# Sparse Motion Segmentation using Propagation of Feature Labels

Pekka Sangi, Jari Hannuksela, Janne Heikkilä and Olli Silvén
*Center for Machine Vision Research, Department of Computer Science and Engineering, University of Oulu, Oulu, Finland*

Keywords: Motion Segmentation, Block Matching, Confidence Analysis.

Abstract: The paper considers the problem of extracting background and foreground motions from image sequences based on the estimated displacements of a small set of image blocks. As a novelty, the uncertainty of local motion estimates is analyzed and exploited in the fitting of parametric object motion models which is done within a competitive framework. Prediction of patch labels is based on the temporal propagation of labeling information from seed points in spatial proximity. Estimates of local displacements are then used to predict the object motions which provide a starting point for iterative refinement. Experiments with both synthesized and real image sequences show the potential of the approach as a tool for tracking based online motion segmentation.



## 1 INTRODUCTION

Detection, segmentation, and tracking of moving objects is a basic task in many applications of computer vision such as visual surveillance and vision-based interfaces. In absence of a priori appearance models, solutions must be based on observed image changes or apparent motions. In the case of a moving camera, one approach is to perform motion segmentation where scene objects are detected based on their motion differences (Tekalp, 2000).

One particular approach to motion-based segmentation is to estimate or track the motion of a set of feature points whose association provides the approximate segmentation of regions of interest and corresponding parametric motions (Wong and Spetsakis, 2004; Fradet et al., 2009). Due to the potential unreliability of local motion estimation, such approaches either use point detectors to find regions with suitable texture, and/or incorporate various mechanisms for detecting or analyzing unreliability (Wills et al., 2003; Wong and Spetsakis, 2004; Kalal et al., 2010; Hannuksela et al., 2011).

When processing is done for long image sequences mechanisms are needed for maintaining the coherence of the motions, segmentations, and appearances of the objects (Tao et al., 2002). One approach here is to use dynamics based filtering such as Kalman filter (Tao et al., 2002). (Tsai et al., 2010) optimize energy functions which model the coherence within and across frames. (Karavasilis et al., 2011) maintain temporal coherence by performing the clustering of feature trajectories. In (Lim et al., 2012), background/foreground segmentations which are based on a regular block grid are linked according to displacements obtained by block matching. (Odobez and Bouthemy, 1995b) propagate dense segmentation information using the parametric motion estimates of the segmented regions.

Various principles have been used to implement motion segmentation algorithms as discussed in (Tekalp, 2000; Zappella et al., 2009), for example. The competitive approach, implemented typically with the Expectation Maximisation (EM) algorithm to find a maximum likelihood solution, has been widely used (see (Karavasilis et al., 2011; Pundlik and Birchfield, 2008; Tekalp, 2000; Wong and Spetsakis, 2004)). In our first contribution, we consider a method based on this approach, and derive a technique where the temporal propagation of feature segmentation information is integrated into competitive refinement. Prediction is based on the segmentation of the previous frame and estimated block displacements. The approach is reminiscent to propagation in (Odobez and Bouthemy, 1995b) and (Lim et al., 2012) who also use motion estimates in some form for temporal propagation; in our case, coarse feature-based segmentation is considered. As a second contribution, we use the results of directional uncertainty analysis of block matching and show experimentally that use of such uncertainty information can improve the performance of online sparse segmentation.


Sangi P., Hannuksela J., Heikkilä J. and Silvén O..
Sparse Motion Segmentation using Propagation of Feature Labels.
DOI: 10.5220/0004281203960401
In *Proceedings of the International Conference on Computer Vision Theory and Applications* (VISAPP-2013), pages 396-401
ISBN: 978-989-8565-48-8

## 2 PROPOSED METHOD

In motion estimation, correspondences for regions observed in the *anchor* frame are sought for in the *target* frame which correspond to temporally earlier and later frames in forward estimation. Based on tracking and assumptions on spatial coherence, the prediction of segmentation can be based on the alternation of temporal and spatial propagation. This idea provides the basis of the proposed method.

