MSFD: Multi-scale segmentation based feature detection for wide-baseline scene reconstruction

Armin Mustafa, Hansung Kim, and Adrian Hilton

***Abstract*—A common problem in wide-baseline matching is the sparse and non-uniform distribution of correspondences when using conventional detectors such as SIFT, SURF, FAST, A-KAZE and MSER. In this paper we introduce a novel segmentation based feature detector (SFD) that produces an increased number of accurate features for wide-baseline matching. A multi-scale SFD is proposed using bilateral image decomposition to produce a large number of scale-invariant features for wide-baseline reconstruction. All input images are over-segmented into regions using any existing segmentation technique like Watershed, Mean-shift, and SLIC. Feature points are then detected at the intersection of the boundaries of three or more regions. The detected feature points are local maxima of the image function. The key advantage of feature detection based on segmentation is that it does not require global threshold setting and can therefore detect features throughout the image. A comprehensive evaluation demonstrates that SFD gives an increased number of features which are accurately localised and matched between wide-baseline camera views; the number of features for a given matching error increases by a factor of 3-5 compared to SIFT; feature detection and matching performance is maintained with increasing baseline between views; multi-scale SFD improves matching performance at varying scales. Application of SFD to sparse multi-view wide-baseline reconstruction demonstrates a factor of ten increase in the number of reconstructed points with improved scene coverage compared to SIFT/MSER/A-KAZE. Evaluation against ground-truth shows that SFD produces an increased number of wide-baseline matches with reduced error.**

***Index Terms*—Feature detection, Segmentation, Matching, Sparse Reconstruction.**

I. Introduction

Finding reliable correspondences between images is a fundamental problem in computer vision applications such as object recognition, camera tracking and automated 3D reconstruction. In this paper we focus on the problem of wide-baseline matching for general indoor and outdoor scenes. Existing feature detectors such as Harris [1], [2], SIFT [3], SURF [4], FAST [5], KAZE [6] and MSER [7] often yield sparse and non-uniformly distributed feature sets for wide-baseline matching and reconstruction, as seen in later sections. Gradient-based detectors (Harris, SIFT, SURF, STAR [8]) locate features at points of high-image gradient in multiple directions and scales to identify salient features which are suitable for affine-invariant matching. This results in sparse features with no detections in uniform regions. Existing segmentation based detectors such as MSER identify salient regions which can be reliably matched across wide-baseline views. However, this results in relatively few features. Existing feature detectors produces a highly sparse non-uniform distribution of scene features. Whilst this may be sufficient for camera estimation and sparse point reconstruction using bundle-adjustment, the detected feature set often results in poor scene coverage.

In this paper we introduce features based on over-segmentation of the image. We propose a new multi-scale segmentation based feature detector MSFD which uses the segmentation boundary (local maximal ridge lines of the image) rather than the segmentation regions. MSFD feature point detections are located at the intersection points of three or more region boundaries. The intersection points represent local maxima of the image function in multiple directions giving stable localization. The key advantage of this approach over previous gradient and region based feature detectors is that it does not require any global thresholds to be set. Features are detected at local maxima of the image or image gradient function. This enables feature detection throughout the image according to the local image variation giving an increased number of feature detections and accurate localisation for widely varying views.

A comprehensive performance evaluation of MSFD is performed with respect to previously proposed feature detection approaches. Evaluation of SFD feature point detections across wide-baseline views demonstrates that the region intersection points are stable and accurately localized, an example is illustrated in Figure 1. SFD feature points are also demonstrated to give improved scene coverage with computational cost similar to existing efficient wide-baseline feature detectors (SURF/FAST/A-KAZE). Our preliminary work on SFD has been previously published at [9] and a video is available online[1]. As compared to our previous paper, this paper: (a) Introduces a multi-scale SFD feature detection (MSFD) based on bilateral decomposition of the image for scale invariant feature detection; (b) Presents a comprehensive performance evaluation of SFD and MSFD on a wide variety of indoor and outdoor datasets against 11 existing feature detectors demonstrating significant performance improvement; and (c) Presents comparative evaluation of MSFD against single scale SFD [9] and previous feature detectors. Contributions of this paper are:

- A novel multi-scale segmentation based feature detector MSFD for wide-baseline matching; which gives an increased number of accurately localised features for different viewpoints and improved coverage for natural scenes.
- MSFD using bilateral filter based image decomposition for scale-invariant feature detection without global thresh-

Centre for vision speech and signal processing, Department of Electronic Engineering, University of Surrey, UK

[1] https://www.youtube.com/watch?v=m3NThvKpR5w

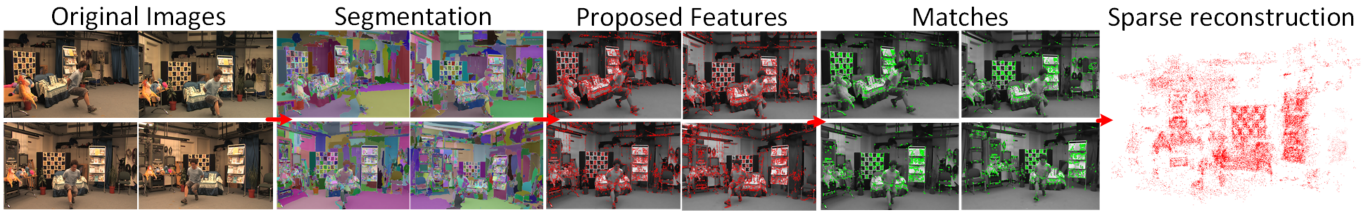


Fig. 1. Segmentation-based Feature Detection SFD for wide-baseline matching and reconstruction for Odzemok dataset.

olds.

- MSFD achieves a factor of ten increase in the number of features detected and matched for a given re-projection error for wide-baseline images.
- A comprehensive performance evaluation of SFD for wide-baseline matching on benchmark datasets against existing feature detectors (Harris, SIFT, SURF, FAST, MSER, ORB, A-KAZE) and descriptors(SIFT, BRIEF, ORB, SURF) showed improved performance in number of features and matching accuracy.

Application of multi-scale SFD to reconstruction from wide-baseline camera views demonstrates an order of magnitude increase in the number of reconstructed points with improved scene coverage and reduced error compared to previous detectors against ground-truth. Proposed MSFD achieves:

- SFD and MSFD feature detection achieves state-of-the-art performance irrespective of the choice of underlying segmentation method (Watershed, Mean-shift, SLIC).
- Performance evaluation of MSFD against single-scale SFD [9] and existing feature detectors (Harris, SIFT, SURF, MSER, A-KAZE) demonstrates increased matches and stable performance with increasing baseline between camera views at varying scales between 0.25 to 2.

II. Previous Work

Decades of research has developed numerous feature detection techniques and a review into interest-point detection reveals three main approaches [10], [11]: image gradient analysis, intensity templates, and contour analysis.

A. Image Gradient based Features

Early image gradient based approaches, such as Forstner corner detector [12], define an optimal point based on the distances from the local gradient lines and Harris corner detector [1], define an interest-point as the maximum of a function of the Hessian of the image. They used the local autocorrelation function of a signal to measure the local changes of the signal with patches shifted by a small amount in different directions. A scale-invariant extension was achieved by successive application of Gaussian kernels on scale-space representation of image and detecting interest-point as a local maximum both spatially, and across the scale-space [13] to deal with significant affine transformations. Mikolajcyzk and Schmid seek these maxima via the Laplacian-of-Gaussian (LoG) filter, which is a combination of the Gaussian smoothing and the differentiation operation [14].

SIFT implements Difference-of-Gaussians [3] to improve on earlier approaches by transforming an image into a large collection of local feature vectors which are invariant to changes in scale, illumination and local affine distortions. Lowe exploited locations that are maxima or minima of a Difference-of-Gaussian (DoG) function applied in scale space to generate local feature vector. Another detector exploits scale and DoG to extract features [15]. A combination of gradient space with local symmetry was used in [16]. Gradient based techniques offer accurate localization [17], and are robust to many image transformations [13]. However, computation of the image gradients are sensitive to image noise and are computationally expensive. SURF mitigates this via the use of integral images and 2D Haar wavelets [4].

CenSurE achieves even faster operation by approximating the LoG operator with a bi-level filter [8]. These approaches suffer from drawbacks since Gaussian blurring does not preserve object boundaries and smooths details and noise at all scales, spoiling localization accuracy and distinctiveness. To overcome this problem KAZE features were introduced to detect and describe features in non-linear scale-space [6]. The scale-space representation is computed by non-linear diffusion filtering instead of Gaussian smoothing, yielding an improvement in the localization accuracy in [6], thereby increasing repeatability w.r.t SIFT and SURF. The main drawback of KAZE is that it is computationally intense which is addressed by A-KAZE by using efficient diffusion filtering [18]. This claims superiority over all major gradient based methods in terms of computational complexity.

B. Intensity based Features

Intensity template approaches seek patterns that are common manifestations of interest-points [11]. FAST first computes the intensity differences between the central pixel and a circle surrounding it, and then counts the contiguous pixels with a difference above a threshold [5]. A rotation-invariant implementation is proposed in [19], and a multi-scale extension in [20]. An extension to a multi-scale detector by scale selection with the Laplacian function was proposed in [21].

Maximally Stable Extremal Regions (MSER) is a region detector responding to areas conforming to a ‘basin’ template [7]. The word ‘extremal’ refers to the property that all pixels inside the MSER have either higher (bright extremal regions) or lower (dark extremal regions) intensity than all the pixels on its outer boundary. The word ‘maximally stable’ describes the threshold selection process since every extremal region is a connected component of a thresholded image. In contrast SFD and MSFD detect intersection points on the region boundaries of the over-segmented image instead of detecting the regions like MSER, removing the requirement of global threshold. Intensity template methods are usually fast, compared to their gradient based counterparts [11]. However, with the exception of MSER, they are not affine-invariant, which limits their ability to cope with viewpoint variations, as presented in evaluation in [17].

C. Contour based Features

Contour intersections and junctions often result in bi-directional signal changes. Therefore, a good strategy to detect features is to extract points along the contour with high curvature. Curvature of an analog curve is defined as the rate at which the unit tangent vector changes with respect to arc length. Contours are often encoded in chains of points or represented in a parametric form using splines [22]. Hence, image contours give rise to two interest-point definitions: local maxima of the curvature along a contour, and intersections. Mokhtarian and Suomla [23] implemented the former by building a scale-space representation of the contour map for the image, and detecting the local maxima of the curvature. The robustness was improved by using gradient correlation based detector [24]. Structured tensor was exploited along with contour information to extract reliable features [25]. Intersection of contour elements provides an alternative interest-point definition. T-junctions constitute a straightforward example [23] which inspires the proposed feature detector. Performance of curvature based techniques are dependent on the quality of the extracted edges [26]. Although they are generally fast, the scale-space approach introduces a compromise between robustness and accuracy. On the other hand, contours, especially intersections are distinctive. Therefore, they are more robust to viewpoint variation [27], [26]. A recent paper proposed feature points on object contours for application to recognition [28].

