Spatio-Temporal Good Features to Track

Christoph Feichtenhofer and Axel Pinz
Institute of Electrical Measurement and Measurement Signal Processing
Graz University of Technology, Austria
cfeichtenhofer@gmail.com, axel.pinz@tugraz.at

Abstract

This paper presents two fundamental contributions that can be very useful for any autonomous system that requires point correspondences for visual odometry. First, the Spatio-Temporal Monitor (STM) is an efficient method to identify good features to track by monitoring their spatiotemporal (x-y-t) appearance without any assumptions about motion or geometry. The STM may be used with any spatial (x-y) descriptor, but it performs best when combined with our second contribution, the Histogram of Oriented Magnitudes (HOM) descriptor, which is based on spatially oriented multiscale filter magnitudes. To fulfil the real-time requirements of autonomous applications, the same descriptor can be used for both, track generation and monitoring, to identify low-quality feature tracks at virtually no additional computational cost. Our extensive experimental validation on a challenging public dataset demonstrates the excellent performance of STM and HOM, where we significantly outperform the well known "Good Features to Track" method and show that our proposed feature quality measure highly correlates with the accuracy in structure and motion estimation.

1. Introduction

Many computer vision applications rely on good features to track: Visual odometry [22], Structure from Motion [23], Simultaneous Localization And Mapping (SLAM) [8], and Augmented Reality [15] require the robust and precise tracking of a set of feature points. Most of these approaches combine interest point detectors with feature descriptors to yield detector-descriptor-based tracking, while other approaches estimate the motion based on the optical flow equation [11]. In all cases, "Good Features To Track" - GFTT - have to be detected to obtain robust and precise results. However, even highly salient points in individual frames might still be bad features to track. Therefore, GFTT can only be identified by a temporal monitoring of feature quality. This phenomenon has first been addressed in the seminal paper by Shi and Tomasi [25] almost 20 years ago.

In more recent work, it has been a common strategy to remove outliers at higher levels in the processing chain by incorporating geometric constraints, *e.g*. by using RANSAC [22, 15], or to use positional tracking information from previous frames, *e.g*. with Kalman [8], or particle filters [9]. These methods use prior information on scene geometry and motion smoothness, and therefore may restrict the general applicability in dynamic scenarios.

In contrast to these recent developments, we revisit the original idea of filtering GFTT in a bottom-up, model-free manner. We aim at eliminating bad feature tracks at the lowest level of the processing chain, without any priors about the expected motion or the scene geometry. Shi and Tomasi [25] analyse the change of appearance of a feature between the first and the current frame by calculating the target's rms residue and consequently reject a feature if its dissimilarity grows too large. Before the calculation of the rms residue, they apply an affine warping to account for the spatial appearance change of the target. In an extension of the original GFTT method, Tommasini *et al*. [26] estimate the distribution of this residual and automatically select a threshold for feature rejection. Although the GFTT approach is invariant to affine changes of the tracking target (*e.g*. caused by camera viewpoint or orientation change), our experimental results demonstrate that the selection of a good global rejection threshold is difficult for GFTT.

In this paper, we address the problem of finding good features to track by a spatiotemporal appearance-based approach. We monitor the quality of features by analyzing both their spatial and temporal appearance. Our novel method, termed Spatio-Temporal Monitor (STM), generates a spatiotemporal feature representation, which is used to detect incorrect correspondences during tracking, even between ambiguous scene structures. When the same descriptor is used for track generation and monitoring, this detection comes at virtually no additional computational cost.

Please note that there exists a large body of literature on comparison of interest-point detectors/descriptors. In many cases (e.g. [18, 19, 5, 16]), quality is measured based on repeatability and stability of points, often under affine warping, and often just for individual images. While these quality measures are well-suited, for example, to select good points for large baseline stereo, they are all based on the assumption of locally planar support regions.

In contrast, we are not searching for the “best” interest points in an individual frame (for instance, in case of repetitive patterns, highly salient corners might still lead to “jumps” in trajectories). Our emphasis is on the continuous temporal analysis adapting to the individual space-time appearance of each tracked feature. We make two main contributions in this work. First, we introduce a Spatio-Temporal Monitor (STM) that generates a spatiotemporal representation on top of any existing spatial descriptor. This allows monitoring the features with descriptors that are invariant to common visual challenges such as a change in the lighting conditions. The space-time representation is used to calculate a feature quality measure, which adapts to each feature individually in order to facilitate the selection of a global feature rejection threshold. Second, we introduce a novel spatial descriptor called Histogram of Oriented Magnitudes (HOM) that provides a large degree of invariance to local deformations and therefore delivers superior results when used with the STM to detect bad features during tracking. Experiments on a challenging public dataset, with a large number of ambiguities in the scenes, reveal the efficacy of both contributions.

2. The Spatio-Temporal Monitor

To model the quality of the features for tracking, we propose an effective representation, termed STM, that models the spatial appearance change of the tracked features in a small temporal window of scale α. For a given track, the STM $\{\mathbf{H}, \mathbf{d}\}$ holds the descriptors $\mathbf{H}$ of all preceding frames and additionally quantifies the change $\mathbf{d}$ of spatial appearance over time, as shown in Figure 1.

