Computer Graphics & Multimedia
Prof. Dr. Leif Kobbelt
RWTH Aachen University

Fast Image-Based Localization using Direct 2D-to-3D Matching

Torsten Sattler, Bastian Leibe, Leif Kobbelt

RWTH AACHEN UNIVERSITY

Problem

- 3D models obtained from **Structure-from-Motion**, every point has ≥ 2 SIFT descriptors [4]
- Need **2D-to-3D correspondences** from query image to 3D model for **pose estimation**
- Feature matching: **Efficient** (fast) and **Effective** (many images registered)

Direct vs Indirect Matching

- **Direct**: Effective but slow
  - Example: approx. kd-tree-based search.
- **Indirect**: Efficient but not as effective
  - Example: **Image retrieval-based** method from Irschara *et al.* [1]

c 3D point representations

- Different possibilities to represent 3D points by their descriptors inside visual words:
  - Use **all descriptors** for each 3D point
  - Compute *mean* / *medoid* of point descriptors, assign it to all visual words (vw) activated by any of the points' descriptors
  - *mean* / *medoid per vw:* Assign descriptors of point to visual words, compute mean / medoid if more than one descriptor of same 3D point assigned to same vw
  - **integer mean per vw**: Round entries of means to nearest integer values

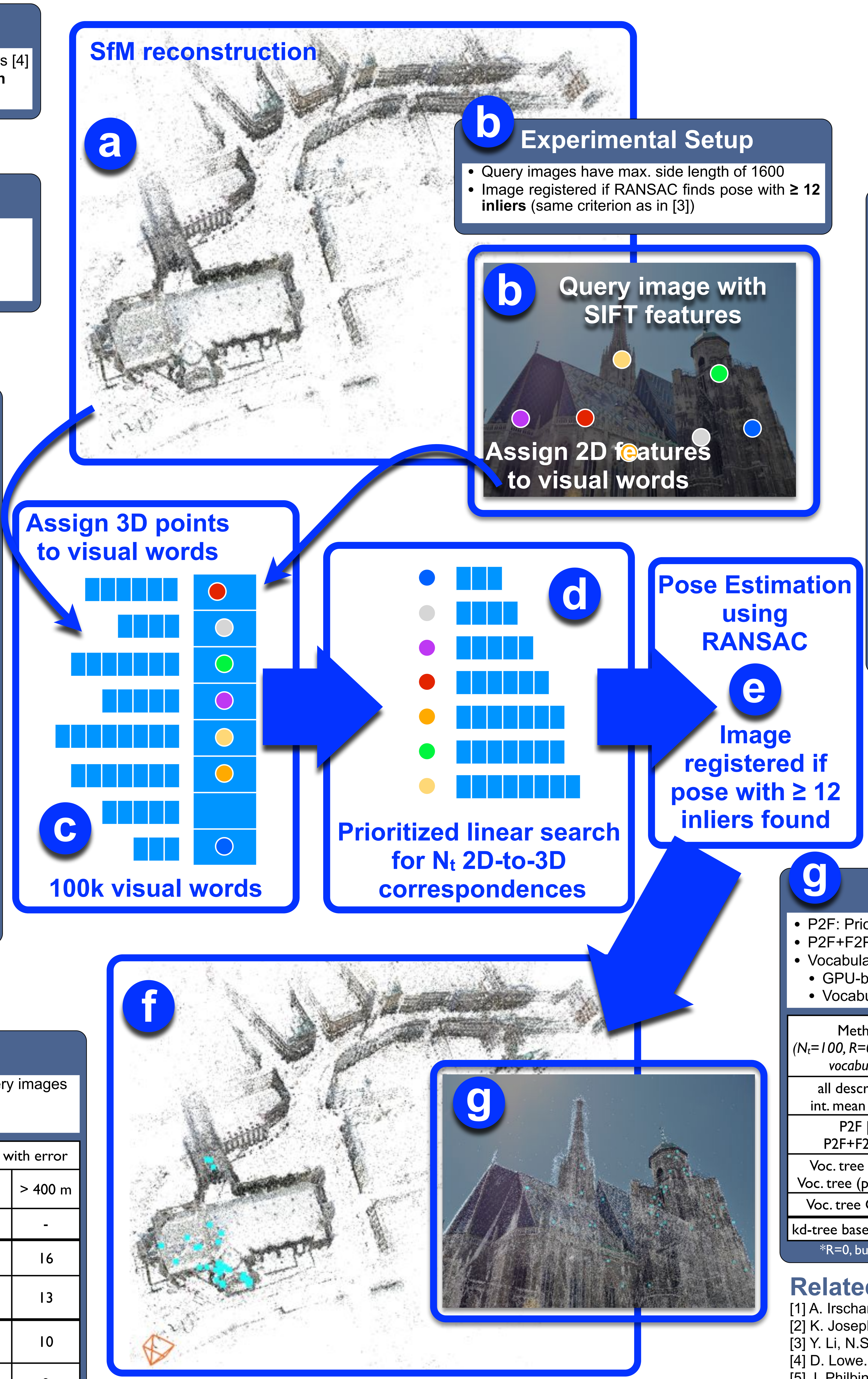
Method	Dubrovnik			Rome			Vienna		
	# reg. images	time registered [s]	time rejected [s]	# reg. images	time registered [s]	time rejected [s]	# reg. images	time registered [s]	time rejected [s]
all descriptors	**785**	**0.81**	**2.19**	**979**	**1.53**	**4.07**	**211**	**1.83**	**9.95**
mean	774	1.61	2.36	972	2.13	1.28	210	2.05	9.19
medoid	762	0.84	1.58	961	1.05	3.74	203	2.23	9.40
mean per vw	782	1.31	5.25	976	2.23	6.50	212	2.46	6.87
integer mean per vw	**783**	**0.87**	**5.35**	**976**	**1.33**	**5.92**	**211**	**2.02**	**7.59**
medoid per vw	778	0.66	4.34	972	1.17	7.27	211	1.81	8.25
kd-tree	*795*	*3.40*	*14.45*	*983*	*3.97*	*6.27*	*220*	*3.44*	*2.72*