D. Learning based features

Although feature detectors have mainly focused on hand-crafted methods, several learning based methods are proposed recently [29], [30], [31], [32]. A classifier was learnt to detect matchable keypoints for Structure-from-Motion (SFM) applications in [30]. They collect matchable keypoints by observing which keypoints are retained throughout the SFM pipeline and learn these keypoints. Although their method shows significant speed-up, they remain limited by the quality of the initial keypoint detector. Method was proposed to identify patch based local convolution features for application to image retrieval [33]. [29] learns convolutional filters through random sampling while searching for the filter that gives the smallest pose estimation error when applied to stereo visual odometry. Efficient features based on Gaussian kernels were proposed for classification by [34]. [32] proposed a deep-learning feature detector and descriptor for various applications. Evaluation of LIFT features is shown using the general SFM pipeline but the images are not wide-baseline. FAST detector [35] was introduced to speed-up the detection using machine learning.Improved repeatability and speed was achieved by FAST-ER [11]. Machine learning based feature detectors and descriptors [36], [37], [38] were proposed for visual and facial recognition respectively. [36] exploits the contextual information of adjacent bits for more robust feature detection. Binary features were introduced for tracking [39] such that a binary descriptor was generated and optimized for each image patch independently. But none of these features have been designed for the purpose of wide-baseline stereo.

E. Summary and Motivation

To overcome the limitations of existing feature detectors in terms of scene coverage and matching across wide-baseline views a segmentation based feature detector is introduced. It is based on the property that the intersections of contours are robust to changes in viewpoint [27], [26]. The number of features detected by curvature based techniques is quite small [26] and none of them have been proposed and evaluated on wide-baseline image pairs. They are based on only edge detection and vulnerable to the well-known difficulties in producing stable, connected, one-pixel wide contours [2]. To avoid this an over-segmentation based method for stable feature detection is proposed. The idea of using regions for salient feature matching is well known and is exploited in [40] for applications other than wide-baseline stereo. A survey on interest points based on Watershed, Mean-shift and Graph-cut segmentation was presented by [41]. This paper proposed a method that uses boundaries and centres of gravity of the segments for extracting features. Evaluation of region segmentation approaches shows that Watershed is superior to the alternatives in terms of repeatability and Mean-shift segmentation performs best for natural scenes [41]. Watershed is superior in terms of repeatability as it detects the local maxima of the gradient magnitude intensities as the region boundaries. Proposed SFD features are based on the the detection of points at the intersection of local maxima, hence Watershed is chosen as the base segmentation technique.

III. Single-scale SFD

In this section we introduce the segmentation based feature detector. The main motivation for this approach is to obtain feature detections uniformly distributed throughout the image rather than just in areas of high variation which is common with existing feature detectors. Distribution of features throughout the scene is important for applications such as matching across multiple wide-baseline views for reconstruction [42] and photo-tourism [43], as shown in Figure 1. The approach is based on over-segmentation of the image into regions which ensures that detected features are distributed across the entire image as the region boundaries are located along contours of local maxima in the image. Over-segmentation detects regions and boundaries at both strong and weak edges, ensuring that the features are robust to the changes in viewpoint. These points of local maxima are consistent with respect to viewpoint change [41]. The use of local maximal contours overcomes the common problem of setting arbitrary thresholds or scales for feature detection, which is common to most existing feature detectors. SFD feature detection is based on the segmentation of the image such that the features are detected at the boundaries of the segmented regions hence the name ‘Segmentation based features’.

A. Feature Detection

Segmentation of an image results in a large number of small regions with uniform appearance. The region boundaries represent ridge lines corresponding to local maxima of the image function or maxima in gradient if the segmentation is performed on a gradient image. The boundary intersection points where three or more region boundaries meet are

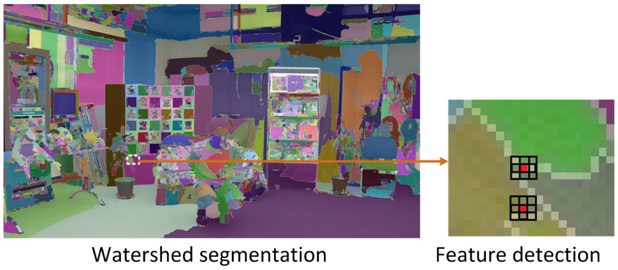


Fig. 2. Illustration of SFD feature detection on the watershed segmentation for Odzemok dataset.

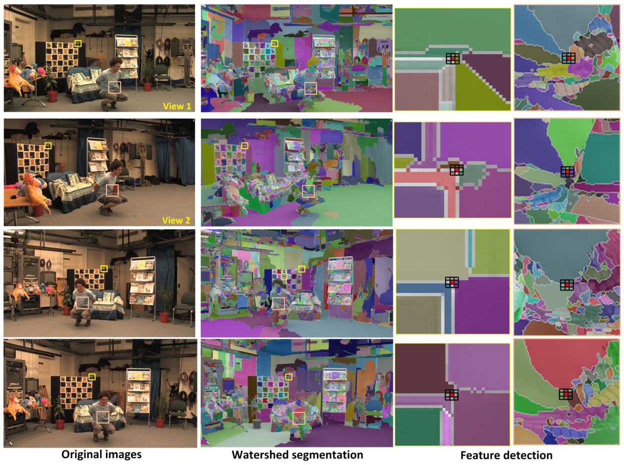


Fig. 3. SFD feature detection on Odzemok dataset for 4 views illustrating the stability of SFD with changes in viewpoint.

local maxima in the image function in multiple directions. Consequently these points are accurately localized, distinctive and stable under changes in viewpoint giving good features for matching across wide-baseline views. This observation forms the basis of our proposed region based feature detector, resulting in an increased number of salient features which are suitable for matching across wide-baseline views.

Over-segmentation is performed on the image using existing segmentation techniques such that the regions in the image are separated by a 1 pixel wide boundary. The intersection points of three or more region boundaries in the over-segmented image are detected as features and a unique intersection can only be obtained with regions of 1 pixel wide boundary. The region intersection points are identified by traversing through the boundary points in the image such that for each point on the contour 3×3 pixel neighbourhood is tested for the number of region labels. If three or more region labels are present the point is detected as a feature as illustrated in Figure 2. These points are detected for the whole image on the region boundary contours. Locating features where multiple region boundaries (3 or more) intersect followed by sub-pixel refinement gives good localization, therefore SFD achieves good localization which is consistent with-respect-to changes in viewpoint, as illustrated in Figure 3 for Odzemok dataset. Segmentation is obtained for different viewpoints with a baseline varying between $0^\circ - 90^\circ$ followed by SFD feature detection. SFD features consistent across views are highlighted in the figure to show the stability of localization with viewpoint.

B. Sub-pixel Refinement

Let us denote the set of features detected for an image as $F = \{f_1, f_2, ..., f_{N_F}\}$, where N_F is the total number of features. These features are integer values of the pixels where intersections of regions are detected. We perform a local sub-pixel refinement to optimize the feature location f_i at a local gradient maxima using the Levenberg-Marquardt method [44]. This refinement is based on the observation that every vector from the feature f_i to a point p_j located within a neighbourhood $\mathcal{N}$ of $f_i = \{x, y\}^T$ is orthogonal to the image gradient $G_j = \{g_x, g_y\}^T$ at $p_j = \{x + \Delta x, y + \Delta y\}^T$, where $\Delta x, \Delta y$ is the shift at the point f_i. In our case a window size of $W \times W$ is chosen for the neighbourhood $\mathcal{N}$, such that $W = \frac{min(N_W, N_H)}{100}$ which is the optimum window size for good localization of the features [45]. N_W and N_H are the width and height of the input image. The cost function is defined as:

$$T(f_i) = \sum_{j \epsilon \mathcal{N}} t_j(f_i), \text{ where, } t_j(f_i) = (G_j^T(f_i - p_j)(1 - e^{-\frac{\Delta x_i^2 + \Delta y_i^2}{2}}))^2 \tag{1}$$

Since the vectors G_j and $f_i - p_j$ are orthogonal, $t_j(f_i)$ is 0 if f_i is at a local maxima, thereby making $T(f_i)$ to be 0. The sub-pixel position of the feature point is the minima of $T(f_i)$. The process is repeated for the entire feature set F to obtain a new solution $F^* = \text{argmin}_{f_i} \{T(f_i)\}$ and the speed is optimized by parallelization. Feature descriptors are then applied to the local image regions of F^* to perform matching and reconstruction.

C. Segmentation

SFD can use different segmentation techniques, in this section we review possible segmentation methods. The performance of SFD for different segmentation methods is evaluated in Section VI.

Segmentation of an image is defined as the process of partitioning an image into multiple segments. Pixels in each region share similar properties and are distinct from the pixels in adjacent regions. The boundary of the segments define contours of local maxima in the image. Our focus is on finding fast, automatic and stable over-segmentation techniques suitable for wide-baseline matching in general indoor or outdoor scenes. The SFD features defined in Section III-A are evaluated on three different segmentation techniques:

Watershed (WA) [46]: The first segmentation technique is based on morphology. Readers are referred to [47] for detailed information on morphological segmentation techniques; the watershed transform [46] is used in this approach because of speed and efficiency. The watershed transformation considers the gradient magnitude of an image as a topographic surface. Pixels having the highest gradient magnitude correspond to watershed lines which represent the region boundaries. Water placed on any pixel enclosed by a common watershed line flows downhill to a common local intensity minimum. Pixels draining to a common minimum form a basin, which represents a segment partitioning the image into two different sets: the catchment basins and the watershed lines.

Implementing the transformation on the image gradient, the catchment basins correspond to homogeneous grey level regions of this image. In practice, this transform produces an over-segmentation due to scene structure, local appearance variation and image noise. We use the modified and more

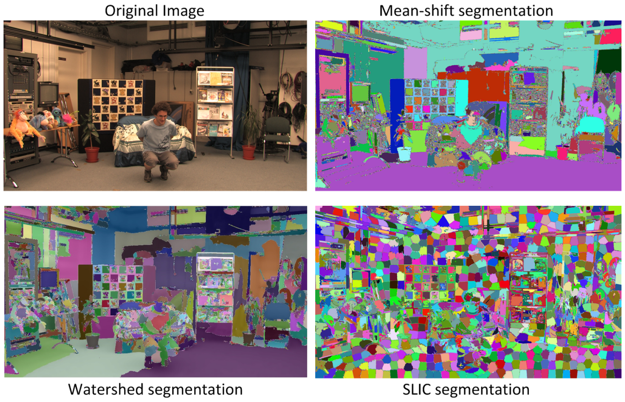


Fig. 4. Different segmentation algorithms for SFD feature detection.

robust version of the watershed algorithm defined in [48]. An example on Odzemok dataset is shown in Figure 4.

Mean-shift (MS) [49]: Mean-shift considers feature space as a empirical probability density function. For each data point, Mean-shift associates it with the nearby peak of the datasets probability density function. For each data point, Mean-shift defines a window around it and computes the mean of the data point. Then it shifts the center of the window to the mean and repeats the algorithm till it converges. After each iteration, the window shifts to a more denser region of the dataset. There are three main parameters considered in this segmentation: Spatial resolution parameter(SRP) which affects the smoothing and connectivity of segments, it is chosen depending on the size of the image, Range resolution parameter (RRP) which affects the number of segments and the third parameter is Size of smallest segment ($S3$). The parameters are initialized automatically and assignments to each of these parameters are: $SRP = \frac{N_W \times N_H}{7.776\times10^4}$, $RRP = \frac{N_W \times N_H}{7.776\times10^4}$, $S3 = w_{min} * h_{min}$,where N_W and N_H are the width and height of input image and w_{min} and h_{min} are the minimum width and height of segmented regions which is set to approx 60×30 respectively.

