On SIFTs and their Scales*

– Extended version –

Tal Hassner
Open University of Israel
hassner@openu.ac.il

Viki Mayzels
Technion
mviki@techunix.technion.ac.il

Lihi Zelnik-Manor
Technion
lihi@ee.technion.ac.il

Abstract

Scale invariant feature detectors often find stable scales in only a few image pixels. Consequently, methods for feature matching typically choose one of two extreme options: matching a sparse set of scale invariant features, or dense matching using arbitrary scales. In this paper we turn our attention to the overwhelming majority of pixels, those where stable scales are not found by standard techniques. We ask, is scale-selection necessary for these pixels, when dense, scale-invariant matching is required and if so, how can it be achieved? We make the following contributions: (i) We show that features computed over different scales, even in low-contrast areas, can be different; selecting a single scale, arbitrarily or otherwise, may lead to poor matches when the images have different scales. (ii) We show that representing each pixel as a set of SIFTs, extracted at multiple scales, allows for far better matches than single-scale descriptors, but at a computational price. Finally, (iii) we demonstrate that each such set may be accurately represented by a low-dimensional, linear subspace. A subspace-to-point mapping may further be used to produce a novel descriptor representation, the Scale-Less SIFT (SLS), as an alternative to single-scale descriptors. These claims are verified by quantitative and qualitative tests, demonstrating significant improvements over existing methods.

1. Introduction

Over the past decade and a half, scale invariant feature detectors, such as the Harris-Laplace [21] and robust descriptors such as the SIFT [18], have played pivotal roles in maturing Computer Vision systems. The key idea is that at each interest point, one (or few) scales are selected based on a scale-invariant function (e.g., the Laplacian of Gaussians). Presumably, local extrema of this function occur at the same scales for the same feature in different images allowing the features to be matched across images in different scales [23]. A typical image, however, often has relatively few pixels for which such scales may be reliably selected. Consequently, matching of scale invariant features has so far been applied mostly to few pixels in each image.

Figure 1. **Dense matches of different objects in different scales. Top:** Left and Right input images. **Bottom:** Left image warped onto Right using the recovered flows: Using DSIFT (bottom left) and our SLS descriptor (bottom right), overlaid on the Right and manually cropped to demonstrate the alignment. DSIFT fails to capture the scale differences and produces an output in the same scale as the input. SLS captures scale changes at each pixel: the output produced by using SLS has the appearance of the Left image in the scale and position of the Right.

When dense correspondences are required, traditional methods restrict themselves to using pixels or pixel patches, filtered or otherwise (see, e.g., [11]). Alternatively, feature descriptors may be computed for all the pixels in the image (e.g., [27]). These are designed to be robust to a range of geometric and photometric image transformation. One such example is the Dense-SIFT (DSIFT) descriptor [29] which is extracted at a single scale for all the pixels in the image. Establishing correspondences between two images is then performed either locally or by using global optimization schemes such as the SIFT-Flow algorithm [16, 17]. Such methods, however, all implicitly assume that features in the

*Additional information available from: www.openu.ac.il/home/hassner/projects/siftscales/

two images share the same, or sufficiently similar, scales. As shown in Fig. 1, when this does not hold, correspondence estimation fails.

In this paper we focus on those pixels for which a method for selecting well defined scales is not known. Making up most of the image, these are the pixels for which local image intensities do not vary sufficiently to provide strong extrema in the scale selection function. This work presents the following contributions:

1. We show that even in low contrast areas of the image, where scale-selection is difficult, descriptors may change their values from one scale to the next. Consequently, selecting an arbitrary single scale may lead to false matches when two images have different scales.

2. We propose representing each pixel by a set of SIFT descriptors extracted at multiple scales and matched from one image to the next using set-to-set similarities. The computational cost of matching more descriptors is balanced by a substantial boost in accuracy.

3. We demonstrate that each such set of SIFTs resides on a low-dimensional subspace. We further show that the subspace-to-point mapping of [4, 5], provides a means of representing these subspaces as a novel feature descriptor, the Scale-Less-SIFT (SLS).

These set-based, multi-scale SIFT representations are tested on dense correspondence estimation problems with images separated by wide scale differences and changing viewing conditions and shown to significantly outperform existing methods both qualitatively and quantitatively.

2. Previous work

Objects and scenes appear in images in different scales. In order to correctly describe features when these scales are unknown, one must consider multiple scales for each feature point. Since the early 90s automatic scale selection techniques have been proposed which seek for each feature point a stable, *characteristic* scale. They thus augment earlier scale-space methods by choosing one scale for each feature for the purpose of both reducing the computational burden of higher level visual systems, as well as improving their performance by focusing on more relevant information (See [14] for more on these early approaches).