The n columns of $\mathbf{H} = [\mathbf{h}_1 \cdots \mathbf{h}_n]$ consist of the track's descriptors in each temporal instance $t = 1, 2, \dots, n$, and $\mathbf{d} = [d_1 \cdots d_n]^\top$ are the distances between each $\mathbf{h}_t$ and a single descriptor, which represents the cumulative appearance of the features during tracking. This space-time descriptor $\mathbf{h}_{\text{ST}}$ is computed via a spatiotemporal weighting of all descriptors of a given track as

$$\mathbf{h}_{\text{ST}} = \mathbf{H}\mathbf{w}, \tag{1}$$

with the spacetime weights $\mathbf{w} = [\gamma_1 \tau_1 \cdots \gamma_n \tau_n]^\top$. The spatial factors γ_t are calculated as the inverse of the feature distances, which represent the spatial appearance change

$$\gamma_t = \frac{1}{d_t}, \tag{2}$$

and the temporal locality adaptor τ_t is modelled by a Gaussian

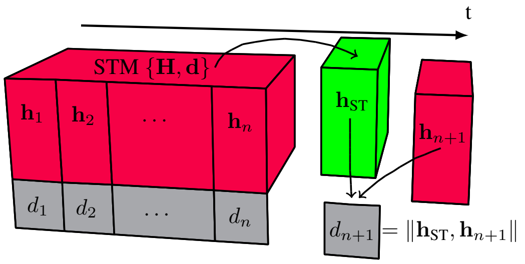


Figure 1. Overview of STM-based feature monitoring. For each feature to track, we generate a spatiotemporal descriptor $\mathbf{h}_{\text{ST}}$ from the collected spatial descriptors of a given track. At time $t = n$, the feature descriptors $\mathbf{H}$ of a track are collected with their corresponding distances $\mathbf{d}$ to form the spatiotemporal representation $\mathbf{h}_{\text{ST}}$ of the features. At time $t = n + 1$, we compute the distance between the next descriptor $\mathbf{h}_{n+1}$ and $\mathbf{h}_{\text{ST}}$ to quantify temporal feature similarity.

$$\tau_t = \exp\left(-\frac{(t-n)^2}{2\alpha^2}\right), \tag{3}$$

to give higher importance to recent features, where the temporal scale α controls the speed of the weighting decay for older features. The spatial weights (2) give low influence to imprecisely located tracks, in contrast to the temporal weighting (3), which allows a continuous spatial appearance change of the tracking target over time. We subsequently normalize the weights, such that $\|\mathbf{w}\|_1 = 1$.

Given $\mathbf{h}_{\text{ST}}$ for $t = 1, 2, \dots, n$, for each new candidate feature $\mathbf{h}_{n+1}$, at time $t = n+1$, the matching distance d_{n+1} to $\mathbf{h}_{\text{ST}}$ is calculated by using a distance metric (*e.g.* χ^2). This tuple $\{\mathbf{h}_{n+1}, d_{n+1}\}$ is then added to the STM $\{\mathbf{H}, \mathbf{d}\}$, only if the quality-ratio q_{n+1} to the existing spatiotemporal representation of the features is reasonably high:

$$q_{n+1} = \frac{\bar{d}}{d_{n+1}} > Q, \tag{4}$$

where Q is a feature quality threshold and $\bar{d}$ is a weighted mean of all distances in the STM, computed by using the temporal weighting τ_t:

$$\bar{d} = \frac{\sum_{t=1}^{N} \tau_t d_t}{\sum_{t=1}^{N} \tau_t}. \tag{5}$$

Eq. (4) allows us to detect bad features to track. For a steady adaptation of the feature to track, and to maintain very low memory usage, we propose to discard old STM entries $\{\mathbf{h}_t, d_t\}$, for $t < n - 3\alpha$, during monitoring, as their influence on the spatiotemporal representation is very low due to their temporal weights $\tau_t < 0.01$.

Our feature quality measure q exhibits some major advantages over common measures such as the GFTT's affine deformation model. First, because we use distance ratios instead of absolute distances, a global selection of one quality threshold is much easier (*e.g.* absolute inter-frame feature distances may change drastically from one trajectory to another). Second, our quality measure adapts to the spatial change of each feature to track individually and is invariant to the descriptor type used for monitoring. Third, because of the temporal weighting τ, our quality measure also adapts to temporal changes in feature distances (*e.g.* caused by an indirect influence of varying illumination on the magnitude of the feature vector).

The STM has 2 parameters: *1*) The feature quality threshold Q, which decides if a new feature point is considered as a correct correspondence. *2*) The temporal scale α of the spatiotemporal descriptor within the STM. The STM can be used with all kinds of (invariant) spatial descriptors (see the extensive comparative evaluation in section 4), but it is especially effective with our novel Histogram of Oriented Magnitudes (HOM) descriptor.