The mean-shift segmentation method is based on connectedness criterion and is proved to give stable and repeatable segments for natural scenes [41]. This is an unsupervised over-segmentation technique performed on image pre-processed using Bilateral filter to remove noise. An example is shown in Figure 4 on Odzemok dataset.

Simple Linear Iterative Clustering Super-pixels (SLIC) [50]: This segmentation technique is a super-pixel method and it clusters pixels in the combined five-dimensional color and image plane space to efficiently generate compact, nearly uniform super-pixels with a low computational overhead. This approach generates super-pixels by clustering pixels based on their color similarity and proximity in the image plane. SLIC is demonstrated to achieve good quality segmentation at a lower computational cost over state-of-the-art super-pixel methods.

The segmentation requires the number of regions (S) as input and this is calculated it using the following equation in this work: $S = \frac{W \times H}{w_{min} \times h_{min}}$, where W and H are the width and height of input image and w_{min} and h_{min} are the minimum width and height of segmented regions which is set to approx 60×30 respectively to avoid very small segments as shown in Figure 4.

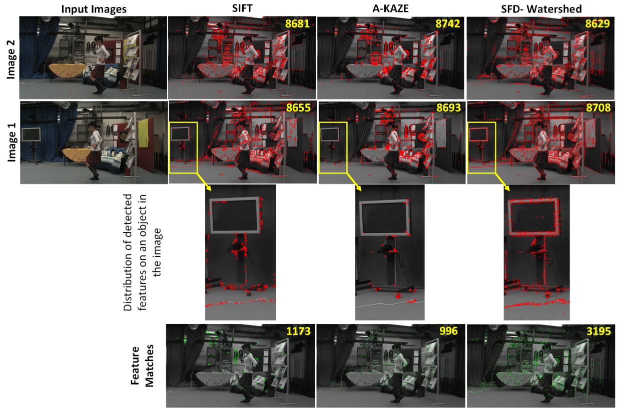


Fig. 5. Results for Dance1 dataset: Top 3 rows: Features detected on pair of images from SIFT, A-KAZE and SFD approach using watershed segmentation. Last row: Matched features. The number of features and matches are shown on the top right of each image respectively.

In this section SFD feature detection introduced for uniform scene coverage in the wide-baseline scene reconstruction was described in detail. Various segmentation techniques like Watershed, Mean-Shift and SLIC used for proposed feature detection were explained. These Single-scale SFD features work well for wide-timeframe matching as shown in [51]. Although SFD depends on over-segmentation of the image it should be noted that increasing the segments to a denser level will lead to higher computational complexity and exponentially reduced matching accuracy. When images undergo transformations or deformations the segmentation approaches are able to retrieve consistent edges in the images. The change in viewpoint or lighting causes a slight variance in the strength of the edges but the points of intersection remain robust to these deformations.

Feature detection is performed on pair of multi-view images which is followed by feature matching. An example of comparison of SFD features detected and matched is shown in Figure 5, against SIFT and A-KAZE. The highlighted region in the image shows that SFD gives improved coverage of features as compared to existing state-of-the-art methods. Feature matches are used for camera parameter estimation followed by wide-baseline sparse scene reconstruction, explained in Section V.

IV. Multi-scale SFD

Although SFD points are invariant to rotation and illumination changes, the features are not invariant to the scale. Hence multi-scale SFD (MSFD) is proposed to introduce scale invariance by constructing a bilateral image pyramid. A bilateral filter is used for our multi-scale algorithm because it avoids the halo artefacts commonly associated with the traditional Laplacian image pyramid and it preserves strong edges compared to Gaussian filtering over different scales. In existing scale-space methods either the image size is varied and a filter (e.g., Gaussian filter) is repeatedly applied to smooth subsequent layers, or the original image is kept unchanged, varying the filter size to change the scale. In this paper we choose the second approach to reduce redundancy and to avoid blurring of the edges due to down-sampling at every pyramid level.

Previously bilateral filter decomposition has been used for detail enhancement in images [52]. In this paper a series of

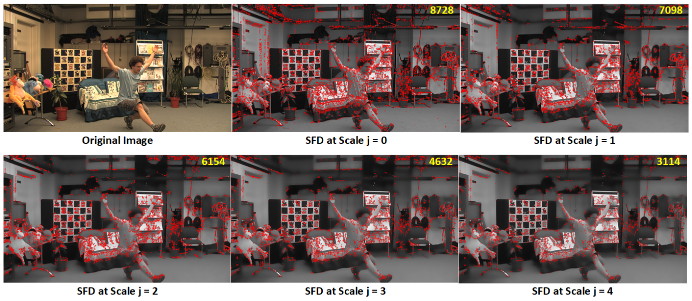


Fig. 6. Multi-scale SFD feature detection on Odzemok dataset for different scales from $j = 0$ to 4.

filtered images I_j are obtained, such that the strongest edges in the input image I are preserved while smoothing small changes in intensity for multi-scale segmentation based feature detection. The bilateral filtering operation at each scale j is defined as follows:

$$I^j_p = \frac{1}{k} \sum_{q \in \Omega} g_{\sigma_{s,j}} \left(\|q\| \right) . g_{\sigma_{r,j}} \left(I^j_{p+q} - I^j_p \right) . I^j_{p+q} \tag{2}$$

where p is a pixel coordinate, $k = \sum_{q \in \Omega} g_{\sigma_{s,j}} \left(\|q\| \right) . g_{\sigma_{r,j}} \left(I^j_{p+q} - I^j_p \right)$, $, g_\sigma(x) = exp(-\frac{x^2}{\sigma^2})$, $\sigma_{s,j}$ and $\sigma_{r,j}$ are the widths of the spatial and range Gaussians respectively and q is the offset window around p. The scale of the filter j varies from 0 to 4, where $j = 0$ is the finest scale such that $I_0 = I$. The number of scales are selected based on the experiments in the previous papers [3], [52].

In single scale SFD, explained in section III-B, features F^* are detected on single scale for image I. MSFD features are extracted by over-segmenting the set of filtered images I_j for each input image I. Let us denote the set of features detected after sub-pixel refinement at each scale j with F^*_j. SFD features are detected to obtain F^*_j, such that $F^*_{j+1} < F^*_j$ due to the reduction in the level of detail at each level as shown in Figure 6. The final set of multi-scale SFD features is defined as: $F^* = \sum_{j=0}^{4} F^*_j$, such that the redundant features are removed from the final set. An evaluation of the proposed MSFD against SFD and existing feature detection methods like SIFT, A-KAZE is presented in Section VI-B.

V. Wide-baseline Scene Reconstruction

Wide-baseline correspondences are obtained for all pairs of images using SFD. These correspondences are used to reconstruct a sparse 3D representation of the scene. We assume that the camera intrinsics are known and camera extrinsics together with 3D point locations are estimated using the correspondences. The fundamental matrix estimation procedure employs RANSAC and the normalized 8-point algorithm [53], to find the epipolar geometry using the intrinsics. The first camera is chosen as the world reference frame to obtain the camera matrix for the second camera from the fundamental matrix. Then, for each image correspondence, the triangulation algorithm [53] seeks the 3D point that minimizes the re-projection error. After the initial pairwise sparse reconstruction is obtained, a new camera is registered to the structure by finding the 2D and 3D correspondences between views and the 3D structure. The view with highest correspondences is selected and pose is estimated for the view from 3D-2D point correspondences using the RANSAC algorithm. The estimated pose minimizes re-projection error and the scene is augmented by triangulating the correspondences. The process is repeated for all the views until the camera pairs are exhausted. The algorithm employs global bundle-adjustment [43] to minimize the re-projection error over the calibration and the structure parameters to get the sparse reconstruction.

Dataset	Resolution	Number of views	Baseline
Odzemok	1920 × 1080	8(2 moving)	15°
Dance1	1920 × 1080	7(1 moving)	15°
Office	1920 × 1080	8(all static)	15°
Magician	960 × 544	5(all moving)	40°-55°
Rossendale	1920 × 1080	8(all static)	25°
Cathedral	1920 × 1080	8(all static)	45°
Patio	1920 × 1080	12(all static)	15°
Juggler	960 × 544	6(all moving)	25°-30°
Building, Books, Cloth & Architecture	800 × 600	119(all static)	15°-30°
Merton	1024 × 768	3(all static)	10°-15°
Valbonne	512 × 768	15(all static)	15°-30°
Castle	1024 × 768	19(all static)	10°-15°
Car	512 × 768	7(all static)	15°-30°

TABLE I
The characteristic properties of datasets used for evaluation.

VI. Results and Evaluation

Evaluation is performed on a variety of datasets: static and dynamic; indoor and outdoor scenes. State-of-the-art feature detection techniques have used the static indoor and outdoor datasets in their evaluation, hence these datasets are included in our evaluation for fair comparison. The characteristics of datasets are presented in Table I. Various benchmark dynamic datasets have been included to emphasize the importance of SFD in wide-baseline dynamic scene reconstruction. Wide-baseline image/video datasets (15-45 degree angle between adjacent cameras) of natural indoor and outdoor scenes under variable lighting are: **1. Static indoor datasets [17]**: Building, Books, Cloth, Architecture. We have chosen datasets from different categories in [17] such that the datasets are relevant to the problem of dynamic wide-baseline scene reconstruction, which is main focus of SFD features. Challenges: Variable lighting and viewpoints; **2. Static outdoor datasets**: Merton CollegeI [2], Valbonne[2], Castle[2], Car[2]. Challenges: Repetitive background, varying lighting condition. ; **3. Dynamic indoor datasets**: Odzemok[3] , Dance1[3], Office[3], , Magician[4]. Challenges: Both scattered and uniform background, stable lighting condition, single and multiple objects. Magician is captured with only hand-held cameras; and **4. Dynamic outdoor datasets**: Rossendale[3], Cathedral[3], Patio[3], Juggler[4]. Challenges: Both scattered and uniform background, repetitive background, variation in illumination. Juggler is captured with only hand-held cameras.

The SFD feature detector is evaluated based on the properties of good features described in [54]: quantity; efficiency; accuracy; coverage; and reconstruction accuracy in the following sections. Following sections present extensive experimental results obtained on the standard evaluation set of [13] and on practical wide-baseline image matching applications.

