Robust Feature Point Matching in General Multi-Image Setups

Anita Sellent
TU Braunschweig, Germany
sellent@cg.tu-bs.de

Martin Eisemann
TU Braunschweig, Germany
eisemann@cg.tu-bs.de

Marcus Magnor
TU Braunschweig, Germany
magnor@cg.tu-bs.de

ABSTRACT

We present a robust feature matching approach that considers features from more than two images during matching. Traditionally, corners or feature points are matched between *pairs* of images. Starting from one image, corresponding features are searched in the other image. Yet, often this two-image matching is only a subproblem and actually robust matches over multiple views and/ or images acquired at several instants in time are required. In our feature matching approach we consider the multi-view video data modality and find matches that are consistent in three images. Requiring neither calibrated nor synchronized cameras, we are able to reduce the percentage of wrongly matched features considerably. We evaluate the approach for different feature detectors and their natural descriptors and show an application of our improved matching approach for optical flow calculation on unsynchronized stereo sequences.

Keywords: Keypoint matching, motion estimation, multi-view video.

1 INTRODUCTION

In recent years the increased availability of high quality video cameras together with readily available storage space and fast data transfer has led to a growing interest in stereoscopic or, more general, multiple view video. Although multi-view video data actually is highly redundant, many algorithms in the processing pipeline consider only pairs of images. One important processing step is establishing feature point correspondences that are used, e.g. as low-level starting point for motion estimation [SLW+10, BWSS09, BBM09]. Determination of robust feature points and corresponding feature point descriptions has been an intensely investigated area of research for decades [MTS+05, MS05]. In spite of great advances, wrongly matched correspondences are still commonly encountered. If additional information on the images is provided, e.g. by calibration, synchronization or assumption of constant rigid motion, this information can be used to eliminate wrongly matched correspondences [HZ03]. Unfortunately, in practical applications additional information is not always available as, for instance, multiple cameras are hard to synchronize in an outdoor environment and usually images of independently moving objects are recorded.

The goal of our work is to develop a versatile, robust feature point matching method that is generally applicable, e.g. also in the unconstrained multi-view video setup. Our basic idea is to exploit the redundancy in the data of multi-view video sequences with a common field of view. We use it to establish more reliable correspondences to ensure high-quality matches. Feature points are matched by considering loops of images. We introduce *three image consistent matching* and evaluate it by means of the percentage of wrong matches.

Additionally, we show how a stereo-video optical flow algorithm [SLM10] can benefit from incorporating our robustly matched features. Recent research has shown that optical flow can be improved if ideas from feature matching are included into the approach, [BBM09, XJM10]. In contrast to variational based optical flow algorithms that require an iterative approach to cope with large distances [BBPW04], features can be matched independently from their position in the image and thus deal with arbitrary distances, as long as their descriptor is sufficiently robust to the corresponding changes in perspective or object deformations. For the inclusion of feature matching, optical flow approaches pay careful attention to outier matches as these are able to prevent convergence to the desired motion fields. In this work we show that our robust loop matching strategy which exploits the data modality given for multi-view video is able to improve optical flow estimations without further outlier treatment.

2 RELATED WORK

Usually features are matched between two images from synchronized cameras and spurious matches can be discarded using epipolar geometry [SZ02, HZ03]. Generally, the assumption of global affine motion between

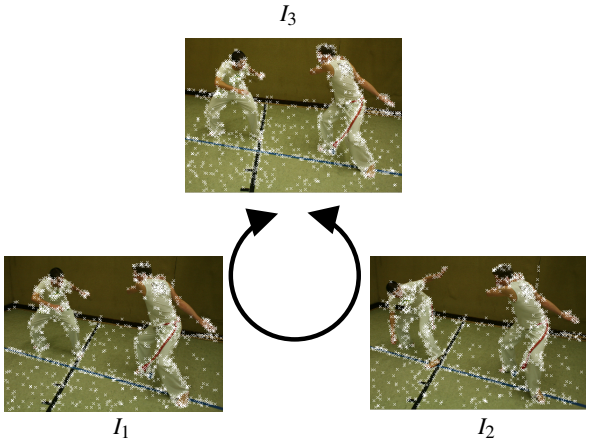


Figure 1: Three images with detected features (SIFT) of a multi-view video sequence: our algorithm accepts three images with some common field of view acquired by one or several unsynchronized and uncalibrated cameras. By requiring consistency of matches on a loop of three images, false matches are eliminated and correspondences between images can be established robustly.

two images can be used to validate matches [BGPS07]. But also game theoretic approaches exploiting local similarity transforms are used to establish reliable matchings between two images [ART10].