3. The Histogram of Oriented Magnitudes

Derivative of Gaussian and Gabor filters are well suited to resemble biological impulse response functions [7]. With our proposed HOM descriptor, we try to model simple cells in the primary visual cortex. We take Gaussian second-derivative filters G_2 at O equally spaced orientations θ_i, with the filter parameters chosen similar to the ones used in [24], which are based on quantitative studies on the properties of the corresponding simple cells in the primary visual cortex. We use an elongation ratio of 3:1 along the dominant orientations $\theta_i \in \{0°, 45°, 90°, 135°\}$ of the filters, with the five spatial scales $\sigma_j \in \{1, 2, 3, 4, 5\}$, being the standard deviations of the Gaussians. These even-symmetric derivative of Gaussian filters respond best when centred on a line with specific width and orientation. Figure 3 illustrates the G_2 filters in the spatial domain.

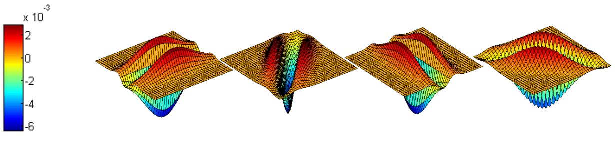


Figure 2. Gaussian second derivative filters at four equally spaced orientations. Our HOM descriptor uses multiscale filters with responses collapsed into a single histogram bin to be sensitive for various spatial scales at a given orientation. Best viewed in colour.

The oriented filter responses M for an image patch P, with coordinates $\mathbf{x} = (x, y)^\top$ are computed by convolution with the orientation selective filters, of orientation θ_i and scale σ_j, according to

$$M(\mathbf{x}, \theta_i, \sigma_j) = |P(\mathbf{x}) * G_2(\theta_i, \sigma_j)|. \tag{6}$$

In order to avoid border effects, we extract larger patches from the image, dependent on the maximum filter size, and symmetrically replicate the image only if the convolution would exceed an image border.

Local Contrast Normalization: Because of the bandpass nature of the filters used, the responses (6) are invariant to additive photometric effects. To additionally provide invariance to locally varying image contrast, we adopt the pixel-wise normalization in [28], by normalizing each orientation selective measurement with respect to the sum of all filter responses at a given scale:

$$\hat{M}(\theta_i, \sigma_j) = \frac{M(\theta_i, \sigma_j)}{\sum_{i=1}^{O} M(\theta_i, \sigma_j) + \epsilon}, \tag{7}$$

where a small constant ϵ is added to the sum of the magnitudes over all orientations. This bias operates as a noise floor and avoids instabilities at low overall magnitudes. Note that equation (7) cancels out multiplicative photometric changes, since these appear in both the numerator and denominator.

Scale Invariant Orientation Measurements: Similar to the SIFT descriptor [17], the descriptive patch is divided into $N \times N = 4 \times 4$ local cells. The normalized filter responses $\hat{M}(\theta_i, \sigma_j)$ are summed and histogrammed in each cell. According to their filter orientations θ_i the responses cast weighted votes in spatial orientation bins. Measurements for different spatial scales σ_j but consistent orientations θ_i are accumulated in identical bins to provide invariance for small scale shifts

$$H^i = \sum_j \hat{M}(\theta_i, \sigma_j). \tag{8}$$

Our scale-invariant, but phase-variant, HOM descriptor is subsequently constructed by concatenation of the O-bin histograms of oriented magnitudes H for the $N \times N$ local cells.

Chromatic Opponency: Visual receptive fields are most sensitive in a small central region, while the surrounding region inhibits the responses of the neurons [14]. Such an antagonistic center-surround behavior has been found for light intensities, red/green and blue/yellow opponencies. Since the HOM descriptor is designed to model simple cells in the visual cortex, we apply the descriptor to these colours. The three considered colour channels, adopted from [14, 27], are: Intensity $I = \frac{r+g+b}{3}$, red/green $RG = \frac{r-g}{\max(r,g,b)}$ and blue/yellow $BY = \frac{b-\min(r,g)}{\max(r,g,b)}$, where r, g and b are the red, green and blue colour channels of the input images. For pixels with $\max(r, g, b) < 10\%$ of the maximum possible value, RG as well as BY are set to zero, because hue variations are not perceivable at very low luminance [14].

Adding the C colour channels, we obtain an $O \times N \times N \times C = 4 \times 4 \times 4 \times 3 = 192$ dimensional feature vector.

Since the HOM descriptor represents a histogram, we $L1$ normalize it, in order to make histogram distance metrics applicable for comparison.

The HOM descriptor exhibits some noteworthy properties for feature monitoring: First, some degree of invariance to small deformations and changes in rotation, due to the broad tuning of the filters. Second, invariance to additive and multiplicative illumination changes, due to the filter's bandpass nature and the local contrast normalization (7), respectively. Third, efficient computations, by performing the filtering with separable and steerable basis filters, as described in [10]. Fourth, the generation of a smooth overlap between the $N \times N$ local cells of the patch, due to the convolution operation, and therefore the avoidance of boundary effects between them; consequently, we do not require any additional normalization and interpolation steps between neighbouring cells.