[2] http://www.robots.ox.ac.uk/∼vgg/data/

[3] http://cvssp.org/data/cvssp3d/

[4] http://www.inf.ethz.ch/personal/lballan/datasets.html

Dataset	SIFT	AKAZE	MSER	SURF	Harris	FAST	S-WA	S-MS	S-SLIC
Odzemok	8101	8102	8066	8043	8005	8029	8169	7908	8093
Dance1	7929	8018	7996	8121	8215	8197	8242	7956	8305
Office	7948	8176	8027	8036	7910	8032	8074	7822	7998
Magician	8018	7994	7889	7969	7915	7904	7921	7878	7909
Rossendale	6462	6543	6312	6498	6614	6590	6576	6349	6542
Cathedral	7911	7831	7845	7964	7894	7776	7806	7747	7983
Patio	7231	7390	7197	7207	7120	7225	7207	7156	7176
Juggler	5445	5498	5209	5467	5327	5476	5231	5196	5435
Building	4954	5029	4499	4962	4855	4838	4981	4809	4943
Books	4898	4907	4714	4813	4705	4853	4877	4796	4814
Cloth	4467	4591	4361	4487	4339	4502	4532	4321	4559
Architecture	4736	4818	4545	4775	4721	4681	4790	4683	4897
Merton	9882	10076	9941	9910	9897	9862	9947	9817	10336
Valbonne	3158	3223	3095	3178	2994	3169	3251	2939	3065
Castle	5597	5846	5416	5582	5737	5713	5674	5547	5848
Car	7851	8002	7761	7914	7812	7809	7764	7939	8065

TABLE II
THE NUMBER OF FEATURES DETECTED WITH EXISTING FEATURE DETECTORS (COLUMN 2^{nd} TO 7^{th}) AND PROPOSED SFD FOR THREE DIFFERENT SEGMENTATION APPROACHES (COLUMN 8^{th} TO 10^{th})

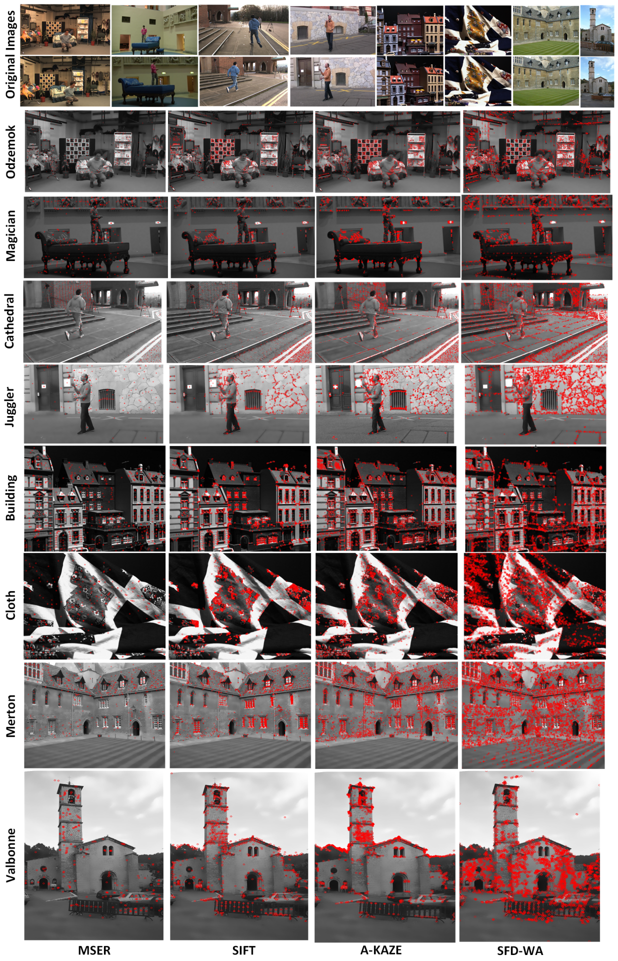


Fig. 7. Results for all datasets: Top two rows: Pair of images from each dataset, Bottom 8 rows: Column $1^{st} - 3^{rd}$ - Features matched between each image pair using MSER, SIFT and A-KAZE respectively, Column 4^{th} - Features matched between pair of images using proposed SFD features.

A. Benchmark Evaluation of Detector-Descriptor

To evaluate the performance of the proposed segmentation based feature detection approach for wide-baseline matching we present a comprehensive comparison with existing state-of-the-art feature detector and descriptor combinations. Comparison is performed with binary (FAST [18], ORB [19], BRIEF [55]) and floating point (Harris [1], GFTT [56], SIFT [3], SURF [4], STAR [8], MSER [7], KAZE [6], A-KAZE [18]) detectors. These detectors are combined with feature descriptors (BRIEF [55], ORB [19], SIFT [3], SURF [4]). We evaluated performance of different segmentation techniques. Adjacent pairs of images are taken from each dataset and segmentation is performed using Watershed, Mean-Shift and SLIC giving three variants SFD-WA, SFD-MS and SFD-SLIC respectively.

A single orientation value is used for each similar descriptor assignment for all detectors including SFD and MSFD for fair comparison. We have paired descriptors to detectors based on their optimum performance. For example: SIFT, SURF, KAZE, A-KAZE, BRIEF, and ORB are paired with their respective descriptors. Detectors for which no specific descriptors are available, has been paired according to the floating point and binary classification. For example: FAST with BRIEF, MSER with SIFT. SFD is tested with both binary (BRIEF) and floating point (SIFT) descriptor. For SIFT, SURF, STAR, ORB, FAST, MSER, Harris and GFTT we use the OpenCV based implementation. For KAZE and A-KAZE the implementation available from the paper [6], [18] is used. The feature detection thresholds of the different methods are set to values to detect approximately the same number of features per image as shown in Table II. Feature matches for MSER, SIFT and A-KAZE feature detectors are shown in Figure 7 against SFD-WA.

Scene Coverage: The distribution of the features across the scene is shown in Figure 7 for different detectors: Proposed SFD with WA, SIFT, MSER. SFD gives improved scene coverage with higher quantity and improved distribution of features across the scene for all the datasets.

Segmentation	Watershed			Mean-Shift			SLIC		
Dataset	$\lvert F^*\rvert$	TC	RC	$\lvert F^*\rvert$	TC	RC	$\lvert F^*\rvert$	TC	RC
Odzemok	8169	6543	3717	7908	5913	3547	8093	7812	**4921**
Dance1	8242	6372	3394	7956	5652	3087	8305	7499	**4459**
Office	8074	6501	3508	7822	5908	3321	7998	7667	**4768**
Magician	7921	5057	2844	7878	4524	2698	7909	6629	**3066**
Rossendale	6576	4528	2332	6349	4075	2118	6542	4786	**2972**
Cathedral	7806	6324	3452	7747	6450	3601	7983	6161	**3882**
Patio	7207	5215	3270	7156	5309	3431	7176	5642	**3986**
Juggler	5231	4478	2342	5196	4563	2267	5435	4657	**2878**
Building	4981	3531	1983	4809	3467	2067	4943	3791	**2240**
Books	4877	3348	2019	4796	3259	1984	4814	3429	**2319**
Cloth	4532	3424	1732	4321	3349	1689	4559	3568	**1981**
Architecture	4790	3213	1654	4683	3563	1780	4897	3664	**2091**
Merton	9947	7644	4533	9817	6899	4485	10336	8899	**5920**
Valbonne	3251	2715	1135	2939	2217	1252	3065	2952	**1981**
Castle	5674	4043	2351	5547	4146	2474	5848	4420	**2559**
Car	7764	4915	3435	7939	5017	3552	8065	5152	**3975**

TABLE III
FEATURE DETECTION AND MATCHING OF SFD FOR THREE DIFFERENT SEGMENTATION APPROACHES (BEST HIGHLIGHTED IN BOLD): F^* IS THE NUMBER OF FEATURES DETECTED, TOTAL COUNT (TC) IS THE NUMBER OF MATCHES WITH BRUTE FORCE MATCHING USING A SIFT DESCRIPTOR AND RANSAC COUNT (RC) IS THE NUMBER OF CORRESPONDENCES THAT ARE CONSISTENT WITH THE RANSAC BASED REFINEMENT.

Feature matching evaluation The proposed SFD detection is performed on each pair of images for each segmentation method followed by feature matching using a descriptor. An exact nearest-neighbour matching algorithm is applied, followed by a ratio test as explained in [3] is used to evaluate the feature detector. All of the matches whose distance ratio is greater than 0.85 are rejected, which eliminates 90% of

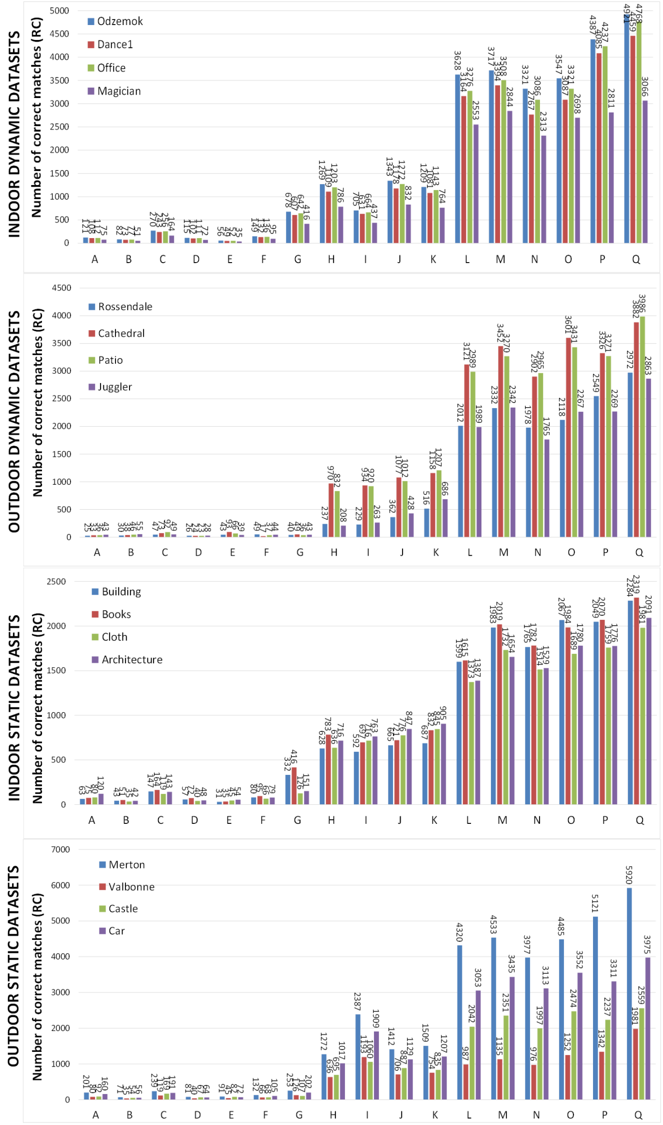


Fig. 8. Evaluation of number of correct matches on all datasets. The number of matches are shown on top of each bar.

Dataset	SIFT	A-KAZE	KAZE	SURF	MSER	SFD-WA
Odzemok	1244	1179	1322	689	313	**1615**
Dance1	1087	1064	1159	622	301	**1540**
Office	1198	1121	1263	655	304	**1519**
Magician	772	757	821	427	264	**1376**
Rossendale	226	496	237	212	125	**1277**
Cathedral	965	1136	969	919	424	**1447**
Patio	824	1204	1117	905	423	**1503**
Juggler	206	1298	669	255	128	**1612**
Building	618	640	629	581	257	**1538**
Books	772	814	683	682	372	**1615**
Cloth	611	831	755	703	439	**1432**
Architecture	705	893	827	749	447	**1454**
Merton	1230	1487	1387	1653	879	**1791**
Valbonne	606	741	689	1129	339	**1655**
Castle	648	819	851	982	365	**1531**
Car	993	1197	1043	1630	463	**1735**

TABLE IV
NUMBER OF MATCHES (TC) OBTAINED FOR EXISTING FEATURE DETECTORS AND SFD, WHEN THE INITIAL NUMBER OF DETECTED FEATURES WERE CONSTRAINED TO 2000 FOR ALL DATASETS.

