Localization Accuracy of Interest Point Detectors with Different Scale Space Representations

Kai Cordes, Bodo Rosenhahn, Jörn Ostermann
Institut für Informationsverarbeitung (TNT)
Leibniz Universität Hannover, Germany
{cordes, rosenhahn, ostermann}@tnt.uni-hannover.de

Abstract

The detection of scale invariant image features is a fundamental task for computer vision applications like object recognition or re-identification. Features are localized by computing extrema of the gradients in the Laplacian of Gaussian (LoG) scale space. The most popular detector for scale invariant features is the SIFT detector which uses the Difference of Gaussians (DoG) pyramid as an approximation of the LoG. Recently, the alternative interest point (ALP) detector demonstrated its strength in fast computation on highly parallel architectures like the GPU. It uses the LoG scale space representation for the localization of interest points. This paper evaluates the localization accuracy of ALP in comparison to SIFT.

By using synthetic images, it is demonstrated that both localization approaches show a systematic error which is dependent on the subpixel position of the feature. The error increases with the scale of the detected feature. However, using the LoG instead of the DoG representation reduces the maximum systematic error by 77 %. For the evaluation with natural images, benchmark data sets are used. The repeatability criterion evaluates the accuracy of the detectors. The LoG based detector results in up to 16 % higher repeatability. The comparisons are completed with a reference feature localization which uses a signal based approach for the gradient approximation. Based on this approach, a new feature selection criterion is proposed.

1. Introduction

Scale invariant features play an important role in many computer vision applications and surveillance tasks, such as scene reconstruction [13] or object re-identification [1]. Examples for scale invariant features detected by different methods are shown in Fig. 1. The *scale invariant feature transform* (SIFT) technique [10] provides very good results for feature localization and matching. The detected features

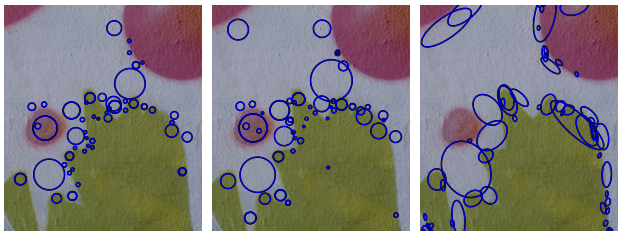


Figure 1: Scale invariant features detected by SIFT, ALP, and HALF SIFT (from left to right).

are robust to changes in illumination, rotation, scale, and to surprisingly large viewpoint changes.

Scale invariant features are detected as scale space extrema. The scale space [9] is represented by a *Difference of Gaussians* (DoG) pyramid, which is an approximation of the Laplacian of Gaussian (LoG) pyramid. The features are localized by an approximation of the gradients in the scale space using a 3D quadratic function [2]. However, the interpolated gradient signal does not have the shape of a 3D quadratic [3]. In [8], Lindeberg characterizes the neighborhood of interest points as blob-like structures with the shape of a Gaussian.

An alternative to SIFT, called *alternative interest point* (ALP) detector [6], is designed for the usage in MPEG. An important reason for its development is that the usage of SIFT is restricted due to its patent [11]. Thus, the commercial utilization is limited. The ALP detector follows a similar approach, but uses different tools for scale invariant feature extraction. Additionally, it is suited for highly parallel architectures like the GPU. The main application is the identification of objects captured with a mobile device.

Although the processing pipeline is similar to SIFT, the ALP detector changes the feature localization task slightly. Instead of the DoG pyramid, the ALP detector builds the LoG pyramid. This should lead to better localization accuracy than using the DoG pyramid. An image feature is localized with subpixel accuracy which is obtained by an in-

terpolation of the gradient values of the scale space. The localization accuracy is mainly determined by the subpixel localization scheme. SIFT uses a three-dimensional quadratic function [2] for the gradient approximation. The subpixel and subscale estimation of ALP differs from SIFT in two attributes. Firstly, the scale estimation is decomposed from the subpixel estimation. Secondly, the subscale is computed as maxima of a polynomial of third degree which approximates the LoG scale space. The subpixel coordinates are then determined as the maximum of a two-dimensional quadratic. The question arises if the localization accuracy is still comparable to the accuracy of the original SIFT method. As shown in [3], the quadratic function is only a coarse approximation of the gradient signal, which leads to a systematic error. Instead of the 3D quadratic function, a signal adapted approximation function for the DoG gradients is proposed. This approach, called HALF SIFT (*highly accurate localized features*) eliminates the systematic error for the input signal of Gaussian shaped features.

