LOCAL MULTI-MODAL IMAGE MATCHING BASED ON SELF-SIMILARITY

C. Bodensteiner, W. Huebner, K. Juengling, J. Mueller, M. Arens

Fraunhofer Institute of Optronics, System Technologies and Image Exploitation

ABSTRACT

A fundamental problem in computer vision is the precise determination of correspondences between pairs of images. Many methods have been proposed which work very well for image data from one modality. However, with the wide availability of sensor systems with different spectral sensitivities there is growing demand to automatically fuse the information from multiple sensor types.

We focus on the problem of finding point and local region correspondences in an inter-modality imaging setup. We use a Generalized Hough Transform to determine small regions with a similar geometric relationship of local image features to robustly identify correct matches. We additionally optimize region correspondences by a fast non-linear optimization of a self-similarity distance measure. This measure outperforms standard multi-modal registration approaches like mutual information or correlation ratio in case of local image regions. The method is evaluated on Visible/Infrared (IR) and Visible/Light Detection and Ranging (LiDAR) intensity image data pairs and shows very promising results. Potential applications are numerous and include for instance multi-spectral camera calibration, multi-spectral texturing of 3D-models, multi-spectral segmentation or multi-spectral super-resolution.

Index Terms— Image matching, Feature extraction, Multi-Modal Image Registration, Self-Similarity Distance, Local Multi-Modal Correspondence

1. INTRODUCTION

A fundamental task in remote sensing and computer vision is the precise determination of correspondences between pairs of images. The involved registration task is mostly formulated as the determination of a geometric transformation which maps corresponding features onto each other. The optimal transformation is usually found by minimizing a proper distance measure between the image pairs. Registration approaches can be divided into either global or local methods depending on the spatial support of the distance measure. The registration problem becomes very challenging if the underlying image data is multi-modal, e.g., the data stems from sensors with different spectral sensitivities. Local feature based methods only work well if the image data stems from the same image modality. In this work we propose a local method which employs geometric constraints between image features to robustly determine multi-modal correspondences. Local methods are advantageous when significant changes of the underlying scenery hamper global methods. Besides, they allow for an effective parallelization, which is very favorable with respect to modern multi-core architectures.

1.0.1. Previous Work and Contribution

Many robust feature-based correspondence methods have been proposed in literature [1, 2]. Most local approaches match common image features based on gradient information between the template I_T and reference image I_R. However, the problem of finding accurate local feature correspondences across different image modalities is known to be difficult [3]. Successful multi-modal registration applications include mostly medical applications, e.g., the fusion of MR/CT or CT/PET images based on maximization of the globally computed information theoretic distance measure MI [4]. Recent work on local multi-modal (Visible/IR) registration and stereo algorithms mostly rely on MI [3, 5] as well. However, newly proposed global intensity-based similarity measures like residual complexity now also deal with complex spatially-varying intensity fluctuations [6] and seem very promising for our applications. However, to the best of our knowledge there is no literature about the accurate determination of local multi-modal correspondences based on self-similarity. We adapt the approach of Shechtman and Irani [7] who proposed self-similarity descriptors for sketch based object and video detection and extend it with ideas from the work of Leibe [8]. We additionally propose to optimize a densely computed self-similarity distance to accurately align small image regions where standard methods like MI or correlation ratio have major difficulties.

The outline of the paper is as follows: first we describe the key parts of the approach and discuss specific details which enable a robust local correspondence search in the multi-modal case. Then we evaluate the method on different image scenarios with a focus on IR/Visible spectrum image pairs. In the end, we discuss the results and give further research directions.

2. METHOD

To enable a local feature based approach for multi-modal images we utilize the concept of simultaneously matching local features inside small image regions with a similar geometric layout. We employ a Generalized Hough Transform similar to the one used in [8] in combination with standard local features like SIFT [9], SURF [10] and recently proposed self-similarity descriptors [7]. Based on these initial local feature correspondences we robustly estimate (RANSAC [11]) a local affine transformation T_a for a corresponding image patch (60x60px). This transformation serves as an starting point for an intensity based multi-modal image registration to accurately align these image patches. The refinement step is intended for applications with very high accuracy requirements e.g., multi-modal camera calibration, structure from motion, or multi-modal super-resolution applications. The method can be decomposed in three main parts:

- Selection of Local Image Patch Regions
- Determination of Coarse Patch Correspondences
- Local Intensity Based Fine Registration

A graphical overview of the method is given in Fig. 1. The following subsections provide a detailed description of the main components:

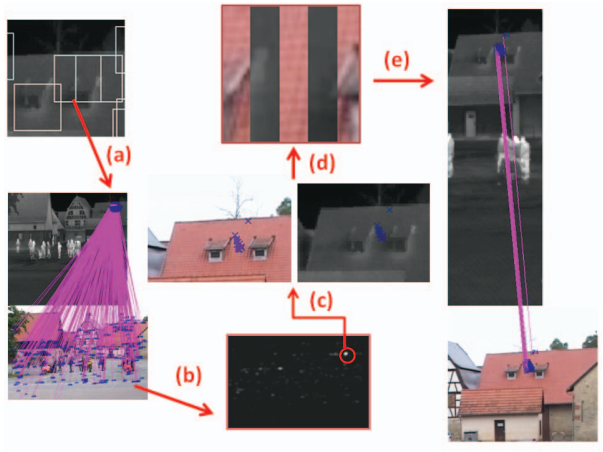
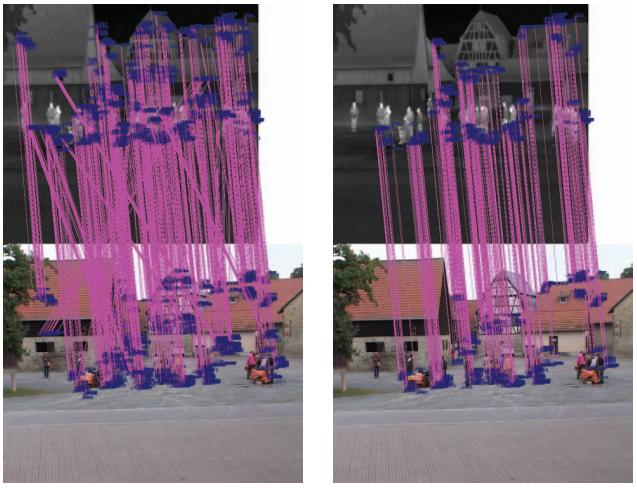


Fig. 1. Algorithm Overview: (a) extracted local image regions, (b) feature matching by searching geometrically consistent feature matches, (c) backprojection of best hypothesis supporting feature matches, (d) intensity based local image region alignment, (e) final point correspondences for one local image region.



(a) IR/Vis (b) IR/Vis

Fig. 2. Self-Similarity matches before (a) and after filtering with the fundamental matrix constraint (RANSAC, 2.25px) (b) for Visible/IR image pairs.

2.1. Selection of Local Image Patch Regions

The selection of local image regions serves as a starting point for the multi-modal correspondence search. Each small region defines a local coordinate frame R_i, where the geometric layout of contained features is determined. Our experiments show that it is favorable to use image regions with strongly distinct features in order to increase the number of correct region matches. To this end we use a simple heuristic based on Harris corners which serve as the origin O_i of R_i and constant region sizes (60x60px).

2.2. Determination of Coarse Patch Correspondences

We extract local features over different scales and use standard descriptors for the initial correspondence search. We also employ a technique called soft-matching for local feature matching which uses the k (e.g., 2-4) nearest neighbors in descriptor space as potential matches. Then we use a Generalized Hough Transformation similar to the implicit shape model (ISM) [8] to filter the correspondences with respect to a consistent geometric layout expressed in the coordinate frame R_i.

Due to fundamentally different object appearances, many initial local feature matches are not correct and would lead to an enormous amount of wrong point correspondences (see Fig. 1 left). Each local feature casts a vote for a corresponding region origin according to the geometric layout in its reference coordinate system [8]. Under the assumption that wrong correspondences spread their votes randomly, we determine the corresponding image region center with a simple maximum search in the voting space. The final correspondences are feature matches which contributed a vote near to the maximum in the voting space. We refer to Leibe et. al. [8] for a detailed description of the voting procedure.

However, due to different object appearances in multi-modal image pairs, the maximum in the voting space does not always correspond to a correct region match. Therefore we use the point correspondences which contributed a vote near the maximum in voting space (backprojection of best hypothesis supporting feature matches) to estimate (RANSAC) a local affine transformation T_a of a corresponding image patch (e.g., 60x60px). We discard all feature matches with a high self-similarity distance (eqn. 6) based on an empirical determined threshold. The remaining correspondences are refined by minimizing an intensity based multi-modal distance measure.