Lindeberg [15] suggested seeking for each feature its "interesting scales"; that is, scales which reflect a characteristic size of a feature. He proposed selecting these scales by choosing the extrema in the Laplacian of Gaussian (LoG) function computed over the image scales. Pixels of local extrema may additionally be rejected if their LoG value is lower than a predefined threshold. This is applied in order to ensure that unstable, low-contrast points are not selected. An efficient approximation to the LoG function is based on differences of Gaussian (DoG) filters (e.g., [18]). For a given image, three sets of sub-octave, DoG filters are produced. The resulting 3D structure (x, y and $scale$) is then scanned, searching for pixels with higher or lower values than their 26 space-scale neighbors. Coordinate localization is then performed in order to obtain more accurate pixel locations as well as, again, reject unstable detections located in low contrast areas or near edges.

Scale selection is sometimes performed along-side spatial localization. The Harris-Laplace detector [21], for example, uses a scale-adapted Harris corner detector to localize points spatially and LoG filter extrema to localize points in scale. These two steps are performed in an iterative procedure which searches for the joint peaks of these two values. Here too, points are rejected if they fail to produce responses stronger than a given threshold.

The methods mentioned above, as well as similar techniques, all typically produce a small set of interest points located near corner structures in the image. Mikolajczyk [20] reports that under a scale change factor of 4.4 the percent of pixels for which a scale is detected is as little as 38% for the DoG detector of which in only 10.6%, the detected scale was correct.

Several existing methods use few invariant features to seed a search for dense matches between different views of a wide-baseline stereo system [7, 26, 31]. As far as we know, however, none of these methods is designed to provide dense correspondences across scale differences. A noteworthy exception is the work of [25] which uses few scale-invariant features to locate an object in an image and then produces dense matches along with accurate segmentations. Their method, however, relies on a global alignment scheme to overcome the main scale differences before dense matching. It is thus unclear how it performs when no such alignment is possible (e.g., several independent scene motions).

In [12] scale invariant descriptors (SID) are proposed without requiring the estimation of image scale. A main advantage of SID is that they are applicable to a broader range of image structures, such as edges, for which scale selection is unreliable. Our experiments here show that SID are less capable of matching across different scenes than the SIFT descriptors underlying our representation. In [28], scale selection is avoided by computing multi-scale fractal features, developed for the purpose of texture classification.

Dense SIFT - no scale selection. When dense matching is required, a common approach is to forgo scale estima-

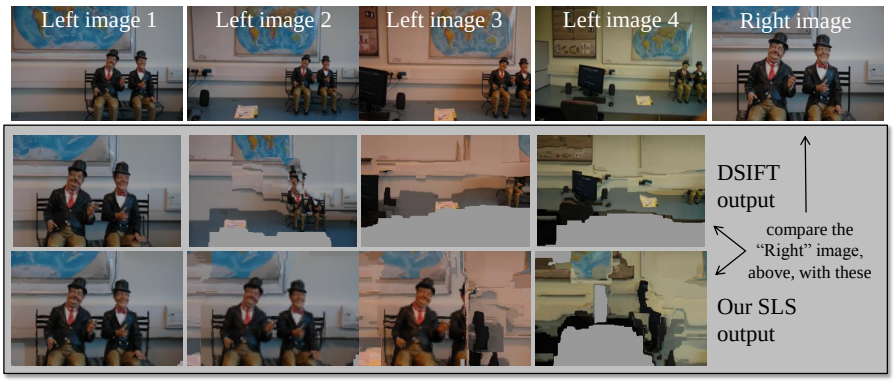

Figure 2. **Effects of scale differences on DSIFT vs. our own SLS descriptor.** Left images warped onto right image using correspondences obtained by the SIFT-Flow algorithm [16, 17] and the DSIFT descriptor, compared against the SLS descriptor (Sec. 3.3). The results in the bottom two rows should appear similar to the top-right image. DSIFT descriptors provide some scale invariance despite a single arbitrary scale selection (left column, middle row). The SLS descriptors provide scale invariance across far greater scale differences (bottom).

tion, producing instead descriptors on a regular grid using constant, typically arbitrarily selected, scales. One such example is the efficient DAISY descriptors of [27] or, more related to this work, Dense-SIFT (DSIFT) descriptors [29].

In object recognition tasks, such regular sampling strategies for descriptor generation have been shown to outperform systems utilizing invariant features generated at stable coordinates and scales [24]. This is may be due to the benefits of having descriptors for many pixels over accurate scales for just a few.

Existing work on dense matching between two images has thus far largely ignored the issue of scale invariance. The SIFT-Flow system of [16, 17], for example, produces DSIFT descriptors at each pixel location. These descriptors are then matched between two images, taking advantage of the robustness of the SIFT representation, without attempting to provide additional scale invariance. Matching is performed using a modified optical flow formulation [8]. Although the DSIFT descriptors used by the SIFT-Flow algorithm provide some scale invariance, this quickly degrades as the scale differences between the two images increase (Fig. 2). An additional related method is the Generalized Patch-Match [2], designed for matching descriptors extracted at each pixel, here, with an emphasis on speed.