### 2.1 Motion Features

Our approach analyzes the observation of interframe motion encoded as so-called *motion features* which are triplets $(\mathbf{p}_n, \mathbf{d}_n, \mathbf{C}_n)$ where $\mathbf{p}_n = [x_n, y_n]^T$ is the location of a block in the anchor frame, $\mathbf{d}_n = [u_n, v_n]^T$ denotes an estimate of its displacement in the target frame, and $\mathbf{C}_n$ is a $2 \times 2$ covariance matrix which measures directional uncertainty related to the displacement estimate, that is, it quantifies the aperture problem associated with the block and its neighborhood.

The image area is divided into $N_{\text{feat}}$ rectangular subregions, and location for a block, $\mathbf{p}_n$, is selected from each region $n$. The minimum eigenvalue of the second moment matrix of local image gradients (2D structure tensor) is used as the basic criterion here. To reduce the amount of computations, this image-based selection technique is complemented with feature tracking which generates points from the motion features of the previous frame pair.

In our experiments, the estimation of the displacements $\mathbf{d}_n$ is based on the evaluation of the sum of squared differences (SSD) or some related measure over the block pixels. The fittings of quadratic polynomials to SSD surface at the minimum are used to obtain an estimate with subpixel accuracy. The match surface is also used as a basis for computing the covariance matrix $\mathbf{C}_n$ as is done in (Nickels and Hutchinson, 2001). For this purpose, we use the gradient based method detailed in (Sangi et al., 2007).

### 2.2 Estimating Parametric Motions

Linear parametric models are used to approximate 2-D motion of background and foreground areas (called *objects* in the following). With such models, the induced displacement $\mathbf{d}$ at image point $\mathbf{p}$ is computed by multiplication $\mathbf{d} = \mathbf{H}[\mathbf{p}]\theta$ where $\mathbf{H}[\mathbf{p}]$ is the mapping matrix, and $\theta$ is the parameter vector.

Weighted least squares (WLS) regression is used to both *predict* and *refine* object motion models. Due to the aperture problem, the estimates of local displacements carry varying amount of information about the local motion. In addition, if the patch is not associated with the object of interest there is no information about the object motion. These notions about informativeness are combined in $2 \times 2$ observation weight matrices $\mathbf{W}_{n,o}^{(i)}$ which are derived from the matrices $\mathbf{C}_n$ and object association weights $w_{n,o}^{(i)} \in (0,1)$ $(o = 1,2)$. Particularly, we use the formulation

$$\mathbf{W}_{n,o}^{(i)} = [w_{n,o}^{(i)}]^a \mathbf{C}_n^{-1} \tag{1}$$

where $a$ is a positive parameter. The superscript $(i)$ refers here to $i$th iteration in refinement; $i = 0$ corresponds to the prediction step.

Let $\mathcal{G}_o^{(i)} = \{(\mathbf{p}_n, \mathbf{d}_n, \mathbf{W}_{n,o}^{(i)})\}_{n=1}^{N_{\text{feat}}}$ be the weighted motion feature set obtained for the object $o$. Then, the associated estimate of $\theta$ is

$$\hat{\theta}_o^{(i)} = (\mathbf{H}^T \mathbf{W}_o^{(i)} \mathbf{H})^{-1} \mathbf{H}^T \mathbf{W}_o^{(i)} \mathbf{z} \tag{2}$$

where $\mathbf{z}$ is a vector composed of feature displacements $\mathbf{d}_n$, $\mathbf{H}$ is a vertical concatenation of matrices $\mathbf{H}_n = \mathbf{H}[\mathbf{p}_n]$, and $\mathbf{W}_o^{(i)}$ is a block diagonal matrix composed of $\mathbf{W}_{n,o}^{(i)}$. Moreover, interpreting $\mathbf{W}_o^{(i)}$ as inverse error covariance matrices we estimate the motion model error covariance as

$$\mathbf{P}_o^{(i)} = (\mathbf{H}^T \mathbf{W}_o^{(i)} \mathbf{H})^{-1} \tag{3}$$

and use it for error propagation in computations.

