Optimal Dominant Motion Estimation using Adaptive Search of Transformation Space

Adrian Ulges[1], Christoph H. Lampert[2], Daniel Keysers[3], Thomas M. Breuel[1]

[1] Department of Computer Science, Technical University of Kaiserslautern
{a_ulges,tmb}@informatik.uni-kl.de
[2] Department for Empirical Inference
Max-Planck-Institute for Biological Cybernetics, Tübingen
chl@tuebingen.mpg.de
[3] Image Understanding and Pattern Recognition Group
German Research Center for Artificial Intelligence (DFKI), Kaiserslautern
keysers@iupr.net

Abstract. The extraction of a parametric global motion from a motion field is a task with several applications in video processing. We present two probabilistic formulations of the problem and carry out optimization using the RAST algorithm, a geometric matching method novel to motion estimation in video. RAST uses an exhaustive and adaptive search of transformation space and thus gives – in contrast to local sampling optimization techniques used in the past – a globally optimal solution. Among other applications, our framework can thus be used as a source of ground truth for benchmarking motion estimation algorithms.
Our main contributions are: first, the novel combination of a state-of-the-art MAP criterion for dominant motion estimation with a search procedure that guarantees global optimality. Second, experimental results that illustrate the superior performance of our approach on synthetic flow fields as well as real-world video streams. Third, a significant speedup of the search achieved by extending the model with an additional smoothness prior.

1 Introduction

We address the estimation of the dominant parametric motion from a sequence of video frames. Such dominant motion is usually equated with background motion, and its precise and robust estimation is required for several applications in the context of video analysis, like motion-based segmentation or motion compensation (which again serves as a building block in modern video encoders, or in video mosaicing).

Like most practical video processing systems, we estimate a global parametric motion from a field of local motion probes – a problem that is difficult due to measurement noise, inaccuracies of the previous motion estimation step, and deviant foreground motion. In terms of dominant motion estimation, such foreground motion probes are "outliers" that have to be recognized and discarded during the fitting process.

We view the problem from a parameter estimation perspective and propose two Bayesian formulations, one of them including a smoothness prior. The resulting optimization problems are solved using the RAST algorithm [3]. While other methods are based on a local sampling of search space and do not guarantee optimal solutions, RAST performs an adaptive, but exhaustive branch-and-bound search and finds the global optimum. This fact is proven by experimental results on synthetic motion fields as well as real-world video data.

Our main contributions are: first, the novel combination of a state-of-the-art MAP criterion with a search procedure that guarantees global optimality up to any accuracy desired. Second, experimental results that illustrate the superior performance of our approach on synthetic flow fields as well as real-world video streams. Third, a novel extension to the RAST algorithm with a smoothness prior that leads to a better search strategy with a significant speedup.

2 Related Work

Motion interpretation has often been called a "chicken-egg" problem: motion estimation is inaccurate without knowledge of motion boundaries due to the aperture problem [1], while on the other hand motion segmentation requires local motion estimates.

Methods to solve this problem can be divided into direct and indirect (or "feature-based" [8]) methods. Approaches from the first category jointly estimate motion and group it into coherent regions. Some estimate a parametric motion over image regions – like regression [1], mixture models [9], clustering methods [16], or formulations imposing additional shape priors [4]. Other direct methods are nonparametric and assume piecewise smoothness of the motion field, which leads to formulations related to Markov Random Fields [12, 17].

In contrast to this, indirect methods are two-step procedures: first, a motion field is estimated using correlation-based techniques [15], feature tracking [14], or optical flow. The result forms the input to a segmentation step, which must cope with local outliers and inaccuracies due to noise in the measurement process, error-prone motion estimation, and foreground objects in motion. For this, greedy local search procedures have been used in the past, like robust least squares, RANSAC [5], least median of squares, or least trimmed squares [10].

Since local errors in the motion estimation step cannot be undone, indirect methods do not reach the robustness of direct ones. Nevertheless, they offer simple and fast alternatives that are more popular in practice, and are applied to several video processing tasks, like in state-of-the-art video codecs or video mosaicing [13]. Our approach belongs to this second category. More precisely, we assume a motion field is given and focus on the motion interpretation step.

3 Statistical Framework

We assume that a motion field $D = \{(x_1, v_1), .., (x_n, v_n)\}$ of 2D positions x_i associated with 2D motion vectors v_i is given. These probes can correspond to

a dense optical flow field, or to sparse probes obtained from block matching or tracked point features.