If several independent objects move in a monocular sequence, e.g. for person or object tracking [YJS06], feature locations from previous frames can also be used to estimate feature locations in the current frame [Zha94]. Assuming that features have at most one correct match in each frame, disjoint tracks of features over multiple frames can be considered to improve correspondences [VRB03, SS05, SSS06]. Thereby, the tracks provide a regularization of the matches over time, but no feedback for the correctness of the tracking is provided.

For static scenes, the trifocal tensor [TZ97] can be used to consider consistency of the matching between more than two images [BTZ96]. Yao and Cham first verify and add matches between image pairs to satisfy the epipolar constraint, before the matches are extended to image triples and the trifocal tensor is computed [YC07]. In contrast, Zach et al. first determine global, invertible transformations between image pairs before they detect wrong transformations on multi-image loops and discard them [ZKP10], enabling more robust multi-image static 3D reconstruction.

If a dynamic scene is recorded by multiple, unsynchronized cameras Ho and Pong work with high density feature points and use assignments of neighboring pixel in a relaxation labeling framework to obtain consistent matchings [HP96]. In the same setup, Ferrari et al. perform consistency checks on loops of images, but require an additional similarity measure that is different from the measure used to establish preliminary matchings [FTV03].

Mathematically the problem of finding consistent correspondences on three sets of equal, finite cardinality is well studied [Spi00] and approximation algorithms to the NP-hard problem have been proposed by several authors [CS92, BCS94].

In Sect. 3 we will adapt these approximation schemes to sets of different sizes. In Sect. 4 we evaluate the results of this new algorithm. We incorporate our consistent matches into a three image spatio-temporal optical flow algorithm, Sect. 5 and show how consistency of flow and features can improve dense correspondence estimation.

3 THREE IMAGE-FEATURE MATCHING

Let $I_1 : \Omega_1 \rightarrow \mathbb{R}$, $I_2 : \Omega_2 \rightarrow \mathbb{R}$ and $I_3 : \Omega_3 \rightarrow \mathbb{R}$ be three images of a multi-view video sequence that have some common field of view on a dynamic scene. In contrast to previous robust matching methods, we do not require epipolar geometry between images to be applicable, nor do we assume a temporal ordering, i.e. the three images can be acquired by one, two or three unsynchronized cameras, Fig. 1. For each image I_i, $i \in \{1,2,3\}$ a feature detector determines features $f_{i,k}, k \in \{1,\ldots,N_i\}$ with corresponding descriptors $s_{i,k}$. We denote the descriptor distance function with $d(s_{i,k}, s_{j,m})$. In our experiments, Sect. 4, we evaluate the algorithm for several detector/ descriptor variants, so we keep the description general in this section.

Usually, after detection the features are matched between two images at a time. Authors of different descriptors propose slightly different matching methods. To keep the results comparable, we follow the approach of [MS05] and use nearest neighbor matching (NN) for all two-matching steps.

A more elaborate two-matching strategy (NNDR) compares the distance of the nearest neighbor to the distance of the second nearest neighbor and only accepts a match if their ratio is below a threshold [Low04]. We also include this matching strategy into our evaluation.

If more than two images are considered, inconsistencies in the matches such as $(f_{1,k}, f_{2,m})$, $(f_{1,k}, f_{3,n})$ and $(f_{2,m}, f_{3,p})$, $p \neq n$ become obvious. In multi-view video, corresponding feature points are supposed to belong to one single scene point, so inconsistent matches indicate false matches. A straightforward approach to reduce the number of false matches is to filter out any match that is not consistent on a three image circle. To eliminate inconsistent matches already during the assignment we formulate the matching problem in a different way.

In our approach we look for *triples* $(f_{1,k}, f_{2,m}, f_{3,n})$ such that each $f_{i,j}$ is present in at most one triple. To each of the triples we assign a cost $\tilde{d}$ that is the sum of

the distances of all three descriptors $\tilde{d}(s_{1,k}, s_{2,m}, s_{3,n}) = d(s_{1,k}, s_{2,m}) + d(s_{1,k}, s_{3,n}) + d(s_{2,m}, s_{3,n})$, i.e. the distance between each pair of features is considered in the cost function, which therefore is independent of the ordering of the images. In contrast to previous approaches this formulation requires the matches in all images to be similar and thus closes the loop between the images, providing a feedback to the matching and avoiding the drift commonly encountered in considering ordered set of images. If all features were present in all three images this is an instance of the classical three-matching problem with decomposable cost-function, a NP hard problem which can be solved approximately with the following algorithm [CS92]:

i. Match the features in I_1 and I_2, e.g. using the Hungarian algorithm, (see [PS98]).

ii. Merge the sets of features on the basis of the matching in (i.) such that the new cost function between features in I_1 and I_3 is $\hat{d}(s_{1,k}, s_{3,n}) = \tilde{d}(s_{1,k}, s_{2,m}, s_{3,n})$.

iii. Match the features in I_1 and I_3 with the new distance function.

iv. Sum up all distances present in the matching.

v. Interchange the role of I_1, I_2, I_3 and restart at (i.).

vi. Of the three matchings thus obtained, return the one with the smallest sum of distances.