4. Experiments

We apply the STM to monitor the features to track. For this task, we generate tracks with two different approaches: In section 4.1, tracks are generated using a detector-descriptor-based approach (SIFT-tracks), and in section 4.2, we show results for optical-flow-based tracks (LK-tracks) on features detected by the minimum eigenvalue method [25].

Dataset: We demonstrate the performance of STM and HOM by extensive experimental validation on the challenging public Robot dataset [1], which provides sufficient complexity, many ambiguities (i.e. highly similar features), and known spatial correspondences of the scene. The ground truth consists of camera pose and scene surface information for 60 different scenes. The images of size 1200×1600 pixels are acquired with a moving camera, mounted on an industrial robot. This setup allows very accurate positioning, with a standard deviation of approximately 0.1mm, which corresponds to 0.2-0.3 pixels when back-projected onto the images [1]. Note that all evaluated methods are independent from the scene dynamics. Therefore, using a dynamic dataset (*e.g.*, the KITTI visual odometry benchmark [12]), would not affect the performance of the evaluated algorithms. Moreover, these datasets do not provide the same degree of localization accuracy.

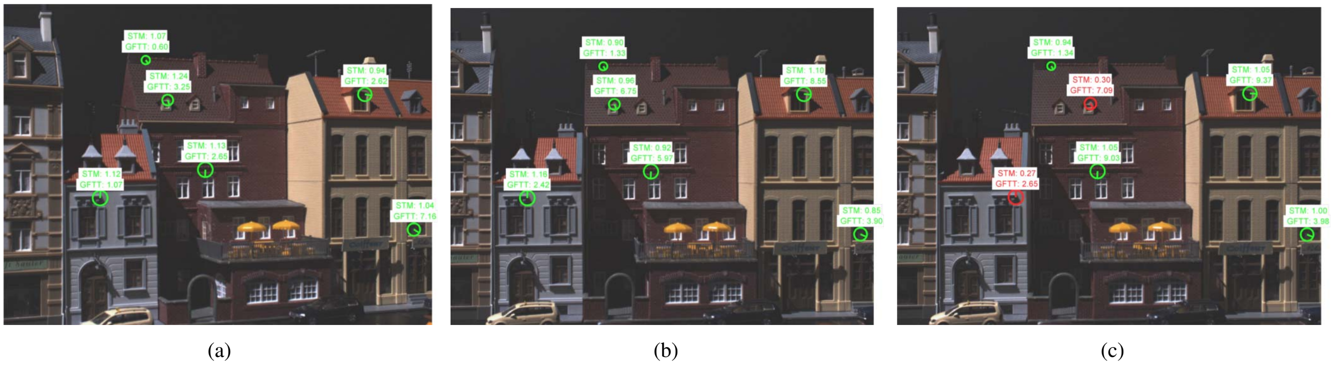
To illustrate a scene of the employed dataset, and to demonstrate STM-based feature monitoring, Figure 3 shows six trajectories, generated by inter-frame SIFT matching [17]. The two wrong tracks (*i.e.* bad features to track) are correctly detected by the STM, based on a significantly decreased quality ratio between the feature in the current frame and the spatiotemporal feature descriptor of the previous frames. Since we use the same descriptor for matching and monitoring, these detections come with very low computational cost.

Descriptors: we use several descriptors within the STM. These describe local patches, which correspond to the keypoint frames for the SIFT-tracks and to the tracking template for LK-tracks. The STM is evaluated with the following 6 descriptors: *1*) The HOM descriptor, as described in section 3, with the normalization bias empirically set to $\epsilon = 0.1$; *2*) SIFT [17], *3*) HOG; [6]; *4*) SURF [4]; *5*) a set of four normalized moment Invariants of order 1 to 4 (INVM) [13] (to facilitate comparison, we bring these moments into a common numeric range by applying a log transformation before comparison); and *6*) colour moments (COLM) [20] invariant to affine geometric deformations and diagonal photometric transformations.

Spatio-Temporal Scale: An in-depth evaluation of multiple temporal scales α for the STM, which included different camera paths with varying inter-frame viewpoint changes, indicated that precise temporal scale selection is not necessary in general, with stable results for $2 \leq \alpha \leq 5$. Further investigations with multiple spatial scales σ for the HOM descriptor have shown a steady increase in performance when increasing the number of scales. We therefore use five different spatial scales $\sigma \in \{1, 2, 3, 4, 5\}$, and $\alpha = 3$ for all experiments in this paper.

Descriptor Comparison: All the histogram-based descriptors are normalized by using the $L1$ norm and compared with appropriate histogram distance measures within the STM; as shown in [2], this yields a higher matching performance in case of the SIFT descriptor. We observe, that using different histogram distance metrics, such as χ^2, only leads to negligible differences in performance. Therefore, all our results are based on the Bhattacharyya distance. The non-histogram-based descriptors (*i.e.* SURF and moments) are compared via the Euclidean distance.

