibit varying affine motion we model a spatially variant distribution, which allows for changing motion at different image locations. We, thus, solve issues 2 to 4. The locally adaptive parameter distributions are introduced into a variational framework which allows for globally optimal segmentation results in case of two regions and near globally optimal results for more than two regions.

2 A Bayesian Approach to Motion Segmentation

Let $v : \Omega \to \mathbb{R}^2, \Omega \subset \mathbb{R}^b, b \in \mathbb{N}$ denote a given motion field. Motion segmentation is the computation of such a labeling function $u : \Omega \to \{1, .., n\}$ assigning a specific label $u(x)$ to each pixel $x \in \Omega$ based on the motion vector $v(x)$, such that the $\Omega_i = \{x \in \Omega | u(x) = i\}$ are disjoint and $\overline{\Omega} = \bigcup_{i=1}^{n} \overline{\Omega}_i$.

2.1 A Parametric Motion Field Representation

Typical motion vectors resemble specific motion patterns. The easiest pattern would be a constant planar motion, more difficult ones are for example rotations,

curls or vortices with spatially varying motion vectors. Only in the first case a simple grouping of motion vectors can be successful. In order to preserve the similarity of different motion vectors belonging to the same object, we describe the motion field by means of model parameters. Similar model parameters then hint at a common motion pattern. In this paper, we assume an affine motion model and compute the affine model parameters $s : \Omega \to \mathbb{R}^6$ by solving the following minimization problem for each pixel $x \in \Omega$

$$s(x) = \arg\min_{s \in \mathbb{R}^6} \int_\Omega G_\sigma(y - x) \left| p(y) \cdot s - v(y) \right|^2 dy, \tag{1}$$

where $p(y) = \begin{pmatrix} y_1 & y_2 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & y_1 & y_2 & 1 \end{pmatrix}$ and $G_\sigma(x)$ is a Gaussian function with variance σ and mean 0. In our experiments, we set σ to the pixel size. The optimization problem can be solved by means of a weighted least squares approach. If we only aim at the identification of particular types of affine motion we can replace the general model $p(y)$ by a more specific one. E.g. in case of vortices we consider skew symmetric affine maps, i.e. looking for a vector of five affine parameters $s : \Omega \to \mathbb{R}^5$ and replacing p by $p(y) = \begin{pmatrix} y_1 & y_2 & 1 & 0 & 0 \\ 0 & -y_1 & 0 & y_2 & 1 \end{pmatrix}$. Alternatively, one could keep s and p as they are and penalize the defect from the desired family of affine motion by an additional energy term, e.g. in our case $|s_2(x) + s_4(x)|^2$.

2.2 The Bayesian Formulation

We want to maximize the conditional probability of u in a Bayesian framework given the motion parameter field s

$$\arg\max_u \mathcal{P}(u \,|\, s) = \arg\max_u \frac{\mathcal{P}(s \,|\, u)\ \mathcal{P}(u)}{\mathcal{P}(s)} = \arg\max_u \mathcal{P}(s \,|\, u)\ \mathcal{P}(u). \tag{2}$$

Assuming that all affine parameter vectors are independent of each other – but in contrast to previous approaches not independent of space – we obtain

$$\arg\max_u \mathcal{P}(s \,|\, u)\ \mathcal{P}(u) = \arg\max_u \left(\prod_{x \in \Omega} \Big(\mathcal{P}(s(x) \,|\, x, u) \Big)^{dx} \right)\ \mathcal{P}(u), \tag{3}$$

where the exponent dx denotes an infinitesimal volume in $\mathbb{R}^b$ and assures the correct scaling for decreasing pixel size. Preserving the dependence of the model parameters on the spatial position is an indispensable prerequisite to cope with objects effected by different and frequently non constant motion patterns. Such important information is entirely lost in the traditional space-independent formulation. Consequently probability density functions that can be easily separated in parameter-location-space can overlap and make the separation of objects impossible if only the parameter space is taken into account.
Since the probability of a parameter vector is independent of labels of other pixels, we deduce from (3) that

$$\prod_{x \in \Omega} \Big(\mathcal{P}(s(x) \,|\, x, u) \Big)^{dx} = \prod_{i=1}^{n} \prod_{x \in \Omega_i} \Big(\mathcal{P}(s(x) \,|\, x, u(x) = i) \Big)^{dx}. \tag{4}$$