Note that step (ii.) enforces the third feature in the triple to be close both to the feature in I_1 and the feature in I_2. Enforcing this condition simultaneously provides the means to transport the information of the other images to the bilateral matching.

The three-match returned by this algorithm can be proved to lie within a certain distance to the actual best solution and in practice it often turns out to be the best solution [BCS94].

Yet, working with real images, we have to deal with occluded and non-detected features as well as with non-distinctive descriptors. We therefore adjust the above algorithm. In step (i.) we use NN matching or optionally NNDR matching. Additionally we match feature points only if they are mutual nearest neighbors. Thus the processing is independent from the ordering of the images and the feature points. For step (ii.) we remove all features from both images that are not matched in the previous step. We are only interested in feature points that can be matched consistently in three images. As the number of feature points differ in every image and we do not require all feature points to be matched, the sum of all matchings is no longer a reliable quality measure and step (iv.) is skipped. Correspondingly, for step (vi.) we do not return the match with the smallest overall cost, as this is dependent on the number of feature points actually matched. Instead we merge the three matches and only return those triples that are found in all three matching directions. Though this last step might seem rather restrictive, in our setup we opt for less matches with high quality instead of a higher number of matches with more questionable quality. This proceeding is in accordance with considering $\tilde{d}$ in (iii.) that enforces the matches to be mutual neighbors. In summary our algorithm looks as follows:

1. (a) Match the features in I_1 and I_2, using NN matching, optionally with distance check to the second nearest neighbor.
   (b) Match the features in I_2 and I_1, using NN matching, optionally with distance check to the second nearest neighbor.
   (c) Accept only symmetrically matched features.
2. Remove unmatched features in I_1 and merge the remaining features on the basis of the matching in (1.) such that the new cost function between matched features in I_1 and features in I_3 is $\hat{d}(s_{1,k}, s_{3,n}) = \tilde{d}(s_{1,k}, s_{2,m}, s_{3,n})$.
3. (a) Match the features in I_1 and I_3 with the new distance function using NN matching.
   (b) Match the features in I_3 and I_1 with the new distance function using NN matching.
   (c) Accept only symmetrically matched features.
4. Interchange the role of I_1, I_2, I_3 and restart at (1.).
5. Merge the three matchings and return only those matches that are assigned in all three matching directions.

4 EVALUATION OF THREE IMAGE-FEATURE MATCHING

A great number of feature detectors [MTS+05] and feature descriptors [MS05] exist in literature. For a comparison of those we refer the reader to these surveys. The aim of our work is to evaluate the impact of three-image matching and so we chose four widely used detector/ descriptor combinations for our evaluations: SIFT [Low04] and SURF [BETV08] are both scale invariant detectors for blob-like structures and with their natural descriptors also invariant to rotation and changes in illumination. We also evaluate our matching algorithm on Harris-corners [HS88] and the more recent accelerated corner detector FAST [RD06] and combine both with the normalized cross correlation (*NCC*) on a 9×9 window. We transform the normalized cross-correlation to a cost function via $d(s_{i,k}, s_{j,m}) = 1 - NCC(f_{i,k}, f_{j,m})$ to obtain a descriptor distance as used in Sect. 3. Using rather advanced and robust detectors as well as rather low level detectors we