We evaluate the systematic error of the LoG based ALP detector. The evaluation compares the systematic errors of ALP and SIFT. For the evaluation, synthetic test signals as well as natural image data are used. The synthetic images consist of one Gaussian feature blob with a priori known subpixel and subscale parameters. Gaussian features are introduced by Lindeberg [8] for the analysis of scale invariant feature detectors. For the comparison using natural images, the repeatability criterion is used [12]. The repeatability measure thresholds the overlap error to decide if detected feature pairs in two images are correct and counts valid feature pairs. In addition to the images from [12], a new, highly accurate and high-resolution benchmark data set is employed [5]. The **contributions** of this paper are:

- the analysis of the localization accuracies of DoG and LoG based detectors
- the comparison of the localization accuracy using natural images from standard benchmarks
- the proposal of a new feature selection criterion

In the following Section 2, the evaluated localization techniques are briefly explained. In Section 3, evaluation methods and the data used for the experiments are derived. Section 4 shows the results on synthetic and natural image data. In Section 5, the paper is concluded.

2. Scale invariant Feature Detection

Methods for the detection of scale invariant features evaluate the scale space as proposed by Lindeberg [9]. A scale space representation is determined by a sequence of Laplacian of Gaussian (LoG) filters. While the ALP [6] detector uses the LoG pyramid, the SIFT detector uses the Difference of Gaussian (DoG) approximation [10]. The determination of subpixel and subscale coordinates of these approaches are explained in the next sections. The SIFT and ALP detectors can be downloaded from the internet [1]. The implementations are called TM7 (*test model 7.0*) for SIFT and TM8 (*test model 8.0*) for ALP. The HALF SIFT (*highly accurate localized features*) approach [3] extends SIFT with a signal adapted localization procedure which assumes a bivariate Gaussian shape for the neighborhood of a feature.

2.1. Feature Localization of SIFT / TM7

The scale space maxima are detected by evaluating, if a pixel value in each DoG scale is bigger or smaller that its 26 neighbors in the $3 \times 3 \times 3$ neighborhood. The subpixel and subscale coordinates $\mathbf{x} = (x, y, s)$ of a feature are found by interpolation with a 3D quadratic [2] function $D^{\text{QUAD}}(\mathbf{x})$. This function has the shape of a parabola in each of the three dimensions:

$$D^{\text{QUAD}}(\mathbf{x}) = D(\mathbf{x}_0) + \frac{\partial D(\mathbf{x}_0)}{\partial \mathbf{x}}^\top \mathbf{x}^\top + \frac{1}{2}\mathbf{x}^\top \frac{\partial^2 D(\mathbf{x}_0)}{\partial \mathbf{x}^2}\mathbf{x} \quad (1)$$

Here, $D(\mathbf{x}_0)$ is the DoG value at the fullpixel sample point $\mathbf{x}_0 = (x_0, y_0, s_0)$. The extremum of $D(\mathbf{x})$ determines the subpixel and subscale localization. It is calculated with the inverse of the Hessian Matrix using the 27 sample points.