2.3 Spatially Varying Parameter Distributions

$\mathcal{P}(s(x)\,|\,x, u(x) = i)$ denotes the conditional probability of a parameter vector $s(x)$ at location x in the motion field provided x belongs to region Ω_i. These spatially varying probability distributions for each object class i are learned from user scribbles. Let $T_i := \{(x_i^j, s(x_i^j)), j = 1, .., m_i\}$ denote the set of user markings consisting of locations x_i^j and corresponding model parameter vector $s(x_i^j)$ for $x_i^j \in \Omega_i$. We can estimate the probability from user scribbles by means of the Parzen-Rosenblatt [8, 9] estimator

$$\hat{\mathcal{P}}(s(x)\,,\, x|u(x) = i) = \frac{1}{m_i} \sum_{j=1}^{m_i} G_\Sigma \left((x, s(x)) - (x_i^j, s(x_i^j)) \right). \tag{5}$$

Here, G_Σ denotes the multivariate Gaussian kernel centered at the origin with covariance matrix Σ. For uniformly distributed samples this estimator converges to the true probability distribution for $m_i \to \infty$ [10]. In case of user scribbles, however, the samples are spatially not uniformly distributed. Therefore, we make use of the separability of the Gaussian kernel and choose Σ such that

$$G_\Sigma \left((x, s(x)) - (x_i^j, s(x_i^j)) \right) = G_\rho(x - x_i^j) G_\sigma(s(x) - s(x_i^j)) \tag{6}$$

Commonly, the spatial variance, $G_\rho(x - x_i^j)$, has been neglected so far. We will call this previous approach the spatially invariant approach.
We now introduce a spatially variable kernel function by choosing the spatial kernel width $\rho(x)$ at image location x proportional to the distance from the k-th nearest sample point $x_{v_k} \in T_i, \rho(x) = \alpha \|x - x_{v_k}\|_2$.

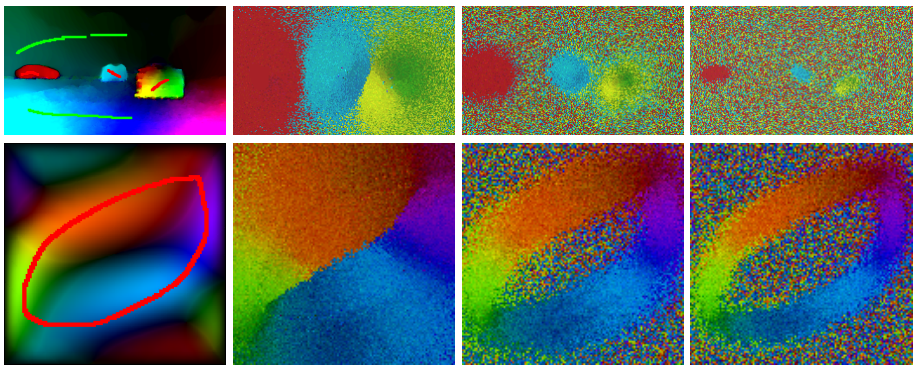
$$\hat{\mathcal{P}}(s(x)\,,\, x|u(x) = i) = \frac{1}{m_i} \sum_{j=1}^{m_i} G_{\rho(x)}(x - x_i^j) G_\sigma(s(x) - s(x_i^j)). \tag{7}$$