		SIFT NN		SURF NN		Harris NN		FAST NN		SIFT NNDR	
		# M	%WM	# M	%WM	# M	%WM	# M	%WM	# M	%WM
art	2IM	1444	53.39	616	64.45	93	49.46	474	45.57	674	10.53
	3IM	603	11.28	177	20.90	44	13.64	220	13.64	506	2.57
books	2IM	1786	15.58	713	38.85	364	21.98	914	27.02	1506	2.52
	3IM	1373	2.26	318	8.81	200	9.00	517	8.70	1315	0.84
dolls	2IM	2206	23.75	809	35.60	134	18.66	812	19.33	1583	2.21
	3IM	1545	4.27	434	7.60	102	2.94	528	4.17	1367	1.02
laundry	2IM	1112	49.64	675	68.89	158	80.38	420	55.58	627	19.94
	3IM	550	15.82	193	28.50	32	40.63	174	17.24	457	7.66
moebius	2IM	1634	24.24	475	38.95	77	20.78	317	35.65	1211	4.54
	3IM	115	5.02	254	14.96	50	4.00	160	6.88	1011	2.47
reindeer	2IM	943	27.78	428	43.69	49	20.78	290	33.79	683	6.88
	3IM	664	7.08	200	14.50	37	8.11	143	11.89	578	2.77
waving	2IM	4345	11.12	1314	24.20	196	26.53	353	19.97	3804	1.26
	3IM	3995	4.76	1069	12.16	156	19.23	135	9.43	3720	0.70
stonemill	2IM	628	34.71	251	62.55	225	49.78	763	49.15	366	2.73
	3IM	427	13.11	114	35.96	133	27.82	452	22.79	324	0.62
RubberW.	2IM	2077	3.85	236	16.53	48	0.00	255	6.67	1975	0.56
	3IM	1585	0.32	107	5.61	25	0.00	153	1.31	1510	0.20
Hydr.	2IM	1111	16.56	432	20.88	176	25.57	576	22.74	853	1.52
	3IM	254	2.76	56	8.93	20	15.00	70	8.57	136	0.74
wall	2IM	7776	25.44	2365	49.26	1693	28.53	6733	33.71	5327	0.56
	3IM	5363	2.50	686	5.10	906	1.21	2892	1.87	4714	0.19
graffiti	2IM	2057	62.52	1385	77.98	265	90.68	822	91.12	689	25.83
	3IM	626	11.50	140	33.57	8	87.50	39	78.95	338	4.14

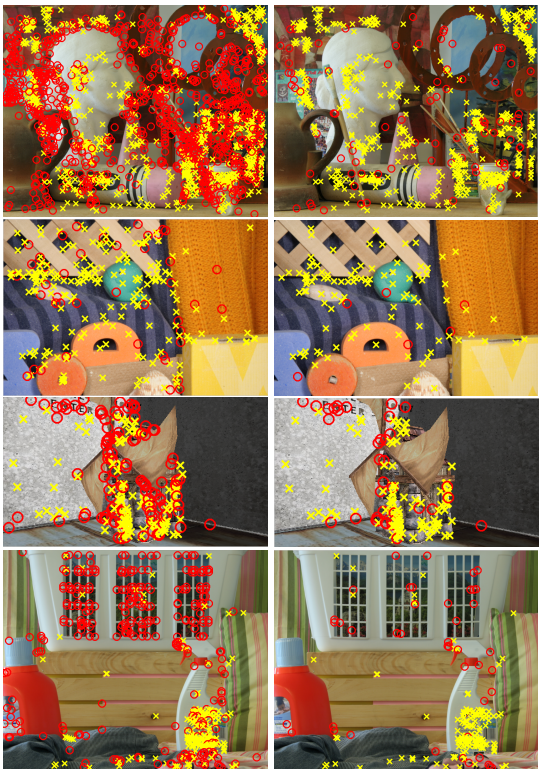
Table 1: As three image matches (3IM) have to satisfy stricter requirements than two image matches (2IM), the total number of matches is reduced while the quality of the matching is increased as the percentage of wrong matches (% WM) is considerably decreased no matter which of the feature detectors (SIFT, SURF, Harris or FAST) or matching strategy (nearest neighbor(NN) or nearest neighbor with threshold on the distance ratio (NNDR)) is used.

want to evaluate our matching scheme independently from the detector used.

For reason of comparison, in our experiments we apply nearest neighbor (NN) matching in all cases [MS05]. Additionally we apply the more advanced NNDR matching that was proposed for SIFT-features, using a threshold of 0.8 on the distance ratio [Low04]. We apply the thresholding step accordingly in the matching step (1.), but found it to have no impact in the matching step (3.) as the combined matching already is sufficiently distinguishing. We therefore do not apply the distance check in (3.).

Using a naïve MATLAB implementation on a 2.66GHz processor, three image consistent matching of 975 FAST features with 81 dimensional descriptors in I_1, 944 features in I_2 and 860 features in I_3 for the *art* scene requires 736ms. With the same setup, independent two-matchings between I_1 and I_2, I_1 and I_3 and I_2 and I_3 last together 126ms.

In our experiments we determine the number of matches and the percentage of matches outside a 5 pixel circle around the ground-truth location in different scenes. The scenes *art*, *books*, *dolls*, *laundry*, *moebius* and *reindeer* are rectified multiple view images of a static scene with known disparity [SP07]. The scenes *waving* [SLM10] and *stonemill* [LLM10] are synthetic, unsynchronized stereo sequences of a moving scene with known ground-truth correspondence fields. The scenes *RubberWhale* and *Hydrangea* are the only monocular sequences of more than two



(a) (b)

Figure 2: The two image-based matching approach (a) results in more outliers (red circles) and a lower relative amount of inliers (yellow crosses) than our three image based-matching (b). From top to bottom: scene *art* with SIFT features, *RubberWhale* with SURF features, *stonemill* with Harris corners, *laundry* with FAST features, all using nearest neighbor matching.

images with independently moving objects and known ground-truth motion from the Middlebury optical flow data set [BSL+07]. In contrast, the scenes *wall* and *graffiti* describe a viewpoint change for a static, mostly planar scene [MS05]. The number of matches and percentage of outliers are shown in Tab. 1, some examples are given in Fig. 2. As expected the number of matches is reduced with our stricter three-matching strategy. But at the same time the percentage of outliers among the assigned matches is also considerably reduced.