2.3. Local Image Patch Correspondence Optimization

Due to small errors in the determined feature correspondences we additionally minimize an intensity based multi-modal distance measure to further refine the region correspondence. Formally, we search for a set of optimal transformation parameters $\hat{\theta}$ which minimize a multi-modal distance measure $D : \mathbb{R}^N \rightarrow \mathbb{R}$ over the support of a local image region A_i around the determined point correspondences:

$$\hat{\theta} = argmin_\theta \, D(\theta), \tag{1}$$

$$D(\theta) = \int_{A_i} D_{(\cdot)}(I_{T_\theta}, I_R). \tag{2}$$

To this date we use parametric (projective) transformations T_θ with 8 degrees of freedom for distance minimization. The local image region A_i should be as small as possible for projective transformations since they inherently imply planarity. The affine transformation T_a serves as a starting point for the image alignment optimization. The nonlinear optimization is based on a specific pattern search method which does not rely on gradient information. Basically, we approximate the distance function $D(\theta)$ with a multi-variate polynomial of degree 2 and recenter/rescale the search pattern at the optimum of the surrogate polynomial. Given a proper initialization, the method needs only a few distance function evaluations to converge to a local optimum and is especially designed [12] for computationally expensive distance functions.

2.3.1. Multi-Modal Distance Measure

Mutual Information [4] is considered the gold standard similarity measure for multi-modal matching. It measures the mutual depen-

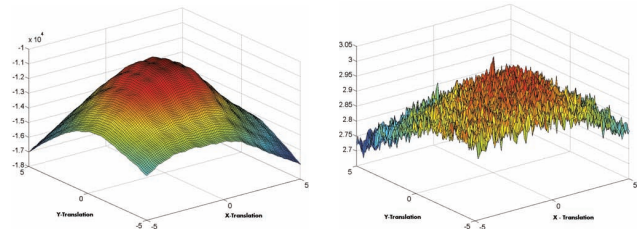
dence of the underlying image intensity distributions:

$$D_{(MI)}(I_R, I_{T_\theta}) = H(I_R) + H(I_{T_\theta}) - H(I_R, I_{T_\theta}) \quad (3)$$

where $H(I_R)$ and $H(I_{T_\theta})$ are the marginal entropies and

$$H(I_R, I_{T_\theta}) = \sum_{X \in I_{T_\theta}} \sum_{Y \in I_R} p(X,Y) log(\frac{p(X,Y)}{p(X)p(Y)}) \quad (4)$$

is the joint entropy. $p(X,Y)$ denotes the joint probability distribution function of the image intensities X, Y in I_R and I_{T_θ}, and $p(X)$ and $p(Y)$ are the marginal probability distribution functions. However, MI is very difficult to estimate (e.g.,see Fig. 3) for small image regions and does not cope well with spatially-varying intensity fluctuations (eqn. (4)). In our application scenario MI failed to correctly align corresponding textured regions with uniform ones (see Fig.4b,e - e.g., uniformly textured cloth in Visible/IR image pairs). Therefore



(a) Self-Similarity (b) MI

Fig. 3. Plots of the values of the Self-Similarity (a) and MI (b) distance measure for deviations (translation in x and y direction) from the found value by the optimization algorithm (same image patch optimization). The optimal value is shown at where all parameters are zero

we propose to minimize a self-similarity distance of corresponding local image regions in I_R and I_{T_θ}. To compute the self-similarity description for an image patch point we compare a small image patch with a larger surrounding image region centered at $q \in R_i$ using simple sum of squared differences (SSD) between image intensities normalized by the image patch intensity variance and noise $c(I)_{noise,variance}$:

$$S_q(x,y) = \exp\left(-\frac{SSD_q(x,y)}{c(I)_{noise,variance}}\right) \quad (5)$$

This correlation image $S_q(x,y)$ is then transformed into a log polar coordinate system and partitioned into bins (e.g., 20 angles, 4 radial intervals) where the maximal correlation value in each bin is used as an entry for the self-similarity dimension description of the vector $S_q^{I(\cdot)}(x,y)$ located at the image position $(x,y) \in A_i$. Each vector is then linearly normalized to $[0,1]$. The distance measure now simply computes the sum of squared distances of the self-similarity description vectors $S_q^{I(\cdot)}$ computed at the region A_i:

$$D_{(SSim)}(I_{T_\theta}, I_R) = \sum_{(x,y) \in A_i} \|S_q^{I_R}(x,y) - S_q^{I_{T_\theta}}(x,y)\|^2. \quad (6)$$