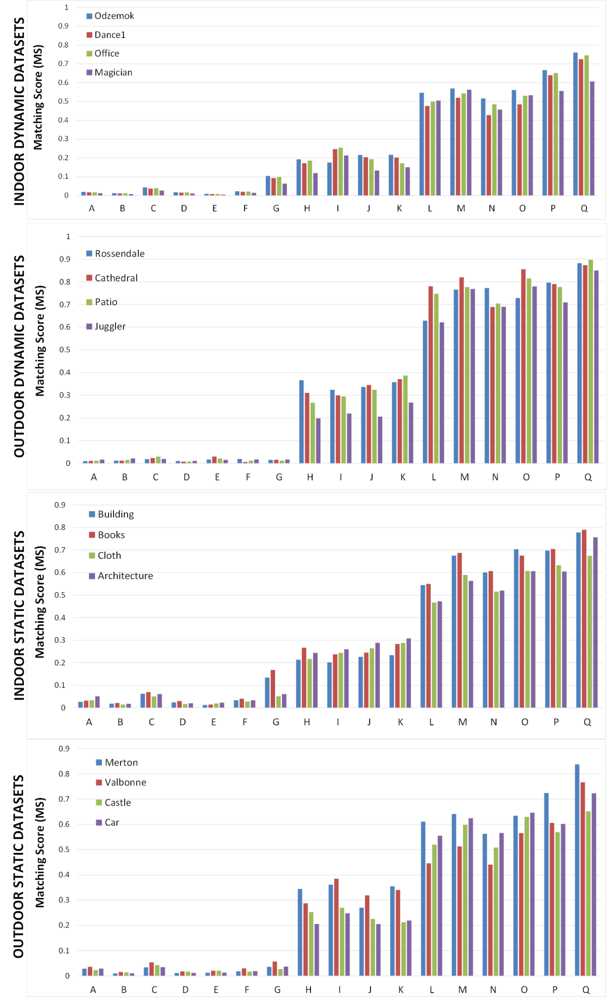


Fig. 9. Evaluation of matching score ($MS = TC/F$) on all datasets.

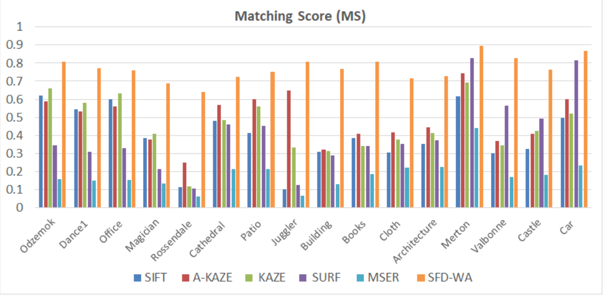


Fig. 10. Evaluation of matching score ($MS = TC/F$) for existing feature detectors and SFD, when the initial number of detected features were constrained to 2000 for all datasets.

false matches and 5% of the correct matches [3]. After obtaining a set of refined matches, a left-right symmetry test is used to further remove inconsistent matches due to repeated patterns. This is followed by RANSAC based refinement [57] of matches without prior knowledge of camera parameters. The fundamental matrix is estimated using RANSAC and the inliers are chosen as the set of matches.

Experimental results for a pair of image for each dataset and all segmentation methods (WA, MS and SLIC) are summarized in Table III. The column headed '$|F^*|$' shows the number of features detected in one of the images. Total count (TC) is the number of matches obtained with brute force matching using

	MSER		SIFT		A-KAZE		SFD-WA	
Dataset	RC	MRE	RC	MRE	RC	MRE	RC	MRE
Odzemok	119	1.390	1269	**1.175**	1209	1.181	3717	1.351
Dance1	102	1.362	1109	1.231	1081	**1.214**	3394	1.251
Office	111	1.431	1203	1.403	1143	1.361	3508	**1.354**
Magician	72	1.255	786	1.104	764	**1.045**	2844	1.195
Rossendale	26	1.411	237	1.323	516	1.318	2332	**1.315**
Cathedral	24	1.386	969	**1.152**	1158	1.154	3452	1.179
Patio	24	1.396	832	1.223	1207	**1.212**	3270	1.256
Juggler	28	1.298	208	1.155	686	**1.110**	2342	1.237
Building	57	1.240	629	**1.103**	689	1.120	1983	1.221
Books	72	1.314	783	1.210	832	1.223	2019	**1.207**
Cloth	40	1.211	636	**1.098**	845	1.201	1732	1.159
Architecture	49	1.273	719	**1.192**	905	1.206	1654	1.211
Merton	81	1.255	1272	1.177	1509	1.196	4533	**1.175**
Valbonne	41	1.258	636	1.181	755	1.187	1135	**1.159**
Castle	67	1.318	695	1.171	835	**1.152**	2351	1.208
Car	65	1.330	1018	1.213	1207	1.200	3435	**1.183**

TABLE V
EVALUATION OF MATCHING ACCURACY OF SFD AGAINST MSER, SIFT AND A-KAZE USING THE GROUND-TRUTH RECONSTRUCTION AND CAMERA CALIBRATION.

a SIFT descriptor and RANSAC count (RC) is the number of correspondences that are consistent with the RANSAC based refinement performed after the ratio and symmetry tests. The number of features detected by all segmentation techniques are similar. The numbers of matches reduces by $30 - 40\%$ after refinement using the symmetry and RANSAC tests (RC).

Figure 8 presents the number of wide-baseline matches (RC) and Figure 9 presents the matching score ($MS = TC/F$) for following detector-descriptor assignments: A: FAST-BRIEF, B: Harris-SIFT, C: GFTT-SIFT, D: MSER-SIFT, E: ORB-ORB, F: STAR-BRIEF, G: BRIEF-BRIEF, H: SIFT-SIFT, I: SURF-SURF, J: KAZE, K: A-KAZE, L: SFD-WA-BRIEF, M: SFD-WA-SIFT, N: SFD-MS-BRIEF, O: SFD-MS-SIFT, P: SFD-SLIC-BRIEF and Q: SFD-SLIC-SIFT. We choose matching score as one of our evaluation parameter for fair evaluation as it removes the bias in having more keypoints.

Performance of the proposed SFD detector combined with WA, MS and SLIC segmentation techniques with BRIEF and SIFT descriptors is shown in bars labelled L - Q, respectively demonstrating that the approach consistently achieves a factor of $3-10$ increase in the number of correct matches and factor $5-8$ in the matching score compared to previous detector-descriptor combinations.

To demonstrate that the proposed SFD features deliver both better results with the same number of keypoints, and achieves a greater number of matches than other methods, we restrict the number of keypoints detected to 2000. The number of matches TC obtained using each method is listed in Table IV and the matching score defined as $TC/2000$ is plotted in Figure 10, demonstrating consistently improved performance obtained using SFD in the matching score.

Matching accuracy evaluation: For further evaluation this section compares the feature matching accuracy of SFD against MSER, SIFT and A-KAZE. We choose only these detectors because they outperform the other feature detection methods for wide-baseline matching. SFD-WA is chosen as our base segmentation technique. The aim is to evaluate the accuracy of each feature correspondence. The ground-truth camera calibration and reconstruction is used for evaluation of the accuracy of the feature matches for all datasets. Ground-truth camera calibration (intrinsic and extrinsic) is known and reconstruction is available for static indoor datasets and for other datasets the reconstruction is computed using existing reconstruction algorithms [53]. The accuracy is evaluated using the projection of a 3D point which gives the ground-truth match for a point in pair of images. The error between the ground-truth match and the match obtained for different feature detection approaches gives the measure of accuracy.

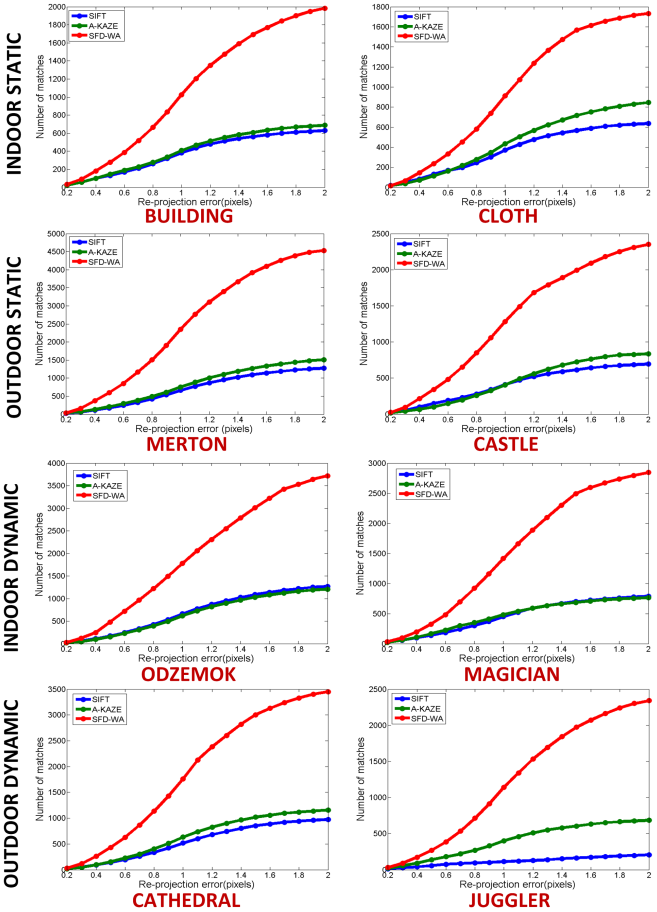


Fig. 11. Accuracy results for dynamic datasets: Re-projection error cumulative distribution of SIFT, A-KAZE and SFD-WA

Ground-truth correspondences are obtained by back-projecting the 3D location of the feature points detected in one image to the other image and evaluating the distance to the estimated feature match. Mean re-projection error (MRE) given in Equation 3 is used for accuracy evaluation of the SFD feature matches again the ground-truth.

$$MRE = \frac{1}{(K+1)} \sum_{0}^{K} \sqrt{(x-x')^2 + (y-y')^2} \qquad (3)$$

where (x, y) is the estimated SFD feature match, (x', y') is the re-projected point, and K is the number of feature matches, here $K = RC$.

Table V presents the results of the ground-truth correspondence for the proposed SFD using Watershed segmentation with a SIFT descriptor for matching and three other detector-descriptor combinations representing state-of-the-art detectors-descriptors (MSER-SIFT, SIFT-SIFT and A-KAZE-A-KAZE). RC shows the number of correspondences obtained with each approach after symmetry and RANSAC consistency tests. The number of matches obtained with the proposed SFD feature detector is greater by an order of magnitude than MSER, and by a factor three greater than SIFT and A-KAZE. The MRE for SFD is lower compared to MSER and comparable with SIFT and A-KAZE within approx. ± 0.2 pixels.

MRE gives an overall comparison of the accuracy of the feature matches, however the distribution of the matches at different pixel errors is not clear. Although the MRE of SFD is comparable with existing feature detectors, it is noted that SFD gives large number of matches and it would be interesting to see the comparison of number of matches at each pixel error against existing detectors. The more the number of matches at lower pixel error the better the accuracy of the feature detector. Hence to evaluate this re-projection error is calculated using the ground-truth reconstruction for each feature match. The errors are ranked from low to high and a graph is plotted for the number of feature matches at each pixel error.

The comparative evaluation of the re-projection errors for all the correspondences obtained by SIFT, A-KAZE and SFD is plotted and results for 2 datasets from each category are shown in Figure 11. This figure shows that the number of wide-baseline matches for a given maximum re-projection error are consistently greater for SFD detection than for SIFT and A-KAZE. Approximately three times more points have less than 1 pixel error for SFD compared to SIFT and A-KAZE depicting the relatively high accuracy of the proposed method. This implies that taking the best N features from SFD will give higher accuracy calibration/reconstruction than for SIFT feature detection. Therefore SFD gives more accurate geometry estimation from wide-baseline views due to the improved accuracy of feature localization demonstrating the suitability of SFD for sparse 3D scene reconstruction.