2.2. Feature Localization of HALF SIFT

The feature localization of HALF SIFT [3] incorporates a signal model. With this model the neighborhood of an input feature $\mathbf{x} = (x, y)$ is described by a bivariate Gaussian distribution G_{Σ_f}. Thus, the output D^{DoG} of the Difference of Gaussian (DoG) filter has again DoG shape,

$$D^{\text{DoG}}(\mathbf{x}) \quad = \quad l \cdot (G_{\Sigma_\sigma}(\mathbf{x}) - G_{\Sigma_{k\sigma}}(\mathbf{x})) * G_{\Sigma_f}(\mathbf{x}), \quad (2)$$

with $\Sigma_\sigma = \begin{pmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{pmatrix}$ and the standard deviation σ of the current scale s. The a priori known k is the distance between two scales. The function (2) of each feature is determined with six parameters for spatial localization $\mathbf{x} = (x, y)$, covariance Σ_σ, and amplitude l as the minimum of the residuum ϵ^{DoG}:

$$\epsilon^{\text{DoG}} = \sum_{\mathbf{x} \in \mathcal{N}} (D^{\text{DoG}}(\mathbf{x}) - D(\mathbf{x}))^2 \quad . \quad (3)$$

As before, 27 sample points of the $3 \times 3 \times 3$ neighborhood $\mathcal{N}$ are incorporated. The parameters are found by Levenberg-Marquardt optimization. As shown in [3], this feature localization approach eliminates the systematic error for features with Gaussian shape.

[1] http://pacific.tilab.com/projects/mpeg-cdvs/

(a) Graffiti (b) Grace (c) Underground (d) Posters (e) There (f) Colors

Figure 2: First images of the benchmark image sequences [5, 12] presented in the results section.

2.3. Feature Localization of ALP / TM8

The ALP detector computes subpixel and subscale coordinates of a feature in separated steps. The scale σ is determined by the initial feature detection procedure which computes the maxima of a polynomial of third degree for each pixel position in the LoG pyramid. The subpixel coordinates $\mathbf{x} = (x, y)$ are estimated by approximating the spatial gradients of this fixed scale with a 2D quadratic function (cf. eq. (1)).

2.4. Feature Selection

After the detection of feature candidates, a number of features is selected. This number is usually given by the user. Thus, a sorted list using a quality measure is generated, and the best n features are selected. The quality measure used in the implementations SIFT / TM7 and ALP / TM8 employs a set of predefined weights for scale, localization, orientation, peak, and curvature [6].

For the HALF SIFT approach, we propose to select features which provide the smallest values for the residuum (cf. eq. (3)), such that those features are selected with maximum similarity to a Gaussian function. The idea is to follow Lindeberg's assumption [8] that the features which are detected by a scale invariant feature detector have Gaussian shape.

3. Experimental Setup

The experiments are devided into two parts. The first part evaluates the synthetic images to reveal the systematic error for the detectors SIFT (TM7 implementation) and ALP (TM8 implementation). The second part uses natural images to show the effect on the repeatability criterion for the benchmarks [5] and [12]. As the latter proved to provide higher accuracy [4], we show the whole data set of [5] (2048×1365) and the most prominent sequence *Graffiti* (800×640) of [12].

In addition to SIFT and ALP, the HALF SIFT method [3] is evaluated in the repeatability comparison as a reference. This approach shows no systematic error on Gaussian input features and provides affine invariant features. It follows that it should perform better for strong viewpoint changes in comparison to SIFT and ALP.

The evaluated algorithms are listed in Tab. 1. The implementations for SIFT and ALP are TM7 and TM8, respectively [6]. The HALF SIFT approach extends the SIFT detector [7] with a new signal adapted localization procedure [3]. It exchanges the subpixel and subscale localization procedure as described in Sect. 2.2. Here, the feature selection scheme as proposed in Sect. 2.4 is used.

Table 1: The compared algorithms, their scale space representation, and their gradient approximation approach.

Detector	Scale Space	Gradient Approx.
SIFT / TM7 [6]	DoG	3D quadratic
ALP / TM8 [6]	LoG	separated spatial / scale
HALF SIFT [3]	DoG	DoG

3.1. Gaussian Feature Input Images

The synthetic images are constructed using a Gaussian distribution with a varying localization in x-direction and a varying standard deviation σ_f. The differences ϵ_x in x-direction are within the interval $[-2.0; 2.0]$ with a step distance of 0.04 px. The standard deviations σ_f of the Gaussians are within the interval $[2.6; 16.0]$ with a step distance of 0.4. With these ranges, the first three pyramid octaves are covered. The resulting error ξ_x^E is defined as the distance between the ground truth ϵ_x-coordinate and the detected x-coordinate of each feature: $\xi_x^E = \epsilon_x - x$. Here, the coordinates are measured relative to an image coordinate system with its origin in the center. Some input image examples are shown in Fig. 3. The image size is 256×256.