We also apply our algorithm to the real multi-video recordings scenes *market*, 421×452 pixel, and *capoeira*, 817×578 pixel, which are recorded using unsynchronized, uncalibrated cameras with automatic gain, while in the scene *outside*, 270×480 pixel, cameras are additionally hand-held. The algorithm is performed on the entire images with all features points found, but for visibility reasons, Fig. 3 shows the results only for 100 randomly selected SIFT-features: matched features are marked with a white x and connected via a yellow line to the location of the corresponding

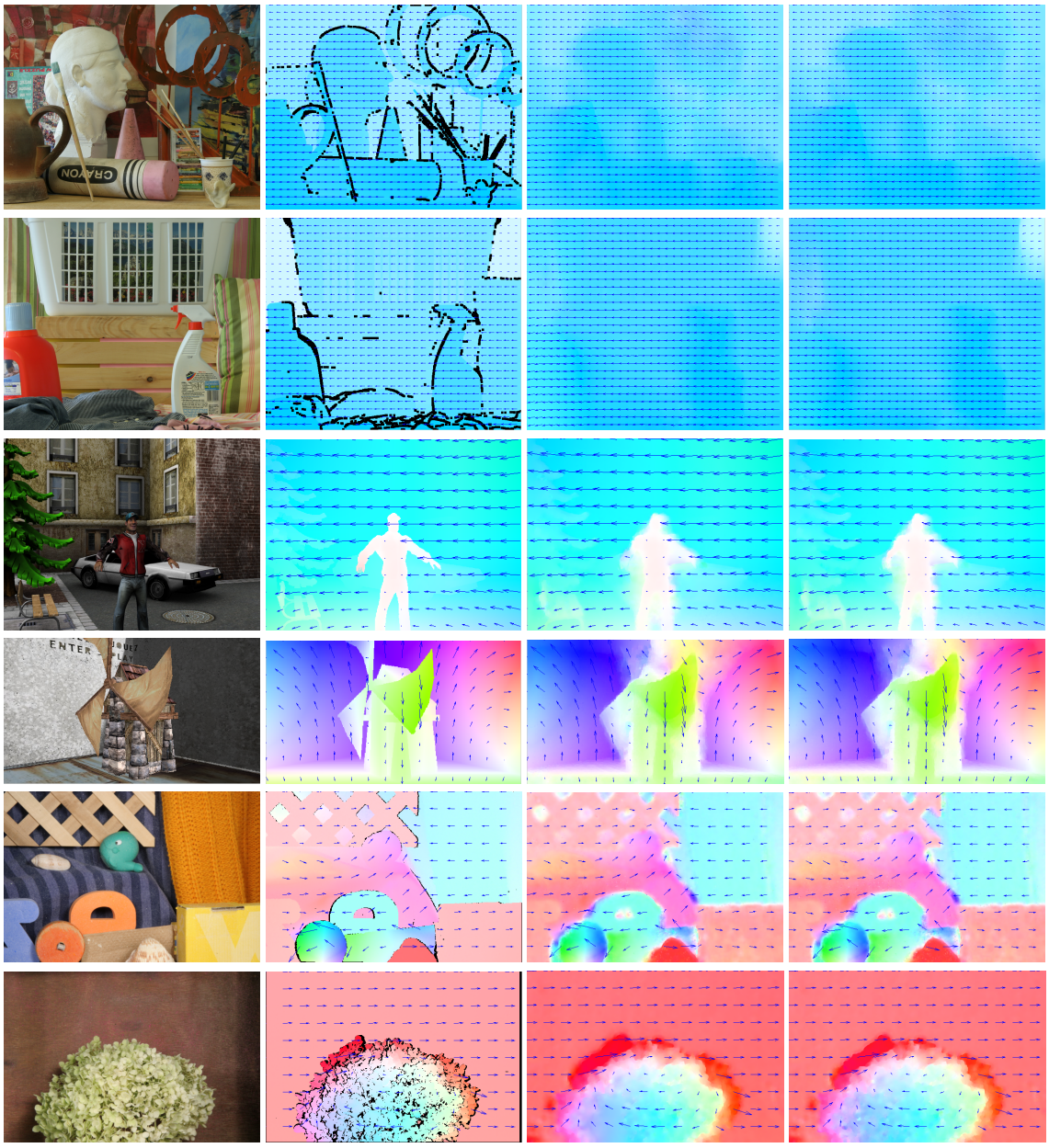
(a) (b) (c)