Influence of the vocabulary

- **Generic** vocabulary from clustering SIFT descriptors from unrelated set of images [5]
- **Specific** vocabularies offer no significant improvement (slightly faster)
- Experimented with 10k, 100k and 1M visual words, best results for **100k**

b Experimental Setup

- Query images have max. side length of 1600
- Image registered if RANSAC finds pose with **≥ 12 inliers** (same criterion as in [3])



f Localization accuracy

- Ground truth: Geo-registered version of Dubrovnik model, containing positions of query images
- Compute the average camera position over 10 repetitions for every image
- For our methods: Report distance of average camera position to ground truth

Method (N_t=100, R=0.2, generic vocabulary)	# reg. images	Mean [m]	Median [m]	1st Quartile [m]	3rd Quartile [m]	Max [m]	#images with error < 18.3m	#images with error > 400 m
P2F [3]	753	18.3	9.3	7.5	13.4	~400	655	-
all desc. (p6p)	783.9 ± 1.60	53.9	1.4	0.4	5.9	7934.3	685	16
int. mean per vw (p6p)	782.0 ± 0.82	47.0	1.3	0.5	5.1	7737.1	675	13
all desc (p6p+p4pfr [2])	783.9 ± 1.60	21.6	0.8	0.2	3.0	2336.1	705	10
int. mean per vw (p6p + p4pfr)	782.0 ± 0.82	17.2	0.8	0.2	3.6	875.6	700	9

a Datasets

- Datasets used in [3], kindly provided by Li *et al.* [3] and Irschara *et al.* [1]
- Query images for Dubrovnik and Rome obtained by removing images from larger reconstructions
- Query images for Vienna obtained from Panoramio

Dataset	# Cameras	# 3D Points	# Descriptors	# Query Images
Dubrovnik	6044	1,886,884	9,606,317	800
Rome	15,179	4,067,119	21,515,110	1000
Vienna	1324	1,123,028	4,854,056	266

d Prioritized Search

- 2D-to-3D correspondences from **linear search** (through vw) and **SIFT ratio-test** $\|d-d_1\|/\|d-d_2\| \leq 0.7$
- **Prioritization**: First search through words with few descriptors, **stop if N_t correspondences are found**