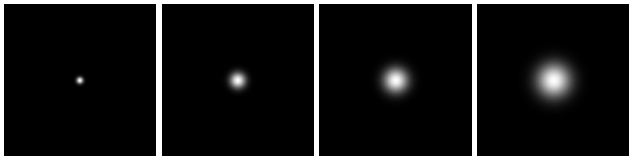
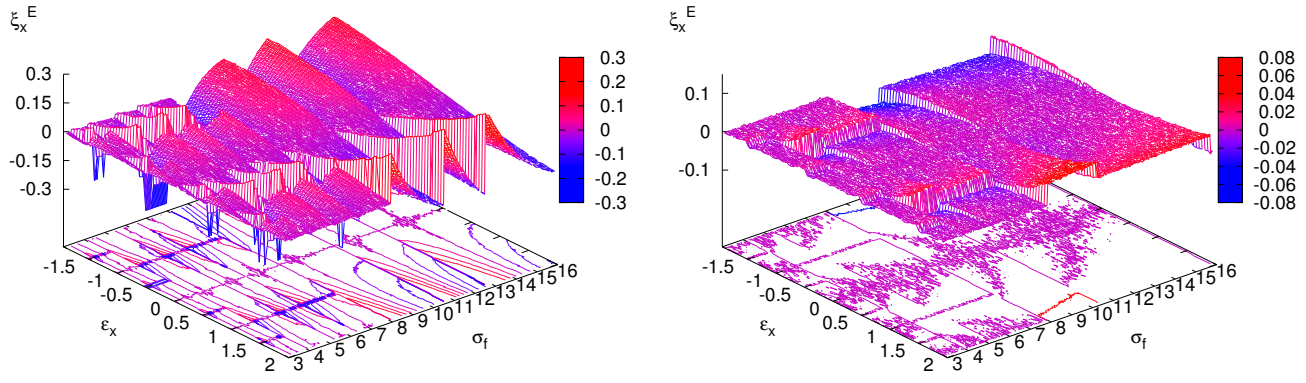


Figure 3: Image examples of the Gaussian features with varying standard deviation $\sigma_f = 3.0, 7.0, 11.0$, and 15.0.



(a) SIFT / TM7 (b) ALP / TM8

Figure 4: Spatial systematic error ξ_x^E in x-direction for Gaussian features. The axes in the ground plane show the ground truth subpixel position ϵ_x and the standard deviation σ_f of the input feature (cf. Fig. 3). The contour lines in the ground plane showcase the periodicity of the error.

Figure 5: Example benchmark sequence *Grace*.

3.2. Natural Input Images

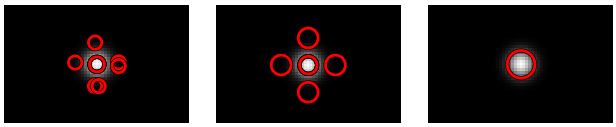
For the evaluation with natural images, two reference benchmark sets [5, 12] are used. The sets consist of sequences of images which show a planar scene captured from a varying viewpoint with increasing angle relative to the first image. An example sequence (*Grace*) is shown in Fig. 5. For the mapping from the first image to the others, ground truth homographies are provided. If the feature which is detected in the first image is also detected in second image, the feature pair is deemed correct. A threshold value $\epsilon_O = 0.4$ defines the maximally allowed distance between two features. This distance is calculated by the *overlap error* [12]. The repeatability is the ratio of correctly detected feature pairs to the maximally possible number of correct feature pairs. It is calculated by the *Matlab* script provided by the authors of [12].

A benchmark sequence consists of six images capturing the same planar scene. The viewpoint perspective increases with the image number. The repeatability is calculated between the first image and each of the other images. The resulting diagram provides a common measure for the localization accuracy of feature detectors [12]. The benchmark [12] provides image resolution of 0.5 megapixels while the data set [5] provides high resolution images.