Figure 3: For three real world scenes *market*, *capoeira* and *outside* (a) we compare different matching strategies. Two-image matches (b) provides a larger number of matches but many outliers among them. Three-image matches (c) reduce the number of outliers considerably. For better visibility here 100 features are randomly selected and connected with the location of their matched features by a yellow line if such a feature is found.

feature. As features are only matched if they are likely correspondences in three images, the three matching algorithm obviously decreases the number of matches as compared to the algorithm that matches features based on two images. But our algorithm renounces to match many inconsistent features so that the percentage of outliers is greatly decreased. As we will show in the subsequent sections, this reduction of the relative amount of outliers allows matching based algorithms to start off much better.

5 APPLICATION TO STEREO-VIDEO CONSISTENT OPTICAL FLOW

Recent optical flow algorithms started to include feature matches into the dense correspondence estimation to faithfully detect large motion also of small objects. More specifically, Xu et al. consider motion vectors of matched features to possibly assign them to pixels all over the image [XJM10], whereas Brox et al. [BBM09] include matched regions as prior into their optical flow algorithm. We adopt the latter idea here and include matched features into the state-of-the-art optical flow for stereo sequences [SLM10]. This optical flow approach is derived from an optical flow algorithm [WTP+09] classified on the Middlebury benchmark [BSL+07]. It considers symmetry and consistency on a three image loop and therefore provides a suitable mean to evaluate the three image based matching. While in the approach of Brox et al. [WTP+09]



(a) (b) (c) (d)

Figure 4: For the scenes *art*, *laundry*, *waving*, *stonemill*, *Rubber Whale* and *Hydrangea* (a) dense ground-truth motion fields are given (b). Compared to the motion fields of the loop-consistent TV-L^2 algorithm of [SLM10], (c) the inclusion of our three-image match as prior results in motion fields with better motion detail (d).

several matches are considered to make sure that the correct correspondence is among them, we incorporate our matched features in their one-to-one fashion. Adopting the notation of $w^r_{i,j}$ for the current estimate of the motion field between image I_i and I_j we simply replace the point-wise energy E_q in [SLM10] with

$$E_f = E_q + \delta_f \|W_{i,j} - w^r_{i,j} - dw_{i,j}\|_2^2 \qquad (1)$$

where for matches $(f_{i,k}, f_{j,n}, f_{h,m})$ and $[f_{i,k}]$ the nearest integer position to the feature location

$$W_{i,j} : \Omega_i \to \mathbb{R}^2, \quad W_{i,j}(x) = \begin{cases} f_{j,n} - f_{i,k} & \text{if } x = [f_{i,k}] \\ 0 & \text{else} \end{cases} \qquad (2)$$

is a function that describes the matching of the features, $\mu, c > 0$ constants and

$$\delta_f(x) = \mu \begin{cases} 1 - \arctan \frac{\tilde{d}(s_{i,k}, s_{j,n})}{c2\pi} & \text{if } x = [f_{i,k}] \\ 0 & \text{else} \end{cases} \qquad (3)$$

a function that assigns values depending on the matching costs or 0 to each point in Ω_i. This new energy is still a quadratic function in the update $dw_{i,j}$, so the updating scheme of [SLM10] is maintained. Note that for all experiments we fix $\mu = 10^3$ and $c = \frac{1}{5}$

To speed up calculations and assist the determination of large flows, loop consistent flow estimation is performed on a factor 0.5 image pyramid. Similar to [BBM09] we down-sample the prior $W_{i,j}$ by considering the 2×2 pixels that are represented by one single pixel in the next coarser level. From the four pixels in the finer level we only pass on to the next coarser level half the motion and the weight of the pixel with the highest weight $\delta_f(x)$. Thus, if no other matches are found in the vicinity, the original match is propagated to the next coarser level or else the match with the smallest cost is used. Having thus established a matching-based prior on all levels of a scale pyramid, we initialize the dense flows on the coarsest level with zero and perform 10 iterations of the updating scheme before proceeding to the next finer level. We use the upscaled flow field from the previous level as initialization on the finer level and thus proceed till the original resolution is reached.