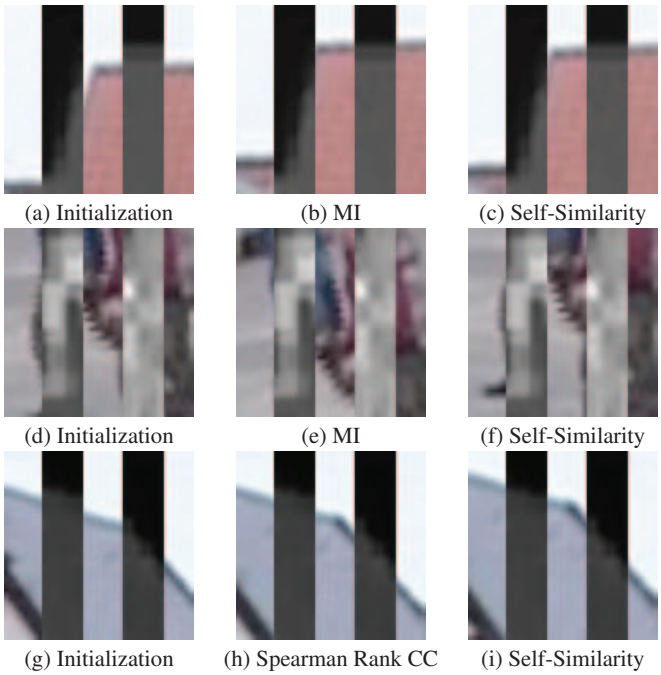
In multiple experiments we plotted the values of the optimization function while varying function parameters as shown in Fig. 3. The plots of this distance measure show unique maxima and relatively smooth and monotonically increasing function shapes.

2.4. Implementation Details and Runtime Information

Since the implemented method requires many feature correspondence searches, it is very important to use fast search structures [13] for nearest neighbors determination in descriptor space (L_2 norm). Since our implementation (C++) is not fully runtime-optimized so far, the reported time measurements provide only a rough estimate for the overall algorithm runtime. The determination of 500 region correspondences ranges from 60-200s on an Intel Q9550 System. The intensity based local image patch refinement (60x60px) needs 10-15s for one image patch optimization (Self-Similarity features and distance measure). All image pairs were transformed to the $L^*a^*b^*$ color space prior to feature extraction and image comparison. The local feature implementations are based on the SiftGPU [14] and OpenSURF [10] libraries.

3. RESULTS

To evaluate the method with different local features we evaluated the point correspondences with respect to a robustly estimated fundamental matrix (8-pt algorithm, RANSAC, 1.25px inlier threshold). When possible e.g., in case of IR/Visible aerial images we also estimated a global homography (RANSAC, 2.5px inlier threshold) to evaluate correct correspondences. In total we used 10 Visible/IR, 50 Visible/IR aerial and 2 Lidar/IR/Visible image pairs. Our experiments show (Tab. 1, Fig. 5) that this method enables the robust determination of multi-modal feature correspondences. The self-similarity descriptors proved to be well suited for this task compared to well established local feature approaches like SIFT [9] or SURF [10] (see Fig. 5) which failed to establish correct correspondences in some cases.

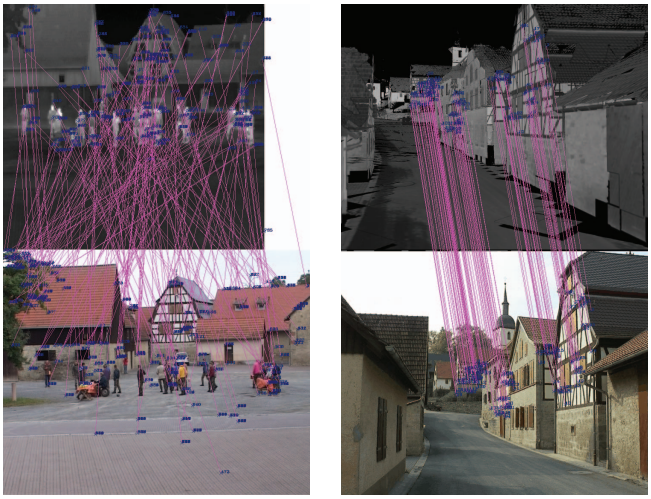


(a) Initialization (b) MI (c) Self-Similarity
(d) Initialization (e) MI (f) Self-Similarity
(g) Initialization (h) Spearman Rank CC (i) Self-Similarity

Fig. 4. (a) Visualization of local image patch correspondence refinements. Before (a,d,g) and after optimization of MI (b,e), Spearman Rank Correlation Coefficient (h) and self-similarity (c,f,i).