Reconstruction accuracy evaluation: The accuracy and the suitability of features for wide-baseline reconstruction is evaluated for complex environments on a variety of datasets in this section. We measure the reconstruction accuracy evaluation (R_A) of SFD, defined as $R_A = \frac{\text{Correct Matches}}{\text{RC}}$ using the ground-truth information for Odzemok dataset. We eliminate the matches from RC with MRE greater that 2.5 pixels to obtain the 'Correct Matches', which is a standard setting to allow noise variance [53]. The comparisons with FAST, MSER, ORB, SIFT and A-KAZE are shown in Figure 12 for dynamic and in Figure 13 for static indoor and outdoor datasets. On left results are shown between testing images 1-2, 1-3, ..., 1-7 with baseline 15°-120°and on the right for adjacent image pairs with baseline 15°-30°.

The reconstruction accuracy evaluation of SIFT, A-KAZE and SFD detector is comparable and greater than other detectors like FAST, ORB and MSER. Watershed segmentation performed consistently better than other segmentation methods. As the baseline between the image pairs increases, the overlap between the images reduce which results in decrease in the number of matches (RC). It is noted that the reconstruction accuracy for each feature detector reduces with the increase in baseline which indicates a drop in the percentage of correct matches from the set of matches (RC). The drop in the reconstruction accuracy is similar for SFD, SIFT, A-KAZE and MSER. However, the percentage of correct matches for SFD is slightly higher than existing approaches, as seen from the Figure. The FAST and ORB detectors does not perform well for wide-baseline images.

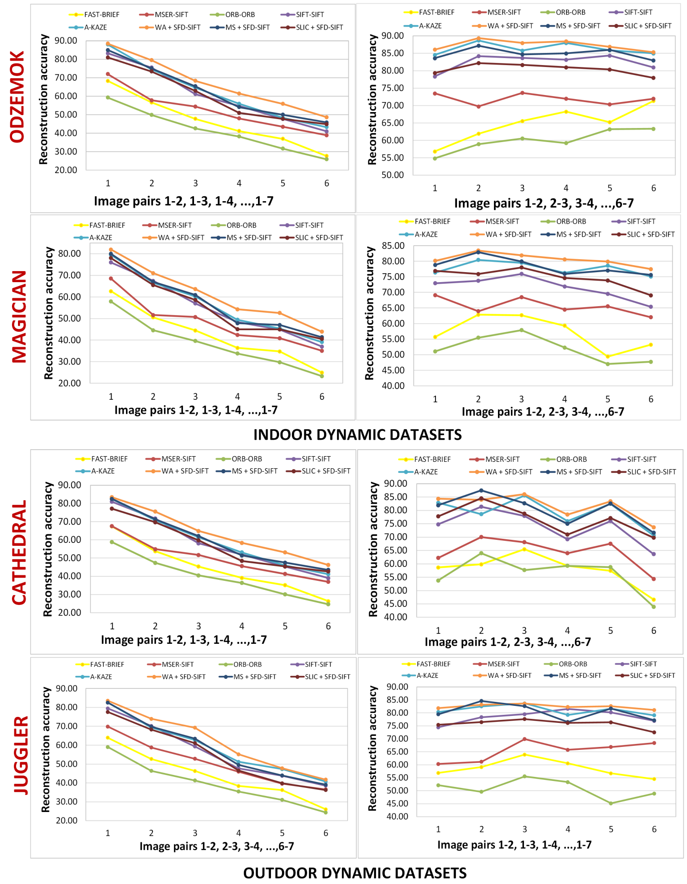


Fig. 12. Reconstruction accuracy evaluation results for dynamic datasets: Left: Comparison for matching of camera 1 to all other views (15-120 degree baseline); and Right: Comparison for matching between adjacent views (15-30 degree baseline).

Time performance: Figure 14 presents the average computation time/frame showing that the computational time is less than floating point detectors and similar to binary detectors. SFD-WA is the fastest detector compared to MS and SLIC, but the number of correct matches are highest for SLIC. MS gives lower number of correct matches compared to both WA and SLIC. The evaluation shows a trade-off between the performance and the number of correct matches for various segmentation techniques. The matching performance of detectors varies with the descriptor assignment. SFD works better with SIFT descriptor compared to BRIEF descriptor. This is expected as SFD is a floating point detector.

B. Multi-scale feature detection evaluation

Multi-scale evaluation is performed by feature detection and matching between pairs of images at difference scales for datasets from each category. The datasets are selected randomly for experimentation. Number of accurate matches (RC) and the percentage (R) of matches is evaluated, such that

$R = 100 \times \frac{\text{Matches from one image at original scale and other downscaled}}{\text{Matches from pair of images at original scale}}$, where $S1 = 0.25$ or $S2 = 2$. The results are shown in

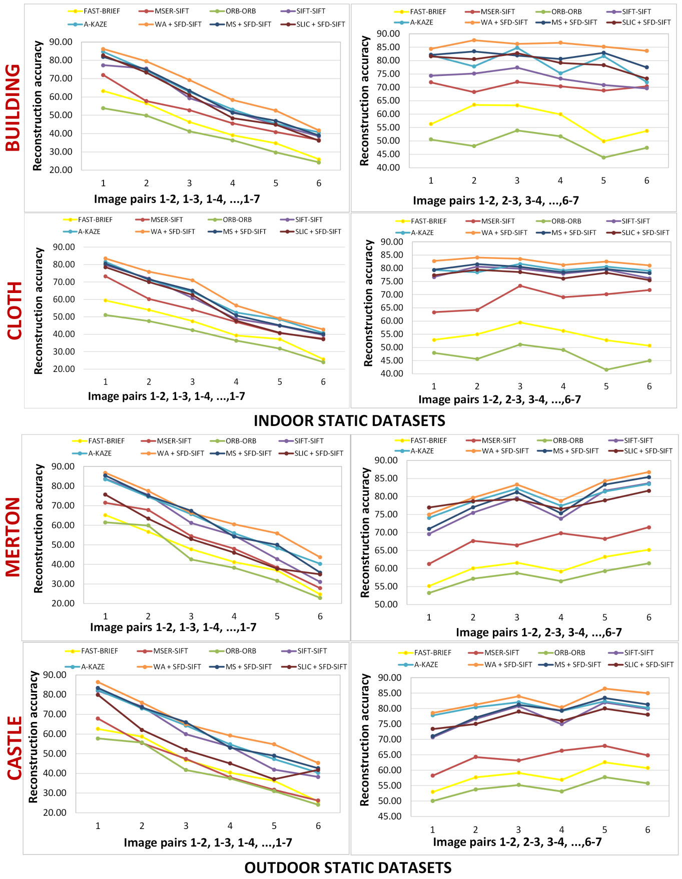


Fig. 13. Reconstruction accuracy evaluation results for static datasets: Left: Comparison for matching of camera 1 to all other views (15-120 degree baseline); and Right: Comparison for matching between adjacent views (15-30 degree baseline).

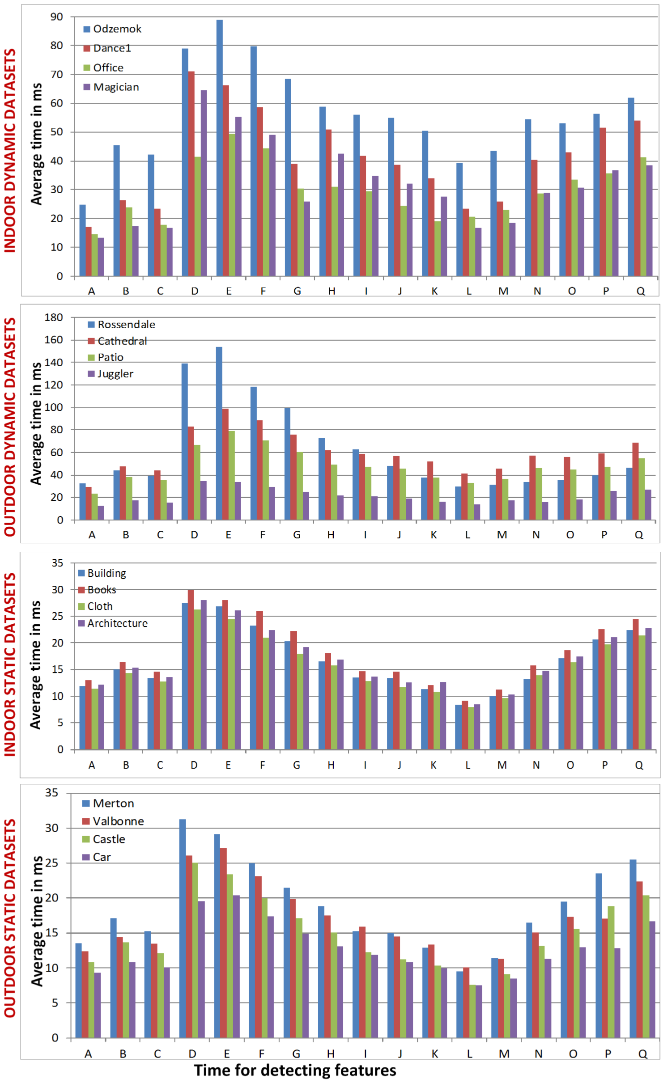


Fig. 14. Evaluation of time for detecting features on a wide-baseline stereo pair for each sequence in ms for all datasets.

		Odzemok	Magician	Patio	Juggler	Building	Books	Merton	Valb
1	S1	952	605	640	151	513	596	979	497
	R	**75.1**	**76.2**	76.9	72.5	**81.5**	76.1	76.9	**78.1**
	S2	1072	658	692	166	547	642	1064	538
	R	84.4	83.7	83.1	79.8	**86.9**	81.9	83.6	84.5
2	S1	901	579	921	517	551	631	1149	558
	R	74.6	75.9	76.3	**75.5**	79.8	75.9	76.2	74.1
	S2	1027	645	1019	583	581	693	1284	611
	R	85.2	**84.5**	83.8	84.9	84.2	**83.4**	85.1	80.9
3	S1	1044	965	1012	797	851	910	1232	344
	R	28.0	33.9	30.9	34.0	42.9	45.0	27.1	30.3
	S2	1851	1354	1471	1174	958	1033	2189	586
	R	49.8	47.6	44.9	50.1	48.3	51.2	48.3	51.6
4	S1	**2896**	**2301**	**2972**	**1894**	**1804**	**1855**	**3681**	**1178**
	R	73.9	75.6	**83.6**	74.8	81.3	**78.2**	**79.4**	76.3
	S2	**3419**	**2561**	**3015**	**2264**	**1919**	**1967**	**4110**	**1337**
	R	**87.3**	84.1	**84.9**	**89.3**	86.4	82.9	**88.7**	**86.6**

TABLE VI
MULTI-SCALE FEATURE DETECTION EVALUATION SHOWING NUMBER OF FEATURE MATCHES AND R IN % AT SCALE S1 = 0.25 AND S2 = 2 FOR DIFFERENT DETECTORS: 1. SIFT, 2. A-KAZE, 3. SFD-WA AND 4. MSFD-WA.

Table VI for two datasets from each category. The number of correct matches are reduced with scale change compared to the original number of matches shown in Table V for respective datasets for all feature detection techniques (SIFT and A-KAZE by $\approx 76\%$, SFD-WA by $\approx 34\%$ and MSFD-WA by $\approx 78\%$ for scale $S1 = 0.25$) (SIFT and A-KAZE by $\approx 84\%$, SFD-WA by $\approx 49\%$ and MSFD-WA by $\approx 87\%$ for scale $S2 = 2$). There is a significant improvement in the number of matches and percentage of original matches (R) with scale change over SFD feature detection. Also, the number of correct matches obtained using Multi-scale SFD are approximately $2-3$ times higher compared to existing feature detection techniques (SIFT and A-KAZE) with approximately same percentage drop (R) in the matches.