The task is now to extract a parametric motion $v_\theta : \mathcal{R}^2 \rightarrow \mathcal{R}^2$ that fits D "well", i.e. $v_i \approx v_\theta(x_i)$. Such parameterized motion has proven a simple and often sufficiently accurate approximation to projected 3D scene motion. From the parameterizations proposed in the literature [8, 13], we choose the similarity transform consisting of a rotation by an angle α, a scaling s (e.g., due to zooming), and a translation $(d_x, d_y)^T$.

As an optimality criterion, we use a statistical formulation of the problem, i.e. we choose the global motion $\hat{\theta} = (\hat{s}, \hat{\alpha}, \hat{d_x}, \hat{d_y})$ that maximizes the posterior:

$$\hat{\theta} = \arg\max_\theta P(\theta|D) = \arg\max_\theta P(D|\theta) \cdot P(\theta) \tag{1}$$

3.1 Criterion Q_1: Local Independence

For our first formulation, we assume a uniform prior $P(\theta)$ and independent motion probes drawn from a distribution $p(v_i|\theta)$. If we also neglect competitive foreground motion and use isotropic Gaussian noise to model inaccuracies of motion estimation and of the capturing process, $p(v_i|\theta)$ is a Gaussian distribution with mean $v_\theta(x_i)$ and diagonal covariance $\sigma^2 I$. In practical flow fields, however, *outliers* occur – again, due to inaccuracies of the motion estimation process, but also due to foreground objects moving in a different direction. Since we do not have prior knowledge about the motion of such objects, we assume a uniform distribution $p(v_i|\theta) = c$ of foreground motion. This gives a more realistic scenario including outliers:

$$p(v_i|\theta) \propto \max\left(\mathcal{N}(v_i; v_\theta(x_i), \sigma^2 I), c\right) \tag{2}$$

We insert this term into the overall likelihood and obtain

$$p(D|\theta) = \prod_i p(v_i|\theta). \tag{3}$$

Maximizing this is equivalent to maximizing the following quality function derived from the log-likelihood (for a detailed derivation, see [18]):

$$Q_1(\theta) = \sum_i \max\left(1 - \frac{(v_i - v_\theta(x_i))^2}{\epsilon^2}, 0\right) =: \sum_i q(v_i, \theta). \tag{4}$$

The only free parameter of this ML criterion, ϵ, determines the allowed deviation of a background motion sample from the parametric motion v_θ. Note that Q_1 consists of local contributions $q(v_i, \theta)$ from the single flow samples, which are in the following referred to as the *support* of a local flow probe v_i for a global motion θ. This support is zero exactly if v_i deviates by ϵ or more from the model motion $v_\theta(x_i)$ (i.e. if v_i is regarded as an outlier). Thus, the evaluation of Q_1 provides a segmentation of the motion field into background and foreground.

3.2 Criterion Q_2: Spatial Coherence Prior

The optimality criterion Q_1 introduced in Equation (4) is derived from the likelihood and neglects the spatial coherence with which motion occurs in real-world videos. Like other researchers before, we use this fact by formulating an additional prior related to formulations in Markov Random Fields [1, 6, 17].

For this, we first introduce a segmentation as a *labeling* of the motion vectors $L : \{x_1, .., x_n\} \rightarrow \{0, 1\}$ such that $L(x_i) = L_i = 1$ iff v_i belongs to the background (which is the case exactly if $q(v_i, \theta) > 0$). Note that – given such a labeling – we can automatically compute a motion $\theta(L)$ as a least squares solution over the motion probes in the background region $L^{-1}(1)$. This is why – instead of searching for a motion θ – we instead search for an optimal labeling by maximizing the posterior:

$$P(L|D) \propto P(D|L) \cdot P(L) = P(D|\theta(L)) \cdot P(L) \tag{5}$$

The first term corresponds to the likelihood criterion from Equation (3). For the prior $P(L)$, we define a neighborhood structure over the motion field sites $\{x_i\}$ (for example, 4-connectedness on a regular grid of sites x_i), which again induces *cliques* of neighbor sites (all pairs of sites (x_i, x_j) which are adjacent). Let $\mathcal{C}$ denote the set of all such cliques. Then we define $P(L)$ as:

$$P(L) \propto \prod_{(x_i,x_j)\in\mathcal{C}} e^{-U(i,j)} \tag{6}$$

with $U(i, j) = L_i L_j \cdot c_1 + (1 - L_i L_j) \cdot c_2$. This leads to the overall posterior

$$P(L|D) \propto \prod_i p(v_i|\theta) \cdot \prod_{(x_i,x_j)\in\mathcal{C}} e^{-U(i,j)} \tag{7}$$

maximizing which is again equivalent to maximizing a simpler quality criterion (a detailed derivation is again given in [18]):

$$Q_2(\theta) = Q_1(\theta) + \gamma \sum_{(x_i,x_j)\in\mathcal{C}} L_i L_j \tag{8}$$

where Q_1 is the quality from Equation (4). The free parameter $\gamma > 0$ determines the weight of spatial coherence relative to the goodness-of-fit term Q_1. It depends on c, c_1, and c_2, and is set manually in practice.

3.3 Optimization using RAST

Both criteria Q_1 and Q_2 can be highly non-convex for motion fields in practice such that techniques based on a sparse sampling of the space of possible motions may get caught in local minima. We present an alternative based on a full search of parameter space. Though more time-consuming, it is made feasible using an *adaptive* search strategy. Our approach is called RAST (*Recognition by Adaptive*

Search of Transformation space) [3][4]. It has been applied in the domain of geometric matching before, but is novel to dominant motion estimation in video. RAST is based on a branch-and-bound strategy: starting with the full parameter space, a parameter subset is iteratively chosen and subdivided into two parts by splitting along one parameter. We obtain subsequently finer subsets until finishing with a sufficiently small region corresponding to our estimate $\hat{\theta}$ (the user can define the accuracy of the solution via this stopping criterion). The search is guided into promising regions of parameter space by managing subsets in a priority queue, i.e. for each subset an *upper bound* $\mathcal{U}$ of the quality is computed and used to reinsert the subset into the priority queue.

The key part of the search is the computation of $\mathcal{U}$. For Q_1, the associated bound is $\mathcal{U}_1 = \sum_i u_i$, i.e. for each motion probe we find out (e.g., using interval arithmetic [2]) if it can contribute to *any* global motion in the subset. For Q_2, $\mathcal{U}_2 = \mathcal{U}_1 + \gamma \cdot \sum_{(i,j)\in\mathcal{C}} u_{ij}$ with $u_{ij} = 0$ if $u_i = u_j = 0$ and $u_{ij} = 1$ otherwise. i.e. after computing $\mathcal{U}_1$, an additional linear sweep through the motion probes is required to increment the bound for each pair of adjacent potential background sites.

4 Experiments

The most important capability of our approach is its optimality: the combination of our statistical framework and the RAST optimization guarantees an optimal solution up to any accuracy desired given a state-of-the-art statistical model – a fact that is proven by quantitative experiments on synthetic motion fields, which provide a controlled framework for evaluation with a well-known known ground truth segmentation and ground truth motion. To validate that our model is adequate in practice, we also present results for real-world video data.

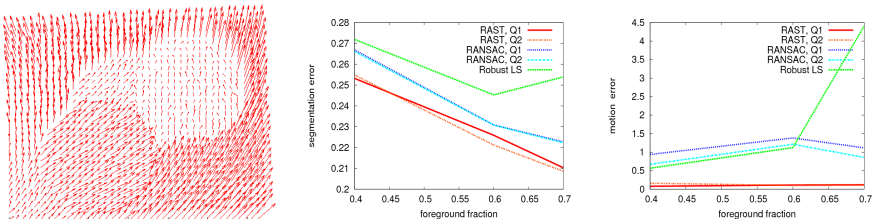
4.1 General Setup

All input motion fields – synthetic or extracted from video – are defined at 16×16 macroblock positions (though our approach is not restricted to this setup). For video streams, motion is estimated using the MPEG-4 video codec XViD[5] [15]. Global motion is parameterized using a similarity transform. The following methods are tested:

1. *Our Framework*: We test our framework for both quality functions Q_1 and Q_2 ($\epsilon = 2.3$, $\gamma = 1$). The 4-dimensional similarity transform space searched by RAST should contain all reasonable motion between adjacent video frames. We choose: $\sigma \in [0.9, 1.1], \alpha \in [-0.1, 0.1], (d_x, d_y) \in [-40, 40]^2$. Search is stopped if the evaluated subset has dimensions smaller than $(0.0002)^2 \times (0.1)^2$. This means, the solution is determined with an accuracy of 0.1 pixels for the translation, 0.0002 *rad* for the rotation, and 0.0002 for the scale.