Features	IR-Vis	IR-Vis Aerial	Lidar-IR
SIFT	-	63 / 54	-
SURF	-	35 / 32	-
S-Sim	3706 / 1816	2185 / 1395	4881 / 1143

Table 1. Averaged rounded (found/correct) point correspondences. The correctness of point correspondences was additionally checked by visual inspection in case of fundamental matrix constraint filtering.



(a) SIFT (b) LiDAR/Vis

Fig. 5. Correspondences using SIFT for a IR/Visible image pair (a). Self-Similarity matches after fundamental matrix constraint (RANSAC, 2.25px) filtering for a LiDAR/Visible spectrum image pair (b).

To this date, we evaluated various intensity based distance measures like MI, Correlation Ratio, Gradient Correlation, Spearman Rank Order Correlation and linear combinations of these. To evaluate the patch registration performance we counted the number of (correct/false) alignments of by visual inspection with respect to MI, Spearman Rank Correlation Coefficient and Self-Similarity for a representative set of 113 local image patches. Spearman Rank Order Correlation led to 39 correct and 74 false, Mutual Information to (36/77), and Self-Similarity to (107/6) alignments. All experiments showed a clear dominance of the proposed self-similarity distance for IR/Visible and Lidar/IR/Visible image pairs in case of small region sizes (see Fig. 4).

4. CONCLUSION AND FUTURE WORK

In this work we proposed and implemented a robust method to determine accurate local multi-modal correspondences. The method is based on simultaneously matching geometrically consistent feature correspondences in combination with an intensity based optimization of small image regions. The registration of multi-modal image patches is inherent difficult due to fundamental differing object appearances. However, we find the approach of matching self-similar structures [7] very intriguing since this methodology does not assume a functional relationship like correlation ratio [15] or clusters in the joint intensity distribution like MI [4]. However, multi-modal distance measures are usually very application dependent and a general suitability of self-similarity as a distance measure remains open.

Future research directions are manifold. The optimal selection of local regions plays a major role to achieve robust initial region matches. Moreover, we work on an extension of the self-similarity descriptors and the voting algorithm to enable wide baseline scenarios. We also plan to annotate subpixel accurate landmarks on a representative set of multi-modal (IR/Vis) image pairs to enable an exact alignment accuracy evaluation.

5. REFERENCES

[1] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *IEEE TPAMI*, vol. 27, no. 10, pp. 1615–1630, 2005.

[2] B. Zitova and J. Flusser, "Image registration methods: A survey," *IVC*, vol. 21, pp. 977–1000, 2003.

[3] S. Krotosky and M. Trivedi, "Registering multimodal imagery with occluding objects using mutual information: Application to stereo tracking of humans," in *Augmented Vision Perception in Infrared*. 2009, pp. 321–347, Springer Verlag.

[4] P. Viola and W.M. Wells, "Alignment by maximization of mutual information," *IJCV*, vol. 24, no. 2, pp. 137–154, 1997.

[5] V. Sharma and J.W. Davis, "Feature-level fusion for object segmentation using mutual information," in *Augmented Vision Perception in Infrared*. 2009, pp. 295–319, Springer Verlag.

[6] A. Myronenko and X. Song, "Image registration by minimization of residual complexity," in *CVPR*, 2009, pp. 49–56.

[7] E. Shechtman and M. Irani, "Matching local self-similarities across images and videos," in *CVPR*, 2007.

[8] B. Leibe, A. Leonardis, and B. Schiele, "Robust object detection with interleaved categorization and segmentation," *IJCV*, vol. 77, pp. 259–289, 2008.

[9] David G. Lowe, "Distinctive image features from scale-invariant keypoints," *IJCV*, vol. 2, pp. 91–110, 2004.

[10] H. Bay, A. Ess, T. Tuytelaars, and Luc Van Gool, "Speeded-up robust features (surf)," *CVIU*, vol. 110, pp. 346–359, 2008.

[11] R. C. Bolles M.A. Fischler, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography.," *ACM*, vol. 24, pp. 381–395, 1981.

[12] C. Bodensteiner et. al, "Motion and positional error correction for cone beam 3d-reconstruction with mobile c-arms," in *MICCAI*, 2007.

[13] M. Muja and D.G. Lowe, "Fast approximate nearest neighbors with automatic algorithm configuration," in *VISAPP*, 2009.

[14] C. Wu, "SiftGPU: A GPU implementation of scale invariant feature transform (SIFT)," http://cs.unc.edu/ ccwu/siftgpu, 2007.

[15] A. Roche et. al, "The correlation ratio as a new similarity measure for multimodal image registration," in *MICCAI*. 1998, pp. 1115–1124, Springer Verlag.