DTU Robotics dataset [17]: We have evaluated MSFD-WA with SIFT descriptor against existing feature detectors (A-M) on the DTU Robotics dataset which consists of 60 sequences as in [32]. We have restricted the number of features to 4000 for each detector and all sequences. The inlier ratio (RC/F) results are shown in Figure 15 and the matching score $(MS = TC/F)$ results are shown in Figure 16 demonstrating the improved performance using MSFD against state-of-the-art feature detectors.

C. Application to Wide-baseline Reconstruction

Wide-baseline sparse scene reconstructions are presented for all the datasets in Figure 17. Reconstructions obtained using the proposed SFD features are compared with those obtained using the SIFT and A-KAZE detectors, in all cases the SIFT descriptor is used for matching. As expected from

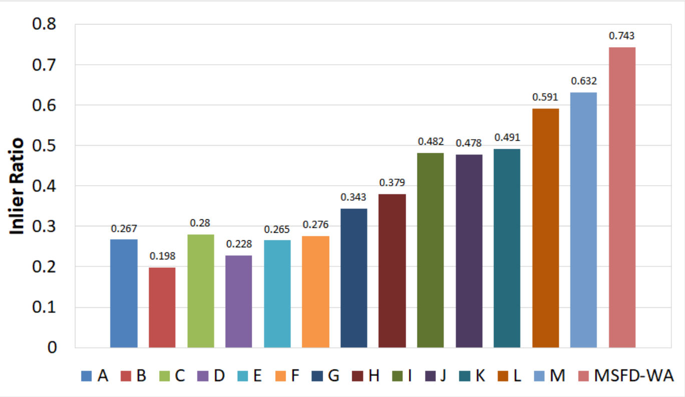


Fig. 15. Inlier-ratio comparison for DTU dataset for proposed MSFD and SFD detector against existing detectors.

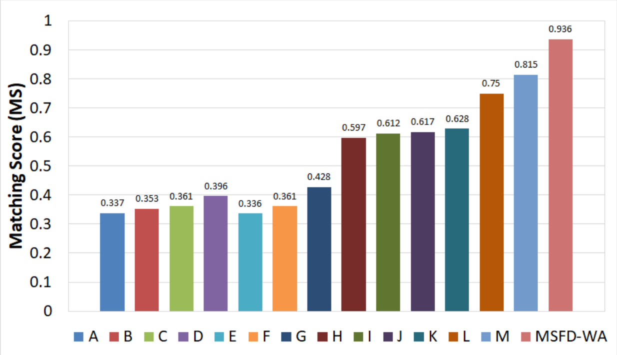


Fig. 16. Matching score comparison for DTU dataset for proposed MSFD and SFD detector against existing detectors.

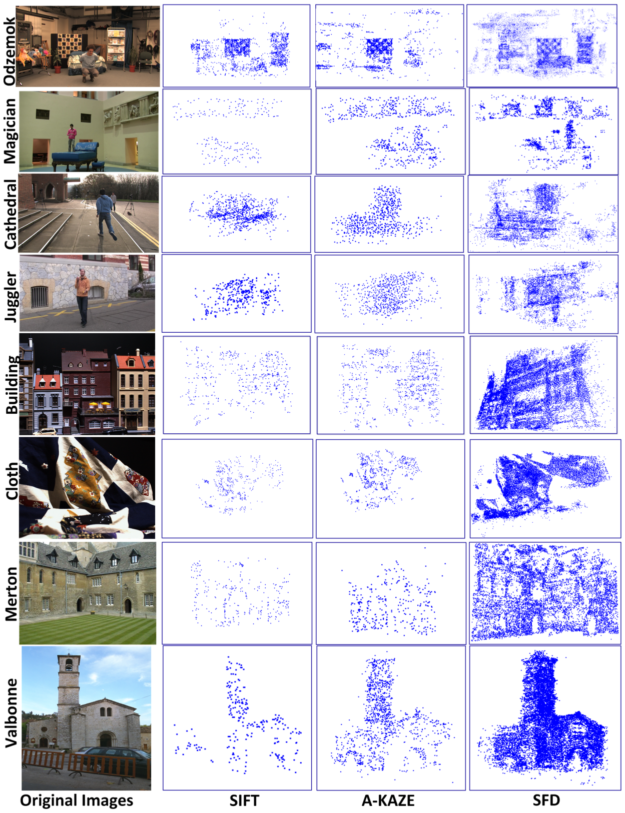


Fig. 17. Results of multi-view sparse reconstruction for all datasets for SIFT, A-KAZE and SFD-WA .

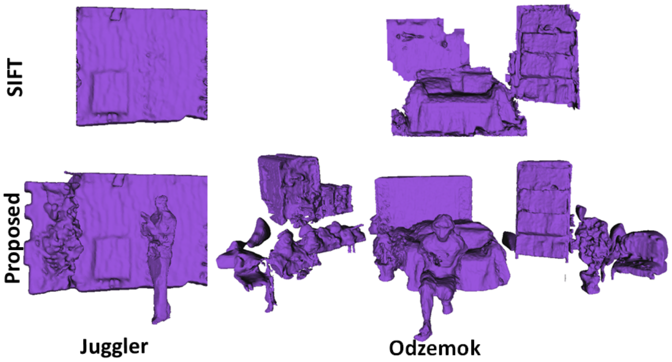


Fig. 18. Results of multi-view dense reconstruction for Odzemok and Juggler datasets for SIFT and SFD-WA .

the evaluation of wide-baseline matching presented above the number of reconstructed points is much higher with the proposed approach as shown in Table VII with WA, MS and SLIC. From Figure 17 it can be observed that sparse wide-baseline reconstruction based on SFD gives a significantly more complete representation of the scene (evaluation of the accuracy against ground-truth reconstruction for all datasets was presented in Table V).

Dataset	MSER	SIFT	A-KAZE	SFD-WA	SFD-MS	SFD-SLIC
Odzemok	171	1884	4025	12385	9087	**14515**
Dance1	153	1652	3599	**13603**	8026	11302
Office	165	1792	3806	11681	8034	**14109**
Magician	128	1171	2544	**9470**	7014	9360
Rossen.	58	526	1145	2213	1017	**3983**
Cathedral	72	2153	2570	10840	9733	**12895**
Patio	61	1847	2679	**8845**	7261	7259
Juggler	67	461	1522	7211	6501	**8102**
Building	249	2788	2984	**8606**	6939	7660
Books	312	3398	3610	**8762**	7009	8610
Cloth	175	2760	3667	7516	5958	**7634**
Archi.	210	3107	3929	**7725**	6019	7178
Merton	316	2760	3274	9619	8118	**10965**
Valbonne	258	1380	1637	4084	3369	**5121**
Castle	261	1508	1811	**5368**	4333	4854
Car	252	2208	2619	**7705**	6755	7184

TABLE VII
EVALUATION OF THE NUMBER OF SPARSE 3D POINTS FROM PAIR-WISE RECONSTRUCTION

Dataset	MSER	SIFT	A-KAZE	SFD-WA	SFD-MS	SFD-SLIC
Odzemok	1197	9246	16163	24876	22474	**26961**
Juggler	335	2535	4849	11277	10918	**13289**
Building	1743	12910	17573	**23218**	22596	22842
Valbonne	2542	12320	15263	27563	25546	**29839**

TABLE VIII
EVALUATION OF THE NUMBER OF SPARSE 3D POINTS FROM VSFM WITH MSER, SIFT, A-KAZE AND SFD ON ONE DATASET FROM EACH CATEGORY ON ALL VIEWS EXCEPT FOR BUILDING, WHERE 16 WIDE-BASELINE VIEWS ARE SELECTED.

For further evaluations we replaced the feature detection in the Visual SFM [58] pipeline with MSER, SIFT, A-KAZE and SFD features. The number of reconstructed points is shown in Table VIII. We have also evaluated the dense reconstruction obtained using SIFT and SFD detectors on a standard pipeline, results are shown in Figure 18.

The feature detection thresholds of the different methods are set to detect approximately the same number of features per image initially. However, the number of feature matches and sparse 3D points is much lower for SIFT and A-KAZE compared to SFD, showing the stability of SFD feature points. Hence, the SFD based dense reconstruction gives more complete coverage of scene compared to other detectors. Dense reconstruction and registration based on SFD features is demonstrated in [59] and [51] respectively for challenging datasets.

D. Limitations

Evaluation has been performed across a wide-variety of indoor and outdoor scenes to identify the limitations of SFD feature detection in the context of wide-baseline matching. As with other feature detection approaches the method is dependent on variation in surface appearance and consequently will produce fewer and less reliable features in areas of uniform appearance, or repetitive background texture like trees, sky etc. However, as demonstrated in the evaluation SFD increases the number of features and scene coverage for wide-baseline matching compared to previous approaches.

VII. Conclusion

In this paper we have proposed a novel multi-scale feature detector MSFD for wide-baseline matching and sparse scene reconstruction. The approach is based on over-segmentation of the scene and detecting features at intersections of three or more region boundaries. This approach is demonstrated to give stable feature detection across wide-baseline views with an increased number of features and more complete scene coverage than popular feature detectors used in wide-baseline applications. MSFD is shown to give consistent performance for different segmentation approaches (Watershed, Mean shift, SLIC), with SFD-SLIC giving a marginally higher number of features. The speed of SFD feature detection is comparable to other methods for wide-baseline matching. A multi-scale segmentation based feature detection is introduced to achieve scale invariance giving improved performance against existing feature detection techniques.

A comprehensive performance evaluation against previous feature detectors (Harris, SIFT, SURF, FAST, ORB, MSER, KAZE, A-KAZE) in combination with widely used feature descriptors (SIFT, BRIEF, ORB, SURF) demonstrates that the proposed multi-scale segmentation based feature detector MSFD achieves a factor of $3-10$ times more wide-baseline feature matches for a variety of indoor and outdoor scenes. Quantitative evaluation against ground-truth of SFD vs. SIFT, MSER, and A-KAZE feature detectors shows that for a given error level MSFD gives a significantly larger number of features. Improved accuracy in feature localisation with SFD results in more accurate camera calibration and reconstruction of sparse scene geometry.

Application to stereo sparse reconstruction from wide-baseline camera views demonstrates that the MSFD feature detector combined with a SIFT descriptor achieves a significant increase in the number or reconstructed points and more complete scene coverage than SIFT detection. Further plans include evaluating the utility of MSFD features in applications such as camera tracking and object recognition and the integration with deep learning approaches to detection and matching to achieve greater generalisation across scenes while maintaining or improving performance.

Acknowledgment

This research was supported by the Royal Academy of Engineering Research Fellowship RF-201718-17177, the European Commission, FP7 Intelligent Management Platform for Advanced Real-time Media Processes project (grant 316564) and EPSRC Platform Grant on Audio-Visual Media Research EP/P022529.

References

[1] C. Harris and M. Stephens, "A combined corner and edge detector," in *Alvey Vision Conf.*, 1988, pp. 147–151.

[2] K. Haris, S. N. Efstratiadis, N. Maglaveras, and A. K. Katsaggelos, "Hybrid image segmentation using watersheds and fast region merging," *IEEE Trans. Image Process.*, vol. 7, no. 6, pp. 1684–1699, 1998.