Thus, the influence of each sample point in T_i at a given location x is determined by the distance of the k-th nearest neighbor to x. If many sample points are close to x, $\rho(x)$ becomes small and the corresponding kernel becomes peaked. Hence, the influence of the samples further away is reduced. In contrast, if no samples are close by $G_\rho(x)$ tends towards a uniform distribution as in the spatially independent approach. Therefore, the spatially variant approach can be understood as a generalization of the original, spatially independent approach. The spatially variant approach yields a different parameter distribution for each location in the motion field, whereas the original, invariant approach yields the same distribution at all locations. Using

$$\hat{\mathcal{P}}(s(x)\,|\,x, u(x) = i) = \frac{\hat{\mathcal{P}}(s(x)\,,\, x|u(x) = i)}{\hat{\mathcal{P}}(x|u(x) = i)} = \frac{\hat{\mathcal{P}}(s(x)\,,\, x|u(x) = i)}{\int_s \hat{\mathcal{P}}(s, x|u(x) = i) ds} \tag{8}$$

we can now derive the conditional probability of a parameter vector $s(x)$ given at location x and label i based on user scribbles T_i.

The parameter α directly determines the variance of the kernel function k and, thus, the locality of the user input. The smaller α the more locally limited is the influence of the user scribbles. This effect can be examined by means of motion synthesis shown in Figure 2. Motion synthesis means that we randomly draw samples from the foreground distribution by means of the inverse distribution function. For all experiments done in Section 3, we set $\alpha = 0.3$ and $k = 10$.



a) original flow b) $\alpha = 0.1$ c) $\alpha = 0.3$ d) $\alpha = 0.5$

Fig. 2. The influence of the user scribbles on their neighborhood is determined by the parameter α and can be examined by means of motion synthesis. Here, for each pixel we randomly draw a motion vector from the spatially varying distribution. The smaller α the more local is the influence of the user scribbles and the more deterministic is the drawn motion vector.

2.4 The Variational Approach

To solve the optimization problem (3) we specify the prior $\mathcal{P}(u)$ favoring shorter boundaries between different regions, i.e. $\mathcal{P}(u) \propto \exp\left(-\frac{\lambda}{2}\sum_{i=1}^{n} \mathrm{Per}(\Omega_i, \Omega)\right)$, where $\mathrm{Per}(\Omega_i, \Omega)$ denotes the perimeter of Ω_i in Ω, cf. [11]. The optimization problem can be solved by minimizing its negative logarithm

$$\mathcal{E}(\Omega_1, \ldots, \Omega_n) = \frac{\lambda}{2}\sum_{i=1}^{n} \mathrm{Per}(\Omega_i, \Omega) + \sum_{i=1}^{n} \int_{\Omega_i} f_i(x)\, dx, \text{ with} \tag{9}$$

$$f_i(x) = -\log \sum_{j=1}^{m_i} k_{\rho(x)}(x - x_i^j) k_\sigma(s(x) - s(x_i^j)) + \log \sum_{j=1}^{m_i} k_{\rho(x)}(x - x_i^j). \tag{10}$$

Using the coarea formula in BV, the function space of bounded variation (cf. [11]), we can replace the sum of the perimeters by the total variation of u and arrive at energy minimization problem $\sum_{i=1}^{n} \int_{\Omega_i} f_i(x)\, dx + \lambda \int_\Omega |D\, u|\, dx \to \min$. To transform this energy minimization into a convex variational problem we apply the multilabel approach [12] in combination with [13] by Pock et al., which

is solved numerically in a primal-dual-minimization scheme. To this end, the multilabel function $u : \Omega \to \{1, \ldots, n\}$ is expressed in terms of its upper level sets, i.e. $\theta_i(x) = 1$ if $u(x) \geq i + 1$ and else 0 for all $i = 1, \ldots, n-1$, where $\theta \in \mathrm{BV}(\Omega, \{0,1\})^n$ and $\theta_0 = 1$ and $\theta_n = 0$. The final energy to be minimized is