To avoid the initial down sampling of the images for the natural data experiment, we set the parameter *resizeMaxSize* in the ALP / TM8 and SIFT / TM7 detector to 2048. As the ground truth data is prepared for the specific image sizes, a down sampling of the images would make it unusable. Furthermore, it would decrease the localization accuracy of ALP and SIFT significantly.

4. Experimental Results

The results with synthetic images are shown in Sect. 4.1. It demonstrates the systematic error of SIFT / TM7 and ALP / TM8. The repeatability results on natural images are shown in Sect. 4.2. It gives a comparison of SIFT / TM7 and ALP / TM8. Additionally, it shows the distance to the reference results of HALF SIFT.

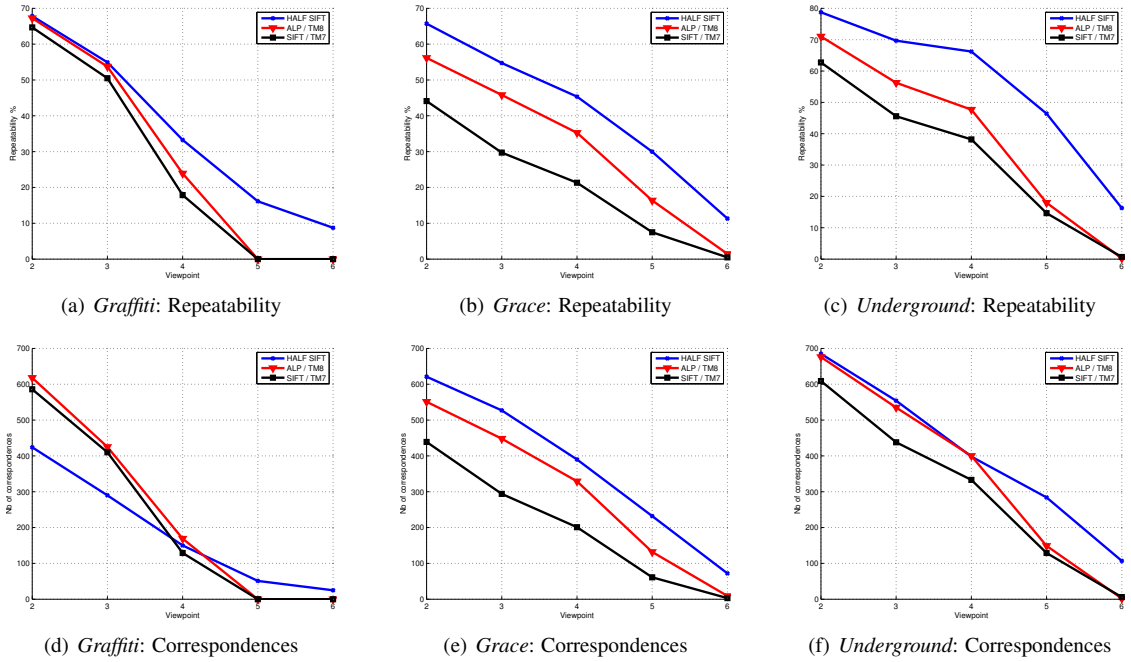


(a) SIFT / TM7 (b) ALP / TM8 (c) HALF SIFT

Figure 6: Results examples for a synthetic Gaussian blob feature. If more than on feature is extracted by the feature detector, the best of them is chosen for the evaluation.