5.1 Evaluation

To evaluate the impact of three image-consistent matching on optical flow estimation, we use all the data sets with known ground-truth motion from Sect. 4 except for the scenes *graffiti* and *wall* which only contain camera motion around a planar scene and are therefore of no interest for dense motion field estimation. We measure the average angular error (AAE) and average endpoint error (AEE) [BSL$^+$07] between the computed and the ground-truth displacement fields. For comparison, we also calculate flow fields with a two-image TV-L^2 approach [SLM10] incorporating standard two image-feature matching as prior and the three image-loop consistent optical flow algorithm [SLM10] without prior. As SURF features provide the best cover of our test scenes with feature points, we here only show the results obtained with SURF. Flow fields incorporating priors obtained with other descriptors behave qualitatively in the same way:

If only two image matches and forward flow are considered, wrong matches have a strong impact and lead to results with high error, Tab. 2. In [SLM10] Sellent et al. show that loop consistent flow improves the results of the TV-L^2 approach. Incorporating feature points that are likewise consistent on three images is able to further improve the results. An improvement is also visible in the flow field, Fig. 4, as small structures such as e.g. the hand in the *waving* scene are better preserved than without the prior matches.

6 CONCLUSIONS AND FUTURE WORK

In our article we show that even in the absence of camera calibration and synchronization, feature points can be matched more robustly if three images are considered simultaneously. By requiring that features are consistent in three images, the quality of the matching improves as the percentage of wrong matches is considerably reduced.

We also combine three-image matching with three image-loop consistent optical flow estimation and obtain dense flow fields that have a smaller error and better preserved motion details than either the loop-consistent flow or basic flow with non-robustly matched features.

In this work we extend the traditional two image approach to three images and obtain more robust results. Future work in this direction compromises to evaluate whether this trend can be continued if four or more images are used and whether there is an optimal number of images to be used.

ACKNOWLEDGEMENTS

This work has been funded by the German Science Foundation, DFG MA2555/4-2.

REFERENCES

[ART10] A. Albarelli, E. Rodolà, and A. Torsello. Robust game-theoretic inlier selection for bundle adjustment. In *Proc. of the International Symposium on 3D Data Processing, Visualization and Transmission*, pages 1–8, Paris, France, May 2010.

[BBM09] T. Brox, C. Bregler, and J. Malik. Large displacement optical flow. In *Proc. of the Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 41–48. IEEE, 2009.

[BBPW04] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert. High accuracy optical flow estimation based on a theory for warping. In *Proc. of the European Conference of Computer Vision (ECCV)*, pages 25–36, 2004.

[BCS94] H. Bandelt, Y. Crama, and F. Spieksma. Approximation algorithms for multi-dimensional assignment problems with decomposable costs. *Discrete Applied Mathematics*, 49(1-3):25–50, 1994.

[BETV08] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool. Speeded-up robust features (SURF). *Computer Vision and Image Understanding*, 110(3):346–359, 2008.

[BGPS07] S. Battiato, G. Gallo, G. Puglisi, and S. Scellato. SIFT features tracking for video stabilization. In *Proc. of the International Conference on Image Analysis and Processing*, pages 825–830, 2007.

[BSL$^+$07] S. Baker, D. Scharstein, JP Lewis, S. Roth, M.J. Black, and R. Szeliski. A database and evaluation methodology for optical flow. In *Proc. ICCV*, pages 1–8. IEEE, 2007.

[BTZ96] P. Beardsley, P. Torr, and A. Zisserman. 3D model acquisition from extended image sequences. In *Proc. of the ECCV*, volume 2, pages 683–695. Springer, 1996.

[BWSS09] X. Bai, J. Wang, D. Simons, and Guillermo Sapiro. Video snapcut: robust video object cutout using localized classifiers. *ACM Trans. Graph.*, 28(3):1–11, 2009.

	TV-L^2		TV-L^2 & 2IM		[SLM10]		[SLM10] & 3IM	
	AAE	AEE	AAE	AEE	AAE	AEE	AAE	AEE
art	1.68	10.62	49.45	84.82	1.32	9.34	1.09	8.70
books	11.23	14.60	31.99	55.37	2.67	6.43	1.62	4.85
dolls	1.93	5.81	32.86	67.66	0.53	2.85	0.51	2.27
laundry	7.83	14.16	40.38	58.08	1.27	9.20	1.03	8.39
moebius	0.96	3.67	16.85	23.77	0.87	3.61	0.87	3.13
reindeer	18.89	25.91	18.70	30.50	2.22	16.35	1.02	8.55
waving	2.95	1.03	26.96	31.39	2.74	0.97	2.58	0.92
stonemill	16.72	4.53	48.59	23.16	11.29	3.81	9.73	3.52
RubberW	6.60	0.20	15.07	5.72	6.46	0.20	6.34	0.20
Hydr.	2.98	0.27	9.28	4.39	2.79	0.25	2.77	0.24

Table 2: Including 2-image SURF matching priors (2IM) into TV-L^2 flow significantly increases average angular (AAE) and average endpoint error (AEE) in comparison to the basic TV-L^2 approach. Under consideration of consistency on a loop of three images, inclusion of 3-image SURF matching priors (3IM) decreases the AAE and AEE of the loop consistent TV-L^2 approach [SLM10].