	N_t	all descriptors (R=0.2, generic vocabulary) # reg.	linear search [s]	RANSAC [s]	total [s]	integer mean per visual word (R=0.2, generic vocabulary) # reg.	linear search [s]	RANSAC [s]	total [s]
Dubrovnik	50	778.9 ± 1.52	0.04	0.05	0.23 ± 0.00	775.8 ± 1.48	0.03	0.05	0.21 ± 0.00
	100	**783.9 ± 1.60**	**0.10**	**0.08**	**0.31 ± 0.01**	**782.0 ± 0.82**	**0.08**	**0.08**	**0.28 ± 0.01**
	150	783.9 ± 1.10	0.16	0.08	0.36 ± 0.01	781.8 ± 1.40	0.12	0.08	032 ± 0.01
	200	784.4 ± 1.26	0.20	0.08	0.40 ± 0.01	782.5 ± 1.35	0.15	0.08	0.35 ± 0.01
	∞	784.6 ± 1.17	0.47	0.08	0.68 ± 0.01	782.5 ± 1.08	0.34	0.08	0.54 ± 0.01
Rome	50	972.0 ± 1.41	0.06	0.02	0.18 ± 0.00	971.3 ± 1.25	0.05	0.02	0.16 ± 0.00
	100	**976.9 ± 1.29**	**0.15**	**0.05**	**0.29 ± 0.00**	**974.6 ± 1.65**	**0.11**	**0.05**	**0.25 ± 0.00**
	150	977.8 ± 1.32	0.23	0.06	0.39 ± 0.01	976.5 ± 1.51	0.17	0.06	0.33 ± 0.01
	200	979.2 ± 1.75	0.30	0.07	0.46 ± 0.01	976.9 ± 1.52	0.22	0.07	0.38 ± 0.00
	∞	980.1 ± 0.88	0.81	0.07	0.98 ± 0.00	976.9 ± 1.20	0.57	0.07	0.74 ± 0.00
Vienna	50	200.4 ± 1.26	0.02	0.13	0.28 ± 0.01	199.1 ± 1.20	0.02	0.10	0.26 ± 0.01
	100	**207.7 ± 1.06**	**0.06**	**0.30**	**0.50 ± 0.02**	**206.9 ± 0.88**	**0.05**	**0.28**	**0.46 ± 0.02**
	150	208.2 ± 0.92	0.09	0.30	0.52 ± 0.03	207.9 ± 0.74	0.07	0.29	0.50 ± 0.03
	200	208.8 ± 1.23	0.11	0.29	0.54 ± 0.04	208.2 ± 1.14	0.08	0.30	0.52 ± 0.03
	∞	207.9 ± 1.29	0.24	0.27	0.65 ± 0.03	208.2 ± 0.42	0.17	0.28	0.59 ± 0.03

e Improving the rejection times

- RANSAC will take many samples if inlier-ratio is low (usually the case for rejected images)
- Assume an inlier-ratio of ***max(12/N, R)*** for *N* correspondences ➔ larger R = faster rejection times
- ***R=0.2***: almost no effect on the registration performance, significantly reduced rejection times
- **Robustness**: Try to register query images from other datasets ➔ None of them can be registered

g Comparison with state-of-the-art

- P2F: Prioritized matching of 3D points to 2D features proposed by Li *et al.* [3]
- P2F+F2P: Match features against points if P2F fails [3]
- Vocabulary tree-based methods: Retrieve the 10 top ranked images, perform pairwise matching and pose estimation
  - GPU-based method by Irschara *et al.* [1], only available for the Vienna dataset
  - Vocabulary tree-based retrieval from Li *et al.* [3] for Dubrovnik and Rome

Method (N_t=100, R=0.2, generic vocabulary)	Dubrovnik # reg. images	Dubrovnik registr. times [s]	Dubrovnik rejection times [s]	Rome # reg. images	Rome registr. times [s]	Rome rejection times [s]	Vienna # reg. images	Vienna registr. times [s]	Vienna rejection times [s]
all descriptors int. mean per vw	**783.9 ± 1.60** **782.0 ± 0.82**	**0.31 ± 0.01** **0.28 ± 0.01**	**2.22 ± 0.26** **1.70 ± 0.18**	**976.9 ± 1.29** **974.6 ± 1.65**	**0.29 ± 0.00** **0.25 ± 0.00**	**1.90 ± 0.10** **1.66 ± 0.10**	**207.7 ± 1.06** **206.9 ± 0.88**	0.50 ± 0.02 0.46 ± 0.02	2.40 ± 0.06 2.43 ± 0.08
P2F [3] P2F+F2P [3]	753 753	0.73 0.70	2.70 3.96	921 924	0.91 0.87	2.93 4.67	204 205	0.55 0.54	1.96 3.62
Voc. tree (all) [3] Voc. tree (points) [3]	668 677	1.4 1.3	4.0 4.0	828 815	1.2 1.2	4.0 4.0	- -	- -	- -
Voc. tree GPU [1]	-	-	-	-	-	-	165	≤ **0.27**	
kd-tree based search *	**795**	3.40	14.45	**983**	3.97	6.27	**220**	3.44	2.72

*R=0, but RANSAC stopped after 1 minute

Related Work

[1] A. Irschara, C. Zach, J.-M. Frahm, and H. Bischof. From structure-from-motion point clouds to fast location recognition. CVPR'09.
[2] K. Josephson and M. Byröd. Pose estimation with radial distortion and unknown focal length. CVPR'09.
[3] Y. Li, N.Snavely, and D.P. Huttenlocher. Location recognition using prioritized feature matching. ECCV'10.
[4] D. Lowe. Distinctive image features from scale-invariant keypoints. IJCV, 60(2):91-110, 2004.
[5] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object Retrieval with large vocabularies and fast spatial matching. CVPR'07.

Source code available at http://www.graphics.rwth-aachen.de/localization

UMIC Ultra High-Speed Mobile Information and Communication

mobile ACcess