4.1. Gaussian Feature Images

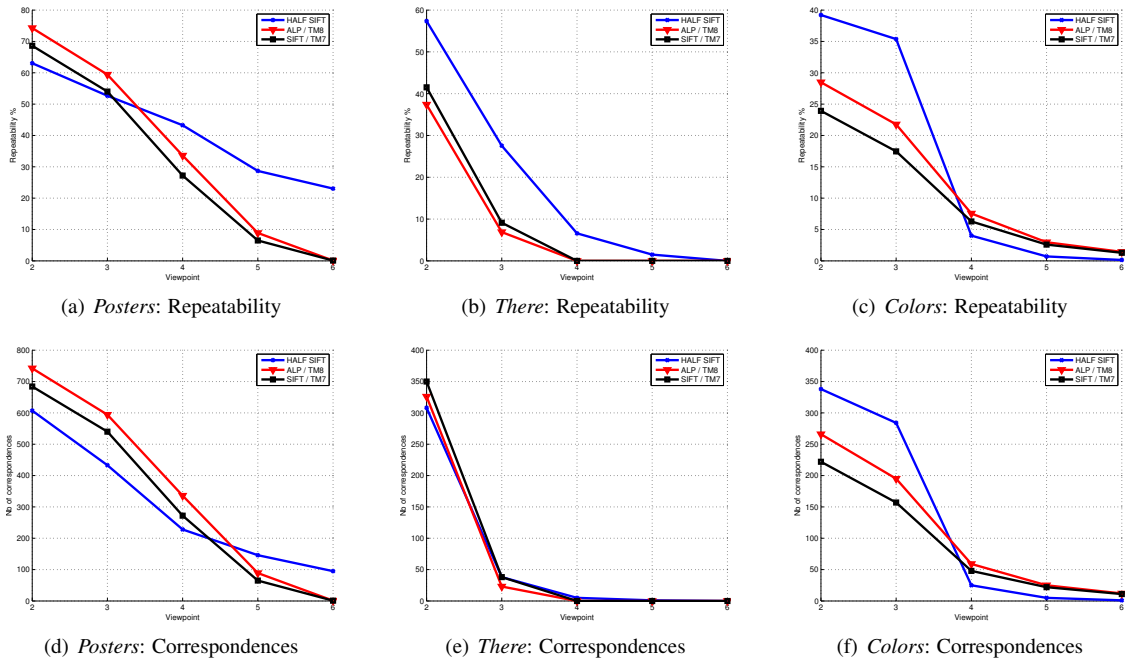
The resulting localization error in Fig. 4 shows the difference $\xi_x^E = \epsilon_x - x$ between ground truth ϵ_x and the detected x-position of the feature. Both localization approaches, SIFT and ALP lead to a systematic error, which increases with the octave of the input signal. The largest errors are located exactly between two pixels in the respective octave. For the first three octaves, the ALP method results in a max-



(a) *Graffiti*: Repeatability (b) *Grace*: Repeatability (c) *Underground*: Repeatability

(d) *Graffiti*: Correspondences (e) *Grace*: Correspondences (f) *Underground*: Correspondences

Figure 7: Repeatability (top row) and absolute number of correct feature pairs (bottom row) for *Graffiti* (800×640) [12], *Grace*, and *Underground* (2048×1365) [5] with 1000 selected features in each of the 6 images.



(a) *Posters*: Repeatability (b) *There*: Repeatability (c) *Colors*: Repeatability

(d) *Posters*: Correspondences (e) *There*: Correspondences (f) *Colors*: Correspondences

Figure 8: Repeatability (top row) and absolute number of correct feature pairs (bottom row) for *Posters*, *There*, and *Colors* (2048×1365) [5] with 1000 selected features in each of the 6 images.

imum absolute error of 0.07 px (for $\sigma_f = 10.4$) while SIFT shows a maximum error of 0.31 px (for $\sigma_f = 12.8$). The result verifies the systematic error of SIFT / TM7 found in [3] for the implementation of Hess [7]. Using the LoG instead

of the DoG decreases the maximum error of the subpixel localization significantly by 77 %.

4.2. Natural Images

The repeatability results for the benchmark sets as shown in Fig. 2 are demonstrated in Fig. 7 and Fig. 8. The repeatability rate (top row) of ALP is higher than for SIFT in all image pairs except for *There* (cf. Fig. 8(b)). Here, SIFT provides slightly better results. For the *Colors* sequence, the strong scale change leads to a feature selection in the images $4, \ldots, 6$, which is very different from the selection in the first image. Thus, all detectors perform very poor for these image pairs.