### 2.3 Association Weights in Prediction

Based on the estimated displacements $\mathbf{d}_n$ of feature points, we can propagate association information from the anchor to the target frame. In addition, the target frame of the previous frame pair is the anchor frame of the current frame pair which provides an approach to propagate association information between frame pairs based on spatial proximity. We expect that if two patches are close to each other then it is likely that they have the same association. This principle is illustrated in Fig. 1.

We formulate this by measuring the proximity of image points with their Euclidean distance. Let $w'_{m,o}$ $(m = 1, \ldots, M)$ be the given probabilistic weights for the association of the seed points $\mathbf{p}'_m$ with the object $o$ $(\sum_o w'_{m,o} = 1)$. The predicted association weight, $w_{n,o}$, for a point $\mathbf{p}_n$ is computed as a weighted average

$$w_{n,o}^{(0)} = \frac{\sum_{m=1}^{M} u(\mathbf{p}_n, \mathbf{p}'_m) w'_{m,o}}{\sum_{m=1}^{M} u(\mathbf{p}_n, \mathbf{p}'_m)} \tag{4}$$

where $u(\cdot)$ is the weighting function. Exponential mapping of the Euclidean distance is used to derive the weights: $u(\mathbf{p}, \mathbf{p}') = \exp\left(r \|\mathbf{p}' - \mathbf{p}\|_2^2\right)$ where $\|\cdot\|_2$ denotes the L2 norm, and $r$ is a scaling parameter.

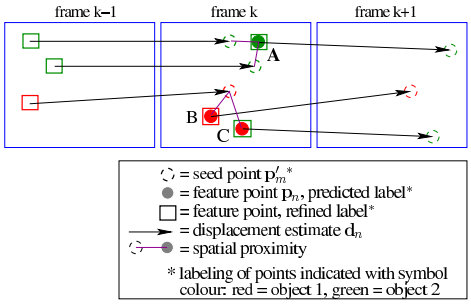



Figure 1: Propagation of feature labeling: motion features of the previous frame pair $(k-1,k)$ provide seed points for frame $k$ which are then used to predict labeling of features A, B, C. Based on seed points, B and C are predicted to be associated with the same object. However, based on the observed motion of these features and refined object motions, the labeling of C changes here.

Using these weights, the prediction of object motions is based on Eq. 2 where small random values are added to the components of $\hat{\theta}_o^{(0)}$ in order to have distinct motion models as a starting point for refinement.

## 2.4 Competitive Refinement

Refinement of predicted motion estimates is based on the competitive paradigm where estimation is performed by iterating two steps, reweighting of data and updating of parametric models. As described above, we use WLS estimation to implement motion model refinement. Bayesian formulation for updating local association weights is used during refinement as follows.

Let the current estimates of object motions and associated covariances be $\theta_o^{(i)}$ and $\mathbf{P}_o^{(i)}$, and let the corresponding association weights be $w_{n,o}^{(i)}$. New association weights, $w_{n,o}^{(i+1)}$, are obtained by weighting old values according to differences between the observed local displacements and displacements induced by object motion models. The estimated errors of motion features and object motion estimates are used to form Gaussian likelihood functions $\mathrm{q}_{n,o}^{(i)}(\mathbf{d})$ whose mean is $\mathbf{H}_n\theta_o^{(i)}$ and covariance $\mathbf{H}_n\mathbf{P}_o^{(i)}\mathbf{H}_n^T$. The Bayes rule is then used to update the association weights according to ($\sum_o w_{n,o}^{(i+1)} = 1$)

$$w_{n,o}^{(i+1)} \propto \mathrm{q}_{n,o}^{(i)}(\mathbf{d}_n) w_{n,o}^{(i)}. \quad (5)$$

It should be noted that if there is no independent foreground motion the motion estimates are close to each other $\mathrm{q}_{n,1}^{(i)}(\mathbf{d}_n) \approx \mathrm{q}_{n,2}^{(i)}(\mathbf{d}_n)$, and the algorithm tends to keep the association weights equal to prediction. This supports maintaining object location information if the independent object motion stops for a moment, and therefore provides a mechanism for handling the temporary stopping problem (Zappella et al., 2009).