Evaluation Methodology: We create Receiver Operator Characteristic (ROC) curves, by varying the quality threshold Q of the STM. If a feature exhibits a low quality-ratio $q_{n+1} < Q$, we declare the corresponding tracking trajectory as faulty at time $n+1$. For comparison, we also show the performance of the GFTT method [25], by varying their feature rejection threshold. The experiments for a single track may have one of the following outcomes: *a*) a True Positive (TP) occurs only if an incorrectly matched feature is detected in the same temporal instance (frame) as the trajectory's first incorrect correspondence in the ground truth;



(a) (b) (c)

Figure 3. Several SIFT trajectories in frames 10 (a), 20 (b) and 21 (c) of Scene 1. The STM's feature quality ratio q and the GFTT's dissimilarity measure are shown above the corresponding keypoints. The STM identifies the bad features (red), delivering incorrect ground truth correspondences, by a low quality ratio, only by analysing the spatial appearance change of the SIFT features over time. The GFTT's dissimilarity measure diverges among the features and is not able to distinguish between the good (green) and bad (red) features to track.

	STM						GFTT
	HOM	SIFT	HOG	SURF	INVM	COLM	[25]
AUC	0.839	0.625	0.619	0.575	0.603	0.583	0.429
F	0.674	0.397	0.425	0.355	0.428	0.416	0.243

Table 1. Performance for the monitoring of SIFT-tracks, averaged for all scenes.

b) a False Positive (FP) occurs for detections in other temporal instances; *c*) a True Negative (TN) occurs if no incorrect match is detected and the ground truth also shows no incorrect correspondence; and *d*) a False Negative (FN) occurs if no incorrect match is detected, but the ground truth shows incorrect correspondences. Along with the ROC curves, which show the True Positive Rate $= \frac{TP}{TP+FN}$ plotted against the False Positive Rate $= \frac{FP}{FP+TN}$, we also report the Area Under Curve (AUC) and the maximum F-measure ($\frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$) as performance measures.

4.1. Detector-Descriptor-Based Tracks

We generate tracks, using Difference of Gaussian keypoints in combination with SIFT descriptors [17], because this combination is known to perform well under changes of viewpoint [21]. For these experiments, we use the longest horizontal camera trajectory, consisting of 49 frames, taken from a circular camera path around the scene, from a distance of 0.5m, and a total viewpoint rotation of 80° (*cf.* [1]). This results in an inter-frame viewpoint change of about 1.6°. For the inter-frame matching we we follow Lowe's ratio criterion of only using matches with a best to second-best match distance ratio of less than 0.8 [17]. Considering this matching criterion, we end up with 0 to 961 trajectories of length 49 for each scene, depending on its contents. Since several scenes, particularly those showing only a single object, generate a small number of trajectories, we only report results for those 35 scenes that generate more than 100 trajectories. These trajectories are evaluated by using the proposed STM with different descriptors, applied to the keypoint ROIs.

The ROC curves for the three scenes with the largest number of trajectories are shown in the first row of Figure 4. Each curve represents the detections of the STM for a different descriptor, with the AUC as reference value for overall performance. The non-monotonicity in the curves results from the true positive requirement of detecting the exact temporal location of the first incorrect match. As can be seen, STM+HOM outperforms the other combinations, and GFTT by a wide margin. This large performance gain of STM over GFTT can be explained by the individual adaptability of our quality measure (4) to each feature, as it is a ratio based on each individual track rather than a firm measure for all tracks. As visualized in Figure 3, the good features can easily be separated from the bad ones, by using a quality threshold of $Q = 0.5$ within the STM; whereas the GFTT dissimilarity measure exhibits large fluctuations and is not able to monitor the features correctly. In contrast to GFTT, we analyse the features over a small temporal window, which provides higher invariance than measuring the similarity between affine warped templates.

The overall results for all 35 considered scenes are reported in Table 1 by listing the mean AUCs and average maximum F-Measures. In combination with the STM, the proposed HOM descriptor significantly outperforms the GFTT method. Also, the SIFT descriptor is able to detect a large percentage of incorrect tracks, only by adding temporal information to the spatial SIFT-tracking procedure.

Please note that the tracks are not generated via brute force matching of the SIFT features. We only use a subset of around 2-10% of the features in the first frame, which satisfy Lowe's ratio criterion over all 49 frames. All other