[3] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *IJCV*, vol. 60, no. 2, pp. 91–110, 2004.

[4] H. Bay, T. Tuytelaars, and L. Gool, "Surf: Speeded up robust features," in *ECCV*, 2006, pp. 404–417.

[5] E. Rosten and T. Drummond, "Fusing points and lines for high performance tracking." in *ICCV*, 2005, pp. 1508–1511.

[6] P. F. Alcantarilla, A. Bartoli, and A. J. Davison, "KAZE Features." in *ECCV*, 2012, pp. 214–227.

[7] J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust wide baseline stereo from maximally stable extremal regions," in *BMVC*, 2002, pp. 36.1–36.10.

[8] M. Agrawal, K. Konolige, and M. Blas, "Censure: Center surround extremas for realtime feature detection and matching," in *ECCV*, 2008, pp. 102–115.

[9] A. Mustafa, H. Kim, E. Imre, and A. Hilton, "Segmentation based features for wide-baseline multi-view reconstruction," in *3DV*, 2015.

[10] H. Moravec, "Obstacle avoidance and navigation in the real world by a seeing robot rover," Tech. Rep., 1980.

[11] E. Rosten, R. Porter, and T. Drummond, "Faster and better: A machine learning approach to corner detection," *TPAMI*, vol. 32, no. 1, pp. 105–119, 2010.

[12] M. A. Föstner and E. Gülch, "A Fast Operator for Detection and Precise Location of Distinct Points, Corners and Centers of Circular Features," in *ISPRS*, 1987.

[13] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. V. Gool, "A comparison of affine region detectors," *IJCV*, vol. 65, no. 1-2, pp. 43–72, 2005.

[14] K. Mikolajczyk and C. Schmid, "Scale &amp; affine invariant interest point detectors," *IJCV*, vol. 60, pp. 63–86, 2004.

[15] D. Weng, Y. Wang, M. Gong, D. Tao, H. Wei, and D. Huang, "DERF: distinctive efficient robust features from the biological modeling of the P ganglion cells," *IEEE Trans. Image Process.*, vol. 24, no. 8, pp. 2287–2302, 2015.

[16] D. C. Hauagge and N. Snavely, "Image matching using local symmetry features," in *CVPR*, 2012, pp. 206–213.

[17] H. Aanæs, A. L. Dahl, and K. S. Pedersen, "Interesting Interest Points - A Comparative Study of Interest Point Performance on a Unique Data Set." *IJCV*, vol. 97, no. 1, pp. 18–35, 2012.

[18] P. F. Alcantarilla, J. Nuevo, and A. Bartoli, "Fast Explicit Diffusion for Accelerated Features in Nonlinear Scale Spaces," in *BMVC*, 2013.

[19] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "Orb: An efficient alternative to sift or surf," in *ICCV*, 2011, pp. 2564–2571.

[20] S. Leutenegger, M. Chli, and R. Y. Siegwart, "Brisk: Binary robust invariant scalable keypoints," in *ICCV*, 2011, pp. 2548–2555.

[21] V. Lepetit and P. Fua, "Keypoint recognition using randomized trees," *TPAMI*, vol. 28, no. 9, pp. 1465–1479, 2006.

[22] T. F. Chan and L. A. Vese, "Active contours without edges," *IEEE Trans. Image Process.*, vol. 10, no. 2, pp. 266–277, 2001.

[23] F. Mokhtarian and R. Suomela, "Robust Image Corner Detection Through Curvature Scale Space." *TPAMI*, vol. 20, no. 12, pp. 1376–1381, 1998.

[24] X. Zhang, H. Wang, A. W. B. Smith, L. Xu, B. C. Lovell, and D. Yang, "Corner detection based on gradient correlation matrices of planar curves." *Pattern Recognit.*, vol. 43, no. 4, pp. 1207–1223, 2010.

[25] C. Vicas and S. Nedevschi, “Detecting curvilinear features using structure tensors,” *IEEE Trans. Image Processing*, vol. 24, no. 11, pp. 3874–3887, 2015.

[26] M. Awrangjeb, G. Lu, and C. S. Fraser, “Performance Comparisons of Contour-Based Corner Detectors.” *IEEE Trans. Image Process.*, vol. 21, no. 4, pp. 4167–4179, 2012.

[27] M. Maire, P. Arbelaez, C. Fowlkes, and J. Malik, “Using contours to detect and localize junctions in natural images,” in *CVPR*, 2008, pp. 1–8.

[28] S. K. Ravindran and A. Mittal, “Comal: Good features to match on object boundaries,” in *CVPR*, June 2016.

[29] A. Richardson and E. Olson, “Learning convolutional filters for interest point detection,” in *ICRA*, 2013.

[30] W. Hartmann, M. Havlena, and K. Schindler, “Predicting matchability,” in *CVPR*, 2014, pp. 9–16.

[31] Y. Verdie, K. Moo Yi, Y. Verdie, P. Fua, and V. Lepetit, “Tilde: A temporally invariant learned detector.” in *CVPR*, 2015, pp. 5279–5288.

[32] K. Yi, E. Trulls, V. Lepetit, and P. Fua, “Lift: Learned invariant feature transform,” in *ECCV*, 2016, pp. 467–483.

[33] M. Paulin, M. Douze, Z. Harchaoui, J. Mairal, F. Perronin, and C. Schmid, “Local convolutional features with unsupervised training for image retrieval,” in *ICCV*, 2015, pp. 91–99.

[34] F. X. Yu, A. T. Suresh, K. Choromanski, D. Holtmann-Rice, and S. Kumar, “Orthogonal random features,” in *NIPS*, 2016, pp. 1975–1983.

[35] E. Rosten and T. Drummond, *Machine Learning for High-Speed Corner Detection*, 2006, pp. 430–443.

[36] Y. Duan, J. Lu, J. Feng, and J. Zhou, “Learning rotation-invariant local binary descriptor,” *IEEE Trans. Image Process.*, vol. 26, pp. 3636–3651, 2017.

[37] X. Tan and B. Triggs, “Enhanced local texture feature sets for face recognition under difficult lighting conditions,” *IEEE Trans. Image Process.*, vol. 19, no. 6, pp. 1635–1650, 2010.

[38] Y. Duan, J. Lu, J. Feng, and J. Zhou, “Context-aware local binary feature learning for face recognition,” *TPAMI*, vol. 40, no. 5, pp. 1139–1153, May 2018.

[39] V. Balntas, L. Tang, and K. Mikolajczyk, “Binary online learned descriptors,” *TPAMI*, vol. 40, no. 3, pp. 555–567, March 2018.

[40] J. Kim and K. Grauman, “Boundary preserving dense local regions,” in *CVPR*, 2011, pp. 1153–1560.

[41] P. Koniusz and Mikolajczyk, “Segmentation based interest points and evaluation of unsupervised image segmentation methods,” in *BMVC*, 2009.

[42] J.-Y. Guillemaut and A. Hilton, “Space-time joint multi-layer segmentation and depth estimation,” in *International Conference on 3D Imaging, Modeling, Processing, Visualization and Transmission (3DIMPVT)*, 2012, pp. 440–447.

[43] N. Snavely, S. M. Seitz, and R. Szeliski, “Modeling the world from internet photo collections,” *IJCV*, vol. 80, no. 2, pp. 189–210, 2008.

[44] K. Levenberg, “A method for the solution of certain non-linear problems in least squares,” *Q. J. Math*, vol. 2, no. 2, pp. 164–168, 1944.

[45] S. Gauglitz, T. Höllerer, and M. Turk, “Evaluation of interest point detectors and feature descriptors for visual tracking,” *IJCV*, vol. 94, no. 3, pp. 335–360, 2011.

[46] J. B. Roerdink and A. Meijster, “The watershed transform: Definitions, algorithms and parallelization strategies,” *Fundam. Inf.*, vol. 41, no. 3, pp. 187–228, 2000.

[47] F. Meyer, “An overview of morphological segmentation,” *Int. J. Pattern Recogn.*, vol. 15, no. 2, pp. 1089–1118, 2001.

[48] P. Perona and J. Malik, “Scale-space and edge detection using anisotropic diffusion,” *TPAMI*, vol. 12, no. 7, pp. 629–639, 1990.

[49] D. Comaniciu and P. Meer, “Mean shift: a robust approach toward feature space analysis,” *TPAMI*, vol. 24, no. 4, pp. 603–619, 2002.

[50] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk, “Slic superpixels compared to state-of-the-art superpixel methods,” *TPAMI*, vol. 34, no. 11, pp. 2274–2282, 2012.

[51] A. Mustafa, H. Kim, and A. Hilton, “4d match trees for non-rigid surface alignment,” in *ECCV*, 2016.

[52] R. Fattal, M. Agrawala, and S. Rusinkiewicz, “Multiscale shape and detail enhancement from multi-light image collections,” *ACM Trans. Graph.*, vol. 26, 2007.

[53] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, 2nd ed. Cambridge University Press, 2003.

[54] T. Tuytelaars and K. Mikolajczyk, “Local invariant feature detectors: A survey,” *Foundation and Trends in Computer Graph. and Vis.*, vol. 3, no. 3, pp. 177–280, 2008.

[55] M. Calonder, V. Lepetit, C. Strecha, and P. Fua, “Brief: Binary robust independent elementary features,” in *ECCV*, 2010, pp. 778–792.

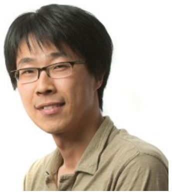
[56] J. Shi and C. Tomasi, “Good features to track,” in *CVPR*, 1994, pp. 593–600.

[57] P. Pritchett and A. Zisserman, “Wide baseline stereo matching,” in *ICCV*, 1998, pp. 754–760.

[58] Y. Furukawa and J. Ponce, “Accurate, dense, and robust multiview stereopsis,” *TPAMI*, vol. 32, no. 8, pp. 1362–1376, 2010.

[59] A. Mustafa and A. Hilton, “Semantically coherent co-segmentation and reconstruction of dynamic scenes,” in *CVPR*, 2017.

Armin Mustafa is a Royal Academy of Engineering Fellow at the University of Surrey. She received the Ph.D. degree from the Centre for Vision, Speech, and Signal Processing (CVSSP), University of Surrey in 2016, where she is currently a Research Fellow. She previously worked at Samsung Research Institute, Bangalore, India for 3 years (2010 - 2013) in Computer Vision and she received the MTech degree from Indian Institute of Technology, Kanpur, India in 2010. Her research interests include 3D/4D Computer Vision and Scene Understanding.



Hansung Kim received the MS and Ph.D degrees in electronic and electrical engineering from Yonsei University, Seoul, Korea, in 2001 and 2005, respectively. He was employed as a Researcher of Knowledge Science Lab (KSL) at Advanced Telecommunications Research Institute International (ATR), Japan, from 2005 to 2008. He is currently a Research Fellow (RA2) at CVSSP, University of Surrey. His research interests include 3-D computer vision, multi-modal data processing, audio-visual data processing and media production.

Adrian Hilton is a Professor of Computer Vision and Graphics and Director of the CVSSP at the University of Surrey. He leads the Visual Media Research (VLab) in CVSSP which is conducting research in video analysis, computer vision and graphics for next generation communication and entertainment applications. He received B.S. (Hons.) and D.Phil. degrees from the University of Sussex in 1988 and 1992, respectively. His research interests include robust computer vision to model and understand real world scenes.