The repeatability rate of ALP is up to 16 % higher compared to SIFT. The absolute number of correct feature pairs is larger for most of the feature pairs (cf. Fig. 7 and Fig. 8, bottom row). The gain is mostly visible for the image pairs with smaller viewpoint changes. Overall, the best results are achieved for the signal based localization approach HALF SIFT, especially for large viewpoint changes. But, it requires more computation time [3].

5. Conclusions

This paper evaluates and compares the localization accuracies of the scale invariant feature detectors SIFT, ALP, and HALF SIFT. While SIFT features are detected in the Difference of Gaussian (DoG) scale space, ALP employs the Laplacian of Gaussian (LoG) scale space. HALF SIFT uses DoG like SIFT, but applies a signal based subpixel and subscale localization technique. Although the detectors are very similar, the evaluation shows significant differences. The ALP detector performs better than SIFT. For Gaussian shaped input features, it is shown that the maximum systematic error is reduced by 77 % compared to SIFT. For natural images, ALP shows a gain of up to 16 % in repeatability compared to SIFT. For large viewpoint changes, the gain decreases. The reference detector HALF SIFT, which shows no systematic error on Gaussian features, performs best, especially for strong viewpoint changes, but requires more computation time.

As the usage of the LoG scale space provides higher localization accuracy for scale invariant features compared to DoG, our recommendation is to combine the signal based localization approach with the LoG scale space for optimal results. This approach will be evaluated in future works.

References

[1] M. Bauml and R. Stiefelhagen. Evaluation of local features for person re-identification in image sequences. In *IEEE International Conference on Advanced Video and Signal-Based Surveillance (AVSS)*, pages 291–296, 2011.

[2] M. Brown and D. G. Lowe. Invariant features from interest point groups. In *British Machine Vision Conference (BMVC)*, pages 656–665, 2002.

[3] K. Cordes, O. Müller, B. Rosenhahn, and J. Ostermann. HALF-SIFT: High-accurate localized features for SIFT. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Workshop on Feature Detectors and Descriptors: The State Of The Art and Beyond*, pages 31–38, 2009.

[4] K. Cordes, B. Rosenhahn, and J. Ostermann. Increasing the accuracy of feature evaluation benchmarks using differential evolution. In *IEEE Symposium on Differential Evolution (SDE)*, pages 1–8, 2011.

[5] K. Cordes, B. Rosenhahn, and J. Ostermann. High-resolution feature evaluation benchmark. In R. Wilson, editor, *Computer Analysis of Images and Patterns*, volume 8047 of *LNCS*, pages 327–334. Springer Berlin Heidelberg, 2013.

[6] G. Francini, M. Balestri, and S. Lepsoy. CDVS: Telecom italia's response to CE1 - interest point detection. In *ISO/IEC JTC1/SC29/WG11 Doc. M31369, Geneva, Switzerland*. MPEG-7 Video Subgroup: Compact Descriptors for Visual Search, 2013. `http://wg11.sc29.org/`.

[7] R. Hess. An open-source SIFT library. In *Proceedings of the International Conference on Multimedia*, MM '10, pages 1493–1496, New York, NY, USA, 2010. ACM.

[8] T. Lindeberg. Detecting salient blob-like image structures and their scales with a scale-space primal sketch: A method for focus-of-attention. *International Journal of Computer Vision (IJCV)*, 11:283–318, 1993.

[9] T. Lindeberg. Feature detection with automatic scale selection. *International Journal of Computer Vision (IJCV)*, 30:79–116, 1998.

[10] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision (IJCV)*, 60(2):91–110, 2004.

[11] D. G. Lowe. Method and apparatus for identifying scale invariant features in an image and use of same for locating an object in an image, Mar. 2004. Patent No. US 6,711,293.

[12] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. V. Gool. A comparison of affine region detectors. *International Journal of Computer Vision (IJCV)*, 65(1-2):43–72, 2005.

[13] N. Snavely, S. M. Seitz, and R. Szeliski. Modeling the world from internet photo collections. *International Journal of Computer Vision (IJCV)*, 80:189–210, 2008.