$$\min_{\theta \in \mathcal{B}} \sup_{\xi \in \mathcal{K}} \left\{ -\lambda \sum_{i=0}^{n-1} \int_\Omega \theta_i \ \mathrm{div}\ \xi_i \ dx + \int_\Omega |\theta_i(x) - \theta_{i+1}(x)| \, f_i(x) \ dx \right\} \tag{11}$$

with $\mathcal{B}$ and $\mathcal{K}$ defined as

$$\mathcal{B} = \{\theta = (\theta_1, .., \theta_{n-1}) \in \mathrm{BV}(\Omega, \{0,1\})^{n-1} \ \big| \ 1 \geq \theta_1 \geq \ldots \geq \theta_{n-1} \geq 0\} \tag{12}$$

$$\mathcal{K} = \left\{ \xi = (\xi_1, .., \xi_{n-1}) \in C_c^1(\Omega, \mathbb{R}^b)^{n-1} \middle| \left| \sum_{i_1 \leq i \leq i_2} \xi_i(x) \right| \leq 1 \ \ \forall i_1 \leq i_2 \right\} \tag{13}$$

where $\xi_i \in C_c^1(\Omega, \mathbb{R}^2)$ denotes the dual variable and C_c^1 the space of smooth functions with compact support.

Proposition 1. *Let $u' \in \mathcal{B}$ be the global minimizer of the original problem (11), $u^* \in \mathcal{B}$ the binarized solution of the relaxed problem and $\tilde{u} \in \tilde{\mathcal{B}}$ the result of the proposed algorithm, where*

$$\tilde{\mathcal{B}} = \{\theta = (\theta_1, .., \theta_{n-1}) \in BV(\Omega, [0,1])^{n-1} \ \big| \ 1 \geq \theta_1 \geq \ldots \geq \theta_{n-1} \geq 0\}. \tag{14}$$

Then an energy bound $\gamma(u^*, \tilde{u})$ exists such that $E(\tilde{u}) - E(u') \leq \gamma(u^*, \tilde{u})$.

Proof. Since $\mathcal{B} \subset \tilde{\mathcal{B}}$, we have $E(u^*) \leq E(u')$. Therefore,

$$E(\tilde{u}) - E(u') \leq E(\tilde{u}) - E(u^*) =: \gamma(u^*, \tilde{u}). \tag{15}$$

3 Results

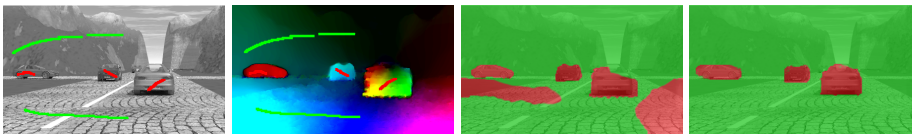
In this section we provide experimental results for the interactive segmentation of real and synthetic motion fields. We compare four settings: model-independent and model-based (see section 2.1), spatially invariant and spatially varying distributions (see section 2.3).

3.1 Model Based vs. Non Model Based

Since motion vectors belonging to the same motion pattern often exhibit very different direction and length, it is important to segment model parameter maps instead of the motion field itself. Difficulties arise next to motion boundaries, where different motion models coincide. These situations lead to large residuals in the least squares approach (1) and can, thus, easily be detected. We set all data terms to 0 in these situations. Figure 3 shows a segmentation example, which demonstrates that segmentations based on affine parameter maps usually yield better results than segmentations based on the motion field itself. It displays four planes varying in depth, which strongly influences the speed at different locations of the planes. The parameter map reduces this effect and even reveals underlying structure and, thus, makes an (almost) correct segmentation possible.

a) image b) HSV c) affine d) no model e) model

Fig. 3. Comparison of segmentation results based on the original flow field and on affine parameter maps using the spatially invariant dataterm, a) underlying image data, b) HSV-coded motion field with user scribbles, c) affine parameter map, d) segmentation based on motion only, e) segmentation based on affine parameter map.



a) image b) HSV c) invariant d) variant

Fig. 4. Segmentation results based on the spatially variant compared to the spatially invariant dataterm, a) underlying image data, b) HSV-coded motion field with user scribbles, c) segmentation based on spatially invariant dataterm, d) segmentation based on spatially variant dataterm.