[4] open source implementation at http://www.iupr.org/~chl/multirast.tar.gz

[5] www.xvid.org



(a) (b) (c)

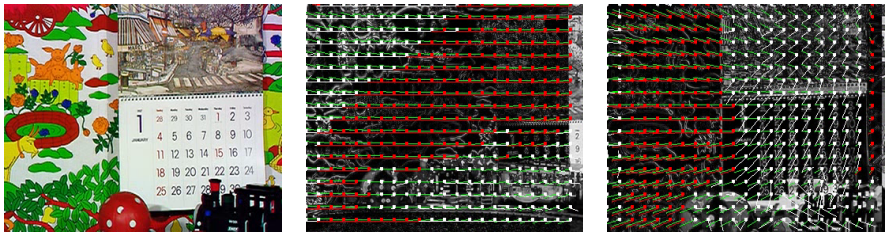
Fig. 1: (a) A synthetic motion field with three blobs each moving in different directions. (b) and (c) Motion estimation results on synthetic blob data. (b) shows the average segmentation error (depending on the fraction of the screen occupied by competitive foreground motion), (b) the squared error of the estimated x translation relative to the ground truth.

2. *Least Squares*: standard least squares regression is equivalent to maximizing a quality function similar to Q_1, but with a pure Gaussian motion vector density instead of a truncated Gaussian one. It is thus expected to perform poorly when competitive foreground motion occurs and serves as a baseline.
3. *Robust Least Squares*: this method alternately computes least squares motion estimates and discards motion samples from D that deviate further than an outlier threshold σ. Our implementation generates a sequence of solutions by decreasing σ according to the schedule $\sigma_{k+1} = 0.95 \cdot \sigma_k$ until $\sigma < 2.3$.
4. *RANSAC*: Random Sample Consensus (RANSAC) [5] is a popular Monte Carlo procedure with excellent robustness to outliers and noise [7, 11]. It is based on an iterated random subsampling of D. The probability of failure decreases with the number of iterations, but never reaches 0, such that optimality is not guaranteed. RANSAC is tested for both Q_1 and Q_2.
5. *XViD Dominant Motion Estimation*: this is the dominant motion estimation component that the XViD codec uses for compression purposes. The implementation is comparable to robust least squares, but with a more greedy outlier rejection strategy.

4.2 Synthetic Flow Fields

In a first experiment, we use synthetic flow fields of blob regions moving in front of a moving background with the purpose of simulating the phenomena of noise and spatial coherence in real-world video frames.

Like the example illustrated in Figure 1(a), all motion fields are derived from a dominant motion and three foreground motions. The background motion is randomly drawn from $[-0.05, 0.05] \times [0.95, 1.05] \times [10, 10]^2$. Also, three blobs are initialized with a random motion from $\{0\} \times \{1\} \times [-16, 16]^2$. All blobs are of the same size such that they – when non-overlapping – occupy a certain fraction $f \in \{0.4, 0.6, 0.7\}$ of the field. Also, isotropic Gaussian noise with standard



(a) (b) (c)

Fig. 2: (a) A frame from the mobile sequence. (b) and (c) Motion segmentation (red vectors belong to the background, white ones to the foreground) and difference between motion-compensated frames for XViD (b) and RAST (c). For XViD, a wrong estimate leads to a poor motion compensation on the upper left part of the frame.

deviation $\sigma \in \{1.0, 1.3, 1.6, 2.0, 2.3\}$ is added to each motion vector, obtaining a total of 1000 motion fields.

Numerical results for all test methods except the XViD codec (which we apply to real-world videos only) and least squares (which performed much worse than all other methods) are given in Figures 1(b) and 1(c). In Figure 1(b), the average segmentation error is plotted against the fraction f occupied by the foreground, reaching from 0.4 to 0.7. Note that some intrinsic segmentation error results from outliers due to noise. The rate of such outliers – and thus the segmentation error – constantly drops with f. Our framework gives lower segmentation error rates than all other methods. The robust least squares method tends to break at high foreground fractions. Between RAST and RANSAC (100 iterations), a difference of about 1 % in segmentation error can be observed.