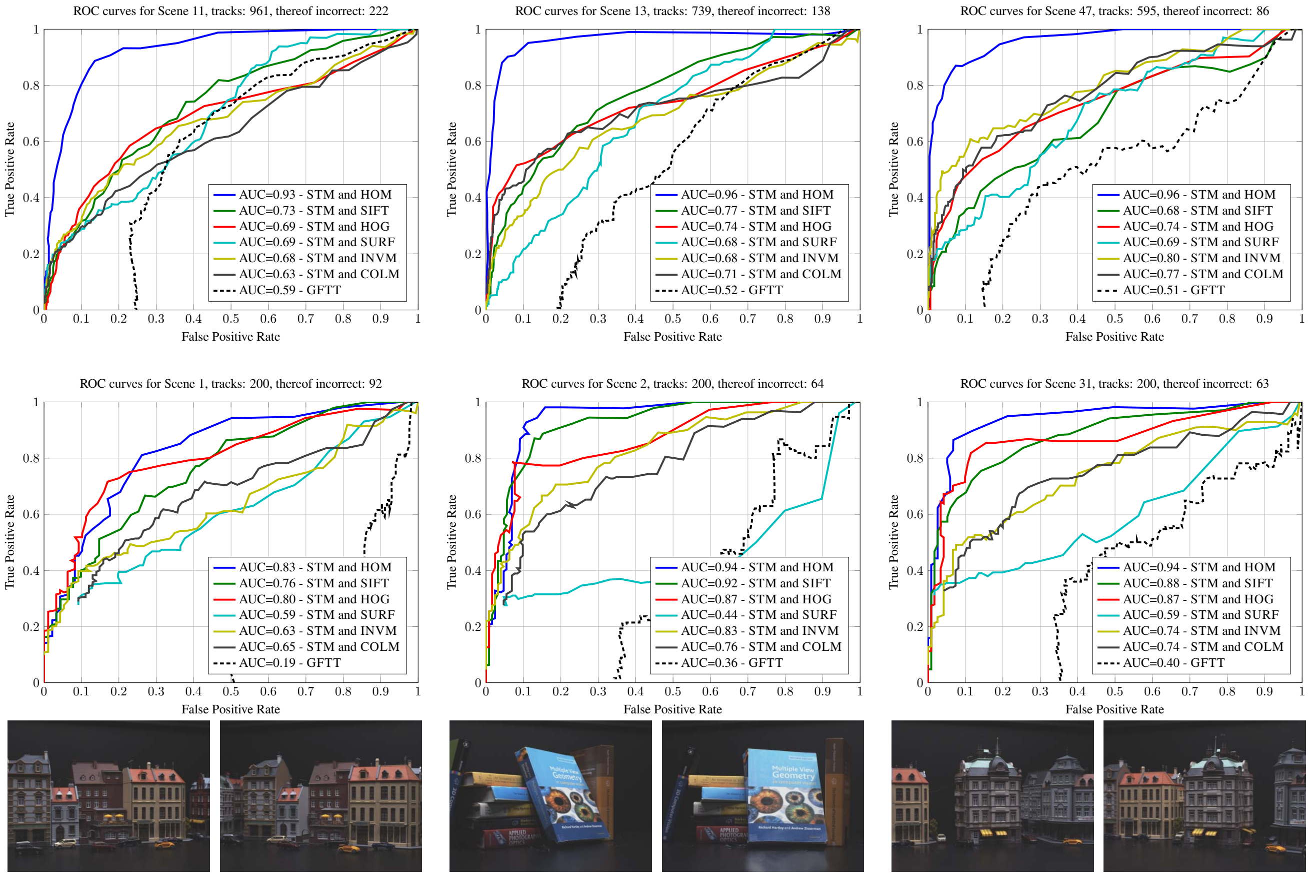


Figure 4. Results for six sample scenes, providing detailed comparison of STM using different spatial feature descriptors with GFTT. The first row shows results for those three scenes that generated the largest number of SIFT tracks (fulfilling Lowe's ratio criterion over all frames). The second and third rows show three selected results for LK-tracks and the corresponding scenes.

	STM						GFTT
	HOM	SIFT	HOG	SURF	INVM	COLM	[25]
AUC	0.825 (0.730)	0.801 (0.704)	0.747 (0.640)	0.548 (0.427)	0.683 (0.556)	0.677 (0.552)	0.261 (0.139)
F	0.755 (0.636)	0.722 (0.607)	0.677 (0.546)	0.487 (0.376)	0.592 (0.461)	0.606 (0.459)	0.276 (0.144)

Table 2. Performance for monitoring LK-tracks on all 56 scenes under diffuse illumination. Parentheses indicate the performance without considering the soft evaluation setting.

trajectories are rejected. Therefore, the remaining tracks exhibit already very consistent appearance, so that it is hard to identify faulty tracks by a purely appearance-based approach (see e.g. the matching to scene ambiguities in Figure 3).

4.2. Optical-Flow-Based Tracks

For these experiments, we extract 200 features, based on the minimum eigenvalue method [25], in the first frame of 56 scenes. We exclude the 4 twigs scenes (57-60) from our evaluation, because they mainly show virtual crossings in front of black background regions, where no ground truth is available from the structured light scans. We use a suppressed region of size 45×45 pixels around each detected corner and a tracking template of 35×35 pixels. The template is tracked by using a robust affine implementation of the Lucas-Kanade (LK) tracker [3]. Because the LK tracker has issues with large viewpoint variations, we now use the images taken from the furthest circular camera path, with a distance of 0.8m from the scene and a total viewpoint change of 40° [1].