3.2 Spatially Variant vs. Spatially Invariant Distributions

There are several situations where the spatial adaptability of the estimated motion distributions is indispensable, e.g. in case of different objects exhibiting similar motion or in case of one or similar objects exhibiting different motion patterns in different locations. Figure 4 shows results for spatially variant compared to spatially invariant distributions on a dataset with three cars on a road exhibiting very different motion direction and speed. These variations are captured by the spatially variant distribution.

3.3 Model Based Spatially Variant Distributions

In order to allow for spatially changing motion models we combine the spatially variant and the model based approach by computing spatially variant parameter distributions. Figure 5 shows original HSV-coded motion fields with user scribbles, the original segmentation result without parameter maps based on spatially invariant distributions and the improved segmentation result based on parameter maps and spatially adaptive parameter distributions. In case of more than two regions, a global optimal solution cannot be guaranteed.

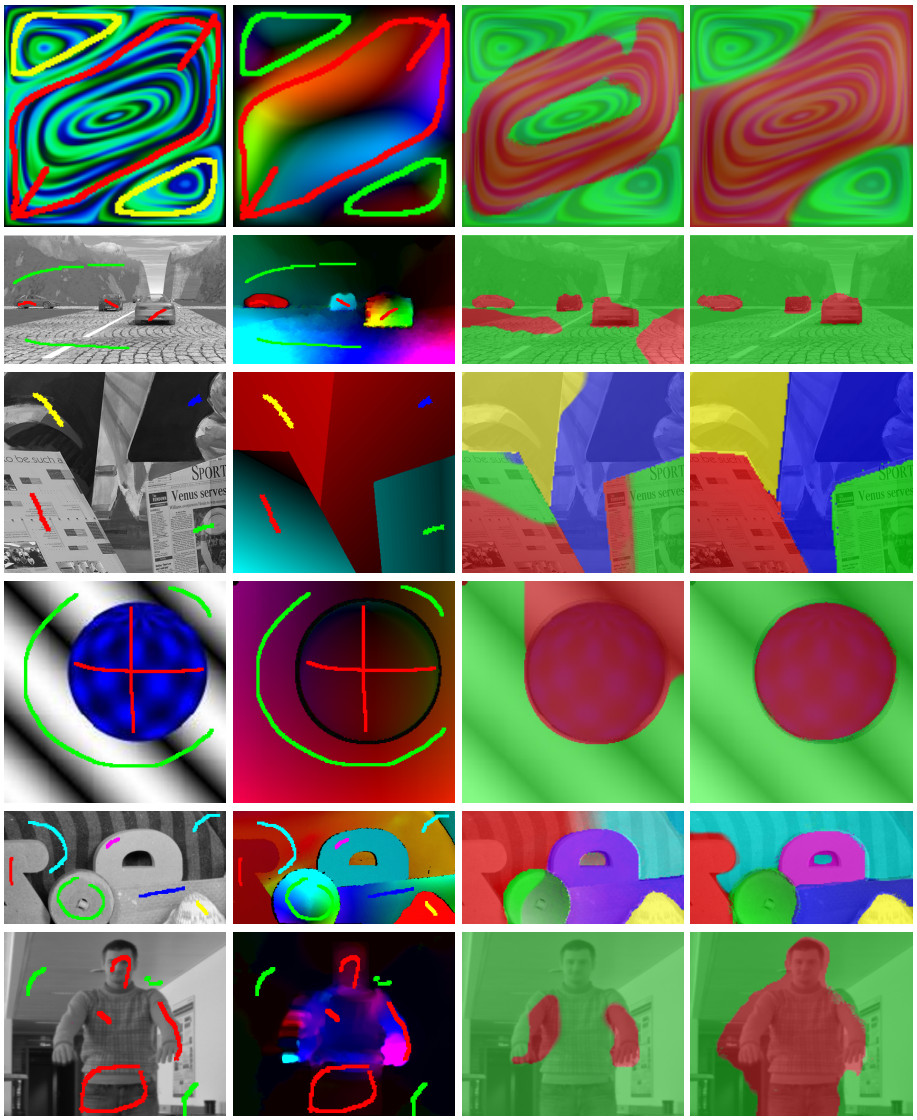
In our experiments, the energy gap between the binarized relaxed and the optimal solution lies between 2 and 4 % of the original energy (numerically evaluated using Proposition 1) and confirms that the solutions for more than two regions are very close to the globally optimal solution.