[CS92] Y. Crama and F. C. R. Spieksma. Approximation algorithms for three-dimensional assignment problems with triangle inequalities. *European Journal of Operational Research*, 60(3):273–279, 1992.

[FTV03] V. Ferrari, T. Tuytelaars, and L. Van Gool. Wide-baseline multiple-view correspondences. In *Proc. of the CVRP*, volume 1, pages 718–725, June 2003.

[HP96] A.Y.K. Ho and T.C. Pong. Cooperative fusion of stereo and motion. *Pattern Recognition*, 29(1):121–130, 1996.

[HS88] C. Harris and M. Stephens. A combined corner and edge detector. In *Proc. of the Alvey Vision Conference*, volume 15, pages 147–151, 1988.

[HZ03] R. Hartley and A. Zisserman. *Multiple view geometry in computer vision*. Cambridge University Press, 2003.

[LLM10] C. Linz, C. Lipski, and M. Magnor. Multi-image interpolation based on graph-cuts and symmetric optic flow. In *Proc. of the International Workshop on Vision, Modeling and Visualization*, page to appear, Siegen, Germany, November 2010. Eurographics, Eurographics.

[Low04] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 2(60):91–110, 2004.

[MS05] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE T-PAMI*, 27(10):1615–1630, 2005.

[MTS+05] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L.V. Gool. A comparison of affine region detectors. *Intern. Journal of Computer Vision*, 65(1):43–72, 2005.

[PS98] C.H. Papadimitriou and K. Steiglitz. *Combinatorial optimization: algorithms and complexity*. Dover Publications, Mineola, New York, USA, 1998.

[RD06] E. Rosten and T. Drummond. Machine learning for high-speed corner detection. In *ECCV*, pages 430–443. Springer, May 2006.

[SLM10] A. Sellent, C. Linz, and M. Magnor. Consistent optical flow for stereo video. In *Proc. ICIP*, Sept. 2010.

[SLW+10] T. Stich, C. Linz, C. Wallraven, D. Cunningham, and M. Magnor. Perception-motivated interpolation of image sequences. *ACM Transactions on Applied Perception*, pages 1–28, 2010.

[SP07] D. Scharstein and C. Pal. Learning conditional random fields for stereo. In *Proc. of the CVPR*, pages 1–8. IEEE Computer Society, June 2007.

[Spi00] F.C.R. Spieksma. Multi index assignment problems: complexity, approximation, applications. *Nonlinear Assignment Problems, Algorithms and Applications*, pages 1–12, 2000.

[SS05] K. Shafique and M. Shah. A noniterative greedy algorithm for multiframe point correspondence. *IEEE T-PAMI*, pages 51–65, 2005.

[SSS06] N. Snavely, S.M. Seitz, and R. Szeliski. Photo tourism: exploring photo collections in 3D. *ACM Transactions on Graphics*, 25:835–846, July 2006.

[SZ02] F. Schaffalitzky and A. Zisserman. Multi-view matching for unordered image sets, or "How do I organize my holiday snaps?". In *Proc. of the ECCV*, volume 1, pages 414–431. Springer, May 2002.

[TZ97] P.H.S. Torr and A. Zisserman. Robust parameterization and computation of the trifocal tensor. *Image and Vision Computing*, 15(8):591–605, 1997.

[VRB03] C.J. Veenman, MJT Reinders, and E. Backer. Establishing motion correspondence using extended temporal scope. *Artificial Intelligence*, 145(1-2):227–243, 2003.

[WTP+09] M. Werlberger, W. Trobin, T. Pock, A. Wedel, D. Cremers, and H. Bischof. Anisotropic Huber-L1 optical flow. In *Proc. BMVC*, London, UK, Sept. 2009.

[XJM10] L. Xu, J. Jia, and Y. Matsushita. Motion detail preserving optical flow estimation. In *CVPR*, San Francisco, 2010. IEEE Computer Society.

[YC07] J. Yao and W.K. Cham. Robust multi-view feature matching from multiple unordered views. *Pattern Recognition*, 40(11):3081–3099, 2007.

[YJS06] A. Yilmaz, O. Javed, and M. Shah. Object tracking: A survey. *Computing Surveys*, 38(4):13, 2006.

[Zha94] Z. Zhang. Token tracking in a cluttered scene. *Image and Vision Computing*, 12(2):110–120, 1994.

[ZKP10] C. Zach, M. Klopschitz, and M. Pollefeys. Disambiguating visual relations using loop constraints. In *Proc. of the CVPR*, pages 1–9. IEEE, June 2010.