Optical-flow-based tracks tend to long term drift, which causes an error accumulation over time. For more reasonable comparisons, we allow the algorithms to detect a faulty track within a window of ±3 frames around the track's first violation of a ground truth criterion and still consider it as

Q	2	5/4	1	5/6	5/7	5/8	5/9	All features
	SIFT-tracks							
Mean back-projection error	0.2317	0.2320	0.2628	0.5025	0.8385	1.0453	1.6299	5.7018
Avg. # good features per scene	252.77	375.31	936.09	3805.2	7512.7	9019.5	9789.5	12264
	LK-tracks							
Mean back-projection error	0.2960	0.2989	0.3382	0.4528	0.6235	0.7782	0.9297	5.9114
Avg. # good features per scene	163.21	255.45	577.95	1786.4	2804.3	3113.3	3255.7	4058

Table 3. Structure and motion estimation with the good feature subset fulfilling the quality Q of STM+HOM. The mean back-projection error of the estimated 3D points in pixels, averaged over all corresponding features and scenes, is shown. The average number of features for all scenes and those, which satisfy the STM's quality threshold Q, is listed below. High quality features generate a significantly lower estimation error.

a correct detection; however, we also provide the results without consideration of this soft evaluation setting, by indicating these scores in parentheses.

Results for the experiments on the optical-flow-based tracks are given in Table 2(a), averaged for 56 scenes. Our methods provide a significant improvement over the GFTT approach, with performance gains of 216% (425%) and 174% (342%) in AUC and F-measures, respectively. The STM performs best combined with the HOM descriptor. The gradient histogram-based methods SIFT and HOG are also competitive.

Typical ROC curves for LK-tracks are shown in Figure 4, where STM+HOM again performs best. Overall, the decent performance of the 4 dimensional invariant moments and the generally poor performance of SURF is also remarkable.

Overall, the significantly higher performance of the proposed STM method over GFTT can be attributed to the design of the STM's feature quality measure. Because it is a ratio rather than a firm threshold, as in GFTT, it adapts to each feature to track individually and therefore is invariant to the appearance of the underlying features to track. For GFTT, choosing a good threshold for all trajectories is very difficult. As reflected in the ROC curves of figure 4, many false detections occur even for high dissimilarity thresholds, because GFTT always falsely detects many correct correspondences too. In contrast, our dissimilarity measure adapts to the features of each single trajectory and furthermore is even invariant to the descriptor type used for monitoring. On the other hand, the better performance of HOM over other descriptors can be explained by its flexibility to deformations and scale variations. This is because, compared to the discrete derivative masks in gradient orientation based descriptors (*e.g.*, SIFT, HOG), the proposed HOM is based on oriented Gaussian derivative filters with multiscale measurements jointly aggregated in histogram bins.

4.3. Feature Quality for Structure and Motion

The estimation of the scene structure and camera position is highly important in autonomous driving tasks to facilitate navigation and collision avoidance. For accurate camera pose estimation and scene reconstruction, correct point correspondences have to be established. In this experiment, we evaluate the effect of our feature quality measure Q on the selection of good feature points for scene structure and motion estimation. For this purpose, we monitor the error in the estimation by using a fraction of good features that fulfil a minimum quality Q. First, 3D points are generated by triangulating the feature points with the corresponding projection matrices of the camera (using a DLT algorithm followed by a Levenberg-Marquardt optimization). Second, to evaluate the quality of our selected features, we back-project the estimated 3D points onto the image plane and calculate the back-projection error. A correct correspondence will deliver a 3D point located on the scene surface, generating a back-projection of around 0.2-0.3 pixels [1].

We evaluate all SIFT- and LK-tracks and use each feature only if its quality ratio is larger than Q. The results are shown in Table 3, where the mean back-projection error for each subset of good features is shown. We further list the error for using all features in the last column and, moreover, the number of features that satisfy the minimum quality Q, averaged for all scenes (*i.e.* "Avg. # good features per scene"). We observe an inversely monotonic behaviour between the STM's feature quality threshold Q and the mean back-projection error. The higher Q, the more bad features are rejected to generate a sparser but more accurate estimation. Overall, we observe a huge benefit in terms of estimation accuracy by the filtering of bad features with STM+HOM. For example, for Q =5/8, we can improve the accuracy by 445% (SIFT tracks) and 658% (LK), by filtering 36% and 25% of the features, respectively.

5. Summary and Discussion

The first main contribution of this paper is the novel Spatio-Temporal Monitor STM that monitors the quality of features to track based on their appearance in space and time. This method combines the temporal dynamics of the features and their spatial appearance in a unified spatiotemporal representation. The major strengths of this approach are: (i) Because the STM works on top of any tracker, and for any spatial descriptor, it can be widely used. (ii) The STM can be used online, in an incremental fashion, to detect the instance in time when a particular feature fails. (iii) It does not use any priors about motion-coherence or scene geometry. (iv) When the same descriptors are used to generate and to analyse tracks, this online analysis can be carried out in the background, at virtually no additional computational cost.