4 Conclusion

In this paper, we proposed an algorithm for interactive motion segmentation, which is based on spatially variant motion model parameter distributions. The suggested segmentation algorithm provides two advancements: 1) it reliably detects regions of difficult motion patterns such as vortices or curls due to its operation in the motion model parameter space, 2) it can handle even spatially varying motion patterns due to the spatial adaptivity of the parameter distributions. Few user indications are sufficient to accurately segment objects with strongly varying motion. The approach is formulated as a convex energy minimization problem, which yields the global optimum for two regions and nearly optimal results for more than two regions.

References

1. Cremers, D., Soatto, S.: Motion Competition: A variational framework for piecewise parametric motion segmentation. Int. J. of Comp. Vis. **62**(3) (2005) 249–265
2. Mumford, D., Shah, J.: Optimal approximations by piecewise smooth functions and associated variational problems. Comm. Pure Appl. Math. **42** (1989) 577–685
3. Vese, L., Chan, T.: A multiphase level set framework for image processing using the Mumford–Shah functional. Int. J. of Comp. Vis. **50**(3) (2002) 271–293
4. Brox, T., Bruhn, A., Weickert, J.: Variational motion segmentation with level sets. In: European Conference on Computer Vision (ECCV). (May 2006) 471–483
5. Bai, X., Sapiro, G.: A geodesic framework for fast interactive image and video segmentation and matting. In: IEEE Int. Conf. on Comp. Vis. (ICCV). (2007)
6. Unger, M., Pock, T., Cremers, D., Bischof, H.: TVSeg - Interactive total variation based image segmentation. In: British Machine Vision Conference (BMVC). (2008)
7. Caselles, V. and Kimmel, R.a.S.G.: Geodesic active contours. Int. J. of Comp. Vis. **22**(1) (1997) 61–79
8. Rosenblatt, M.: Remarks on some nonparametric estimates of a density function. Annals of Mathematical Statistics **27** (1956) 832–837
9. Parzen, E.: On estimation of a probability density function and mode. Annals of Mathematical Statistics **33** (1962) 1065–1076
10. Silverman, B.W.: Density estimation for statistics and data analysis. Chapman and Hall, London (1992)
11. Ambrosio, L., Fusco, N., Pallara, D.: Functions of bounded variation and free discontinuity problems. Oxford University Press (2000)
12. Pock, T., Chambolle, A., Bischof, H., Cremers, D.: A convex relaxation approach for computing minimal partitions. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Miami, Florida (2009)
13. Pock, T., Cremers, D., Bischof, H., Chambolle, A.: An algorithm for minimizing the piecewise smooth mumford-shah functional. In: IEEE Int. Conf. on Computer Vision, Kyoto, Japan (2009)



a) image b) HSV field c) no model, invar. d) model, variant

Fig. 5. Segmentation results based on the spatially variant, affine parameter distributions for HSV coded motion fields. a) underlying image data, b) HSV-coded motion field, c) result of non-model based, spatially invariant approach, d) result of model-based, spatially variant approach.