## 2.5 Computational Cost

The derived algorithm is summarized in Fig. 2. Iteration of Step 4 may also stop after Step 4(a) has been evaluated. In Step 5, sparse segmentation for the target frame is produced using the current feature points $\mathbf{p}_n$ and their displacements $\mathbf{d}_n$. The segmentation is soft and uses the association weights computed in the last iteration. Only one weight, $w_{n,1}^{(N_{\mathrm{iter}})}$, is saved as the sum of object weights is one. These points are also used as the seed points in processing of the next frame pair.

Considering the computational cost with $M = N_{\mathrm{feat}}$ seed points, Step 1 takes time $\mathrm{O}(N_{\mathrm{feat}}^2)$ whereas the cost of other steps is $\mathrm{O}(N_{\mathrm{feat}})$. However, in evaluation of predicted associations it is necessary to consider only seeds in the $3 \times 3$ neighborhood of subregions, and then then the cost of Step 1 is too $\mathrm{O}(N_{\mathrm{feat}})$.[1]

**Inputs**: a set of motion features, a set of seed points
**Outputs**: object motion estimates, sparse segmentation of the target frame

1. Predict associations $w_{n,o}^{(0)}$ of each motion feature using (4) and given seed points.
2. Compute the weight matrices $\mathbf{W}_{n,o}^{(0)}$ using (1).
3. Make the predictions of object motions, $\hat{\theta}_o^{(0)}$, using (2) with added perturbation. Compute error covariances $\mathbf{P}_o^{(0)}$ using (3).
4. Iteratively refine estimates ($i = 1,\ldots,N_{\mathrm{iter}}$):
   (a) Compute new estimates of association weights, $w_{n,o}^{(i)}$, using (5).
   (b) Compute weight matrices $\mathbf{W}_{n,o}^{(i)}$ using (1).
   (c) Compute estimates of object motions, $\hat{\theta}_o^{(i)}$, using (2). Compute also $\mathbf{P}_o^{(i)}$ if needed.
5. Derive sparse segmentation for the target frame as the set of pairs$(\mathbf{p}_n + \mathbf{d}_n, w_{n,1}^{(N_{\mathrm{iter}})})$.

Figure 2: Derived algorithm for two-motion extraction.

# 3 EXPERIMENTS

Experimental work concentrates on showing the efficacy of algorithmic solutions. To make quantitative comparisons, synthesized image sequences were generated and ground truth information about object mo-

[1]Computation of a Matlab implementation takes 29 ms for motion features (implemented partly in C) and 10 ms for the motion extraction (64 $8 \times 8$ blocks, AMD Opteron 2.4 GHz Linux server).

tion models is used as a basis for quantitative measures of performance. Emphasis in the performance analysis here is on the precision of motion estimates. In practice, we match estimated motions with the ground truth motions, and then compute the root mean square error (RMSE) over the ground truth support regions. It can be shown that the RMSE measure is related to performance in motion content analysis: RMSE should stay below about 1 pixel, and on average it should be at the level of $0.1 - 0.2$ pixels with synthetic sequences.

Synthesized sequences were generated by simulating the background motion for eight different scenes, and pasting moving textured objects of various size to those sequences. In addition, we studied the performance of the motion extraction with real sequences visually by checking the association of features to moving objects, and performing post-segmentation using available motion estimates.

## 3.1 Efficacy of Feature-based Prediction

In the first experiment, performance of the proposed WLS based prediction-refinement method (denoted WLSPR) is evaluated against two variants which use Kalman filtering to implement motion estimation. Both variants perform the propagation of segmentation as described in Sec. 2.3. In the first variant, denoted W-KP-KF, the stages of Kalman filtering are substituted for both prediction (Step 3 in Fig. 2) and estimation (Step 4c) of motions. In the second variant, denoted WLSP-KF, prediction uses WLS and only final estimate is computed using Kalman filtering (Step 4c). Motion dynamics in W-KP-KF is based on an assumption about the constant motion of objects.