We have performed a thorough experimental validation of the STM to compare the power of various commonly used descriptors, on the spatiotemporal dataset that provides accurate ground truth, sufficient diversity, and spatial detail. Our results clearly demonstrate a significant gain (*i.e.* more correctly identified wrong tracks) over the GFTT method, independent of the spatial descriptor that has been used.

The second contribution of this paper is a novel spatial descriptor, the Histogram of Oriented Magnitudes HOM. It is based on spatially oriented filter magnitudes, motivated by biological vision systems. The HOM descriptor tolerates slight deformation and rotation of the tracking target, due to the rather broad tuning of the filters used. It is invariant to additive and multiplicative photometric changes and it may be implemented very efficiently by using separable and steerable filters. In combination with the STM, HOM exhibits superior performance for monitoring features.

In the context of vision-based autonomous driving, this combination of STM and HOM will be very useful for many systems. It should be employed to filter good feature tracks at a low level underneath higher level processes that may exploit geometric constraints or motion information. Our own focus will be in the online identification of good features for Multibody Structure and Motion analysis (see *e.g.* [23] for a concise definition of this task). To provide an online analysis of camera pose and independently moving foreground objects, we wish to concentrate on a limited number of a few, but reliable, good tracks. The STM will provide us with exactly these good features to track. Furthermore, we expect to reliably harvest more tracks on the moving foreground objects to be able to produce better object models.

We share the code for our methods online at `http://www.emt.tugraz.at/~pinz/code/`.

References

[1] H. Aanæs, A. Dahl, and K. Steenstrup Pedersen. Interesting interest points. *IJCV*, 97:18–35, 2012. 4, 5, 6, 7

[2] R. Arandjelovic and A. Zisserman. Three things everyone should know to improve object retrieval. In *CVPR*, 2012. 4

[3] S. Baker and I. Matthews. Lucas-Kanade 20 years on: A unifying framework. *IJCV*, 56(3):221–255, 2004. 6

[4] H. Bay, T. Tuytelaars, and L. V. Gool. SURF: Speeded up robust features. In *ECCV*, 2006. 4

[5] H. Comer and B. Draper. Interest point stability prediction. In *ICVS*, 2009. 1

[6] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005. 4

[7] J. G. Daugman. Spatial visual channels in the Fourier plane. *Vision Research*, 24(9):891 – 910, 1984. 3

[8] A. J. Davison. Real-time simultaneous localisation and mapping with a single camera. In *ICCV*, 2003. 1

[9] E. Eade and T. Drummond. Scalable monocular SLAM. In *CVPR*, 2006. 1

[10] W. Freeman and E. Adelson. The design and use of steerable filters. *PAMI*, 13(9):891 –906, 1991. 4

[11] S. Gauglitz, T. Höllerer, and M. Turk. Evaluation of interest point detectors and feature descriptors for visual tracking. *IJCV*, 94:335–360, 2011. 1

[12] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the KITTI vision benchmark suite. In *CVPR*, 2012. 4

[13] M.-K. Hu. Visual pattern recognition by moment invariants. *IRE Trans. Info. Theory*, 8:179–187, 1962. 4

[14] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *PAMI*, 20(11):1254 –1259, 1998. 3

[15] G. Klein and D. Murray. Parallel tracking and mapping for small AR workspaces. In *ISMAR*, 2007. 1

[16] B. Li, R. Xiao, Z. Li, R. Cai, B.-L. Lu, and L. Zhang. Rank-SIFT: Learning to rank repeatable local interest points. In *CVPR*, 2011. 1

[17] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60:91–110, 2004. 3, 4, 5

[18] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *PAMI*, 27(10):1615–1630, 2005. 1

[19] 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*, 65(1):43–72, 2005. 1

[20] F. Mindru, T. Tuytelaars, L. V. Gool, and T. Moons. Moment invariants for recognition under changing viewpoint and illumination. *CVIU*, 94:3–27, 2004. 4

[21] P. Moreels and P. Perona. Evaluation of features detectors and descriptors based on 3D objects. In *ICCV*, 2005. 5

[22] D. Nister, O. Naroditsky, and J. Bergen. Visual odometry. In *CVPR*, 2004. 1

[23] K. Ozden, K. Schindler, and L. Van Gool. Multibody structure-from-motion in practice. *PAMI*, 32(6):1134 –1141, 2010. 1, 8

[24] T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber, and T. Poggio. Robust object recognition with cortex-like mechanisms. *PAMI*, 29(3):411 –426, 2007. 3

[25] J. Shi and C. Tomasi. Good features to track. In *CVPR*, 1994. 1, 4, 5, 6

[26] T. Tommasini, A. Fusiello, E. Trucco, and V. Roberto. Making good features track better. In *CVPR*, 1998. 1

[27] D. Walther and C. Koch. Modeling attention to salient proto-objects. *Neural Networks*, 19(9):1395–1407, 2006. 3

[28] R. P. Wildes and J. Bergen. Qualitative spatiotemporal analysis using an oriented energy representation. In *ECCV*, 2000. 3