In Figure 1(c), we plot the average error of the estimated motion (more precisely, for the x-translation parameter) for the noise level $\sigma = 2.0$ against the foreground fraction f. Again, our framework shows the best performance. The average mean squared error remains below 0.2 pixels. Also, it can be observed that Q_1 and Q_2 give a similar performance.

4.3 Test Sequences "Mobile" and "Snooker"

To validate its performance on real-world video data, we first apply our framework to MPEG-4 motion vectors derived from the "mobile and calendar" test sequence[6]. The sequence shows a textured background behind three foreground objects, each moving in a different direction approximately perpendicular to the optical axis. We subsampled the sequence in the temporal domain at 1 fps, obtaining 11 frames 22×18 macroblocks each. One frame is shown in Figure 2 together with motion estimates for XViD and RAST. The motion visualization is layed over a motion-compensated difference image. For the RAST result, the difference is low except for foreground regions. For the XViD result, it can be

[6] http://www.m4if.org/resources.php

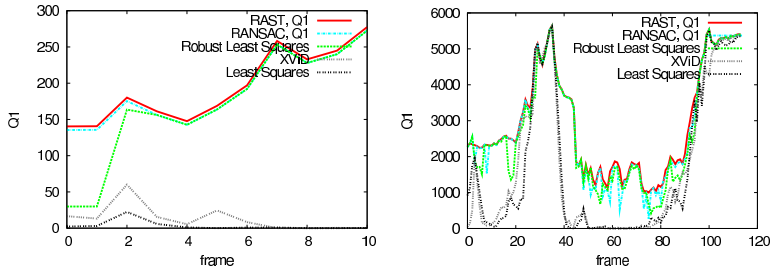


Fig. 3: Motion support results for (a) the mobile test sequence (11 frames) and (b) the snooker test sequence (90 frames).

seen that parts of the background (on the upper left) have been classified as foreground and have thus been poorly compensated for.

Figure 3(a) illustrates the motion support Q_1 for several test methods, plotted over the frames of the mobile test sequence. For RANSAC and RAST, the ML formulation was used.

We also compared the average processing time of RAST for both criteria Q_1 (2.85 sec./frame, 1.6 Ghz Pentium M) and Q_2 (1.07 sec./frame). Interestingly, the spatial prior – though demanding an extra sweep through all motion samples for the evaluation of a subset – leads to a significant speedup (62 %) that can be observed throughout all of our experiments. Obviously, spatial coherence helps to discard bad motion hypotheses early that are scattered over the field, and to guide search into promising regions of transformation space. This insight might be interesting in the geometric matching domain where RAST was developed.

Comparable results can be observed for our second test sequence "snooker" captured from a TV sports broadcast (90 frames), showing a snooker player tracked by a camera with a strong translation. The support Q_1 for the sequence is plotted in Figure 3(b) (for RANSAC, 20 iterations were used). Again, XViD and least squares give relatively poor results. RANSAC and robust least squares perform comparable to our method, but fail occasionally.

For both sequences, the support for our approach serves as an upper bound for the performance of other methods.

4.4 Test Sequence "Foreman"

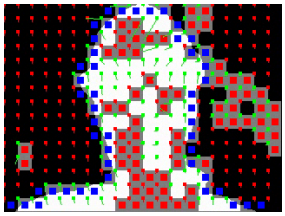
In this experiment, we test the performance of our approach for motion segmentation on a subsampled version of the MPEG-4 test video sequence "foreman" (80 frames) that comes with a ground truth segmentation mask. The sequence shows strong, chaotic camera motion and a highly non-planar background.

Again, we tested several methods, for RAST and RANSAC (100 iterations) including the spatial prior (Q_2). Segmentation results are compared to the ground truth on block basis (mixed blocks showing more than 5 % of both foreground

method	segmentation error rate
RAST	0.24
RANSAC	0.25
Robust LQ	0.26
XViD	0.33
LQ	0.41

(a)

(b)



(c)

Fig. 4: (a) Average segmentation error rates for the foreman sequence. (b) A frame from the foreman sequence, and (c) a typical segmentation result evaluated using MPEG-4 ground truth segmentation masks. Blue blocks are ignored, red blocks are misclassified.

and background pixels are ignored). The resulting error rates are given in Figure 4 (a), a sample segmentation is illustrated in Figure 4 (b) and (c). Our method gives the best results, followed by RANSAC and robust least squares. A high intrinsic error occurs due to two reasons (besides inaccuracies in the motion estimation step): first, the object stands still in some frames and is missed by motion segmentation. Second, the 4D motion model implicitly assumes a planar background surface perpendicular to the optical axis. Since this assumption is heavily violated in the foreman sequence, the optimal motion fit cannot be determined in some frames.