The RMSE precision of the estimates, sorted in ascending order, is shown in Fig. 3(a). It can be seen that the proposed approach provides more precise estimates than its variants on average. The median RMSE value with WLSPR is 0.07 pixels whereas it is 0.10 for W-KP-KF and 0.08 for WLSP-KF. In addition, we note that single iteration can already be sufficient for refinement.

Weakness of dynamics-based motion prediction is observed in situations where the direction of motion changes. This fact is illustrated using real video in Fig. 4 where W-KP-KF does not assign any features with the foreground object after motion change occurring in the video whereas feature-based prediction is not so sensitive. In effect, WLSP-KF and WLSPR produce the same soft segmentation as can be seen from Fig. 4 but weighted combination in filtering tends to increase the error in the final object motion estimates, and therefore WLS also in refinement is preferred.

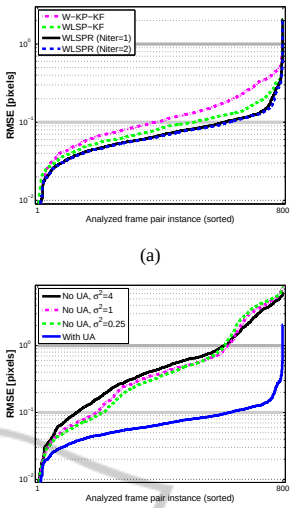



Figure 3: (a) Comparison of the method against Kalman filter configurations. (b) Comparison to estimation which does not exploit uncertainty analysis.

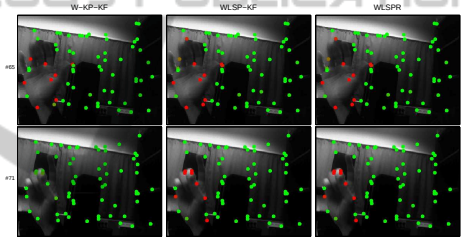



Figure 4: Example of failure of dynamics-based prediction (*Hand* sequence).

## 3.2 Utility of Uncertainty Information

The second experiment checks whether the computation of uncertainty estimates, covariances $\mathbf{C}_n$, done according to the gradient-based analysis (Sangi et al., 2007), is useful in the proposed method. To do this, the weight matrices (see Eq. 1) are set alternatively as $\mathbf{W}_{n,o}^{(i)} = [w_{n,o}]^a \sigma^2 \mathbf{I}$ where $\mathbf{I}$ denotes a $2 \times 2$ identity matrix, and $\sigma^2$ is a constant variance parameter. In Fig. 3(b), the results with synthesized sequences, computed with different choices of $\sigma^2$, are illustrated and compared against the result obtained with WLSPR (2 refinement iterations used in each case). It can be seen that the precision of estimates is improved significantly with uncertainty analysis, and large errors are avoided.

In the experiment with real sequences, the quality of sparse segmentation was evaluated visually by the comparison of feature assignments provided by

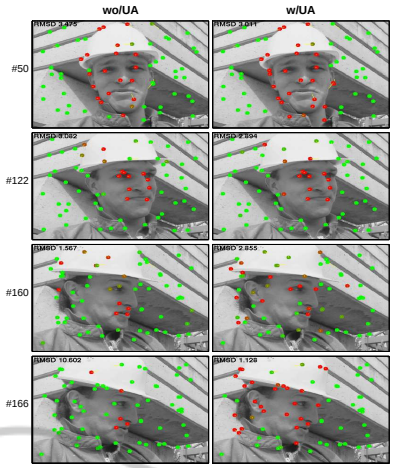
Table 1: Result with real videos based on visual check of quality of feature assignments when motion uncertainty information is used (w/UA) and not used (wo/UA). The 2nd and 3rd column consider absolute quality of assignments whereas 4th and 5th column evaluate their relative quality.

| Sequence [# frames] | # wo/UA good | # w/UA good | # wo/UA better | # w/UA better |
|---|---|---|---|---|
| Hand [201] | 104 | 177 | 12 | 131 |
| Foreman [185] | 148 | 148 | 37 | 60 |
| David [201] | 102 | 156 | 21 | 124 |

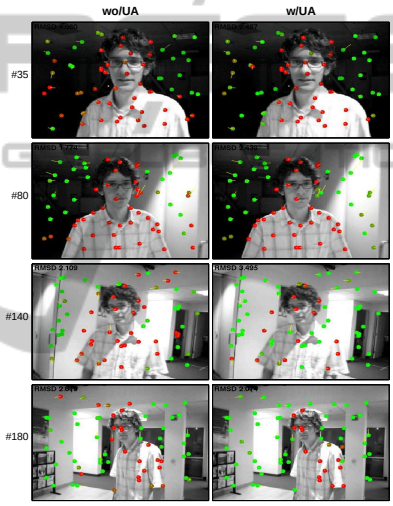
the alternatives ($\sigma^2$ was set to 1.0 when uncertainty information was not used). When the number of mislabellings (64 features used) was observed to be less than three, the segmentation was considered a good one in this experiment. In addition, we compared the segmentation qualities. The figures obtained in this way are given in Table 1, and it can be seen that segmentations obtained using uncertainty information were better on average.

Examples of related segmentations are given in Fig. 5.[2] In the case of the *Foreman* sequence, Fig. 5(a), segmentation of the face area does not typically extend to the area of the helmet and shirt due to the absence of texture and similarity with the background motion, respectively (see Frame 122). In the frames 155-158, there is a moving hand in the view which disturbs segmentation (see Frames 160 and 166). The solution which exploits uncertainty analysis recovers from this situation already in the Frame 161 whereas without uncertainty analysis segmentation is poor until Frame 171.

In the experiment with the *David* sequence, the background tends to get mislabellings more often when uncertainty analysis is not used as illustrated in Fig. 5(b). With uncertainty analysis, the largest errors occur at the beginning of the sequence (Frames 2-7) and when the person turns sideways (Frames 135-160, check Frame 140). However, there are long periods (Frames 64-103, 115-135, 170-200) where the segmentation is very good (see Frame 80).




(a) *Foreman* sequence.




(b) *David* sequence.

Figure 5: Examples of sparse segmentation obtained without and with uncertainty analysis.

## 3.3 Comparison to a Reference Method

We also implemented two-motion extraction based on the dominant motion principle (Tekalp, 2000). A robust multiresolution method for estimating parametric motion models (Odobez and Bouthemy, 1995a) was applied sequntially, first to the whole image, and then to that part of the image which did not support the motion estimate obtained in the first step. In this case, the method uses the whole image area as a basis for estimation which gives significant gain in performance with synthetic sequences observable from Fig. 6. To make the comparison with WLSPR more fair, a fixed grid of blocks was also used as a reduced estimation support, and the same set of blocks was used to provide motion features for WLSPR. The robustness of WLSPR was better with the translational motion model in this experiment (RMSE$>$ 1 pixel in 5 versus 27 out of total of 800 frame pairs). In the case of the four-parameter similarity motion model, used for the results shown in Fig. 6, the robustness of the methods was quite similar.

Finally, masks computed from motion compensated frame differences are compared in Fig. 7 for the *Hand* sequence. Small differences in the masks indi-

[2]See videos at http://www.ee.oulu.fi/research/imag/sms.

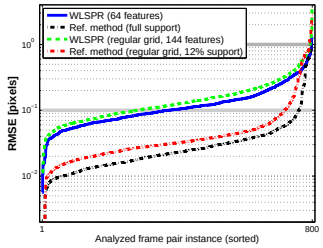



Figure 6: Comparison against the reference method.