5 Discussion

We have presented a framework for the indirect estimation of a global motion from a given motion field. Our method is based on two alternative probabilistic formulations of the problem: an ML criterion assuming independence of motion samples, and an extension with a spatial coherence prior enforcing piecewise-smooth motion. The optimization of the resulting quality functions is done using RAST, an approach novel to dominant motion estimation in video.

The most important capability of our framework is that our method – in contrast to local search procedures used in the past – guarantees an optimal solution up to any user-defined accuracy. We demonstrate this superior performance on synthetic motion data showing blobs moving in front of a noisy background motion, as well as on several real-world video sequences. Though greedy search procedures may be fast, attractive solutions for online processing, they do not guarantee global optimality. In this context, our framework might provide ground truth for benchmarking global motion estimation in video.

Another novelty we present is the combination of RAST optimization with a spatial prior formulation. In our experiments, we measured a significant speed-up using this extension. Obviously, this approach helps to guide the adaptive search into more promising regions of parameter space – an insight that might be interesting for RAST applications in the area of geometric matching and object recognition.

6 Acknowledgements

This work was supported in part by the Stiftung Rheinland-Pfalz für Innovation, project InViRe (961-386261/791).

References

1. M.J. Black and P. Anandan. The Robust Estimation of Multiple Motions: Parametric and Piecewise-Smooth Flow Fields. *CVIU*, 63(1):75–104, 1996.
2. T.M. Breuel. On the Use of Interval Arithmetic in Geometric Branch-and-Bound Algorithms. *Pattern Recogn. Lett.*, 24(9-10):1375–1384, 2003.
3. T.M. Breuel. Fast Recognition using Adaptive Subdivisions of Transformation Space. In *CVPR 92*, pages 445–51, 1992.
4. D. Cremers, S. Soatto. Motion Competition: A Variational Approach to Piecewise Parametric Motion Segmentation. *Int. J. Comput. Vision*, 62(3):249–265, 2005.
5. M.A. Fischler and R.C. Bolles. Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. *Communications of the ACM*, 24(6):381–395, 1981.
6. D. Geman. Stochastic Model for Boundary Detection. *Image Vision Comput.*, 5(2):61–65, 1987.
7. R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, New York, NY, USA, 2003.
8. M. Irani and P. Anandan. About Direct Methods. In *ICCV '99: Intern. Workshop on Vision Algorithms*, pages 267–277, London, UK, 2000.
9. A. Jepson and M.J. Black. Mixture Models for Optical Flow Computation. In *CVPR 93*, pages 760–761, 1993.
10. G. Kühne. *Motion-based Segmentation and Classification of Video Objects*. PhD thesis, University of Mannheim, 2002.
11. V. Lepetit and P. Fua. Monocular Model-Based 3D Tracking of Rigid Objects: A Survey. *Foundations and Trends in Computer Graphics and Vision*, 1(1), 2005.
12. D. W. Murray and B. F. Buxton. Scene Segmentation from Visual Motion using Global Optimization. *IEEE Trans. Pattern Anal. Mach. Intell.*, 9(2):220–228, 1987.
13. A. Smolic. *Globale Bewegungsbeschreibung und Video Mosaiking unter Verwendung parametrischer 2-D Modelle*. PhD thesis, RWTH Aachen, 2001.
14. C. Tomasi and T. Kanade. Detection and Tracking of Point Features. Technical Report CMU-CS-91-132, CMU, 1991.
15. A.M. Tourapis. Enhanced Predictive Zonal Search for Single and Multiple Frame Motion Estimation. In *Proc. SPIE Conf. Visual Communications and Image Processing*, pages 1069–1079, Lugano, Switzerland, 2002.
16. J.Y.A. Wang and E.H. Adelson. Layered Representation for Motion Analysis. In *CVPR 93*, pages 361–366, 1993.
17. Y. Weiss. Smoothness in Layers: Motion Segmentation using Nonparametric Mixture Estimation. In *CVPR 97*, pages 520–526, 1997.
18. A. Ulges. Motion Interpretation using Adaptive Search of Transformation Space. Technical Report, IUPR Research Group, TU Kaiserslautern, 2007.