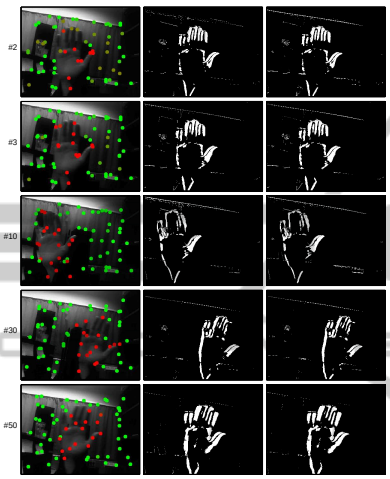



Figure 7: Snapshots from the experiment with the *Hand* sequence, Left: patch/object assignment, Middle: segmentation masks computed from the WLSPR output, Right: masks computed from the reference output.

cate that the object motion estimates provide the same level of performance in post-processing.

## 4 CONCLUSIONS

In this paper, we have proposed an approach to extraction of background and foreground motions where the temporal propagation of probabilistic feature associations is done. This is based on estimated displacements which provides labeled seed points. Spatial proximity of the new feature patches to those points is then used to predict the labelling of features. This propagation technique was integrated with iterative refinement under the WLS estimation framework.

Experiments show that feature-based prediction of motion provides a better starting point for segmentation than the approach using dynamics. In addition, experiments show importance of using directional uncertainty information about the block motion estimates in improving the precision and robustness of the feature-based approach.

## REFERENCES


Fradet, M., Robert, P., and Pérez, P. (2009). Clustering point trajectories with various life-spans. In *Proc. Conf. on Visual Media Production*, pages 7–13.

Hannuksela, J., Barnard, M., Sangi, P., and Heikkilä, J. (2011). Camera-based motion recognition for mobile interaction. *ISRN Signal Processing*, Art. Id 425621.

Kalal, Z., Mikolajczyk, K., and Matas, J. (2010). Forward-backward error: automatic detection of tracking failures. In *Proc. Int. Conf. on Pattern Recognition*, pages 2756–2759.

Karavasilis, V., Blekas, K., and Nikou, C. (2011). Motion segmentation by model-based clustering of incomplete trajectories. In *ECML PKDD*, volume 6912 of *LNAI*, pages 146–161. Springer-Verlag.

Lim, T., Han, B., and Han, J. H. (2012). Modeling and segmentation of floating foreground and background in videos. *Pattern Recognition*, 45(4):1696 – 1706.

Nickels, K. and Hutchinson, S. (2001). Estimating uncertainty in SSD-based feature tracking. *Image and Vision Computing*, 20:47–58.

Odobez, J. and Bouthemy, P. (1995a). Robust multiresolution estimation of parametric motion models. *J. Visual. Comm. Image Repr.*, 6(4):348–365.

Odobez, J.-M. and Bouthemy, P. (1995b). Direct model-based image motion segmentation for dynamic scene analysis. In *Proc. Asian Conf. on Computer Vision*, pages 306–310.

Pundlik, S. J. and Birchfield, S. (2008). Real-time motion segmentation of sparse feature points at any speed. *IEEE Trans. SMC-B*, 38(3):731–742.

Sangi, P., Hannuksela, J., and Heikkilä, J. (2007). Global motion estimation using block matching with uncertainty analysis. In *Proc. European Signal Processing Conf.*, pages 1823–1827.

Tao, H., Sawhney, H. S., and Kumar, R. (2002). Object tracking with Bayesian estimation of dynamic layer representations. *IEEE Trans. PAMI*, 24(1):75–89.

Tekalp, A. M. (2000). Video segmentation. In Bovik, A., editor, *Handbook of Image & Video Processing*, chapter 4.9. Academic Press, San Diego.

Tsai, D., Flagg, M., and Rehg, J. M. (2010). Motion coherent tracking with multi-label MRF optimization. In *Proc. British Machine Vision Conf.*

Wills, J., Agarwal, S., and Belongie, S. (2003). What went where. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, volume 1, pages 37–45.

Wong, K. Y. and Spetsakis, M. E. (2004). Motion segmentation by EM clustering of good features. In *Proc. CVPR Workshop*, pages 166–173.

Zappella, L., Lladó, X., and Salvi, J. (2009). New trends in motion segmentation. In Yin, P.-Y., editor, *Pattern Recognition*, pages 31–46. INTECH.