The methods described above provide means for matching descriptors produced on dense regular grids. In the absence of per-pixel scale-invariant descriptors, they are not designed to handle large scale differences. In this paper we extend these approaches by discussing the utility of multiple SIFT descriptors at each pixel, and their representations.

3. The behavior of SIFT across scales

We begin by considering how the values of multiple SIFT descriptors vary through scales. The scale space $L(x, y, \sigma)$ of an image $I(x, y)$ is defined by the convolution of $I(x, y)$ with the variable-scale Gaussian $G(x, y, \sigma)$ [13], where:

$$\begin{aligned} L(x, y, \sigma) &= G(x, y, \sigma) \star I(x, y) \\ G(x, y, \sigma) &= \frac{1}{2\pi\sigma^2} e^{-(x^2+y^2)/2\sigma^2} \end{aligned}$$

Typically (Section 2), a feature detector selects coordinates in space x, y and scale σ, from which a single SIFT descriptor $h_\sigma = h(x, y, \sigma)$ is then extracted [18]. Although sometimes more than one scale is selected, they are usually treated independently of each other.

Here, we consider instead all the descriptors $h_{\sigma_i} = h(x, y, \sigma_i)$, where σ_i is taken from a discrete set of scales $\{\sigma_1, ..., \sigma_k\}$. Our chief assumption is that corresponding pixels should exhibit a similar behavior throughout scales. In other words, the same pattern of SIFT descriptors $h(x, y, \sigma_i)$ should be apparent when examining corresponding pixels. The challenge then becomes how to effectively capture this pattern of change across scales?

3.1. SIFT sets

Rather than selecting a single scale for each pixel we compute multiple descriptors at multiple scales and represent pixels as sets of SIFT descriptors. Formally, denote by p and p' a pair of corresponding pixels in images I and I', respectively. For a set of scales $\sigma_1, \ldots, \sigma_k$, the two pixels are represented by the sets $H = [h_{\sigma_1}, \ldots, h_{\sigma_k}]$ and $H' = \left[h'_{\sigma_1}, \ldots, h'_{\sigma_k}\right]$.

To match the pixels of two images, a set-to-set similarity definition is required. There are quite a few such measures

available, e.g., [30]. As we will show in Sec. 4, however, highly accurate matching results are obtained by considering the straightforward "min-dist" measure [30], defined as follows.

$$mindist(p, p') = \min_{i,j} dist(h_{\sigma_i}, h'_{\sigma_j}). \quad (1)$$

Comparing two pixels represented as n SIFT descriptors, would require $O(128 \times n^2)$ operations, which may be prohibitive if the sets are large. Often, however, only a few scales are required to provide accurate representations (Sec. 4). This is explained by the following assumption.

Assumption 1 - Corresponding points are similar at multiple scales. Our underlying assumption is that there exist a set of scales $\sigma_1, \ldots, \sigma_k$ for image I and a set of scales $\sigma'_1, \ldots, \sigma'_k$ for image I', such that the descriptors produced at the two pixels are equal (or else sufficiently similar): $h_{\sigma_i} = h'_{\sigma'_i}$. Let $H = [h_{\sigma_1}, \ldots, h_{\sigma_k}]$ and $H' = \left[h'_{\sigma'_1}, \ldots, h'_{\sigma'_k}\right]$, then we can write $H = H'$.

This equality, however, holds only when all the scales $\sigma_1, \ldots, \sigma_k$ and $\sigma'_1, \ldots, \sigma'_k$ correspond exactly. In practice, we do not have these correspondences and instead sample the scales at fixed intervals for all images. Thus, the set of scales in one image may be interleaved with the other. Because SIFT values change gradually with scale, only few scales need to be sampled to provide similar descriptors even in such cases. This is illustrated in Fig. 3 which demonstrates SIFT values in multiple scales of two images separated by a $\times 2$ scale factor. SIFTs in the Right image match the SIFTs in the Left image by a scale offset.

3.2. SIFT subspaces

An alternative, Geometric representation for sets of SIFT descriptors, is obtained by considering the linear subspace on which these SIFTs reside. Subspaces have often been used to represent varying information. Some recent examples are listed in [4, 5]. Here, we show that low-dimensional linear subspaces are highly capable of capturing the scale-varying values of SIFT descriptors.

Assumption 2 - Descriptors computed at multiple scales of the same point span a linear subspace. The SIFT descriptor consists of gradient histograms. In many cases the local statistics of these gradients are equivalent at different scales. For example, in homogeneous, low-contrast regions or areas of stationary textures, the size of the local neighborhood does not change the distribution of gradients. In these cases we get $h_{\sigma_i} = h_{\sigma_j}$ for $\sigma_i \neq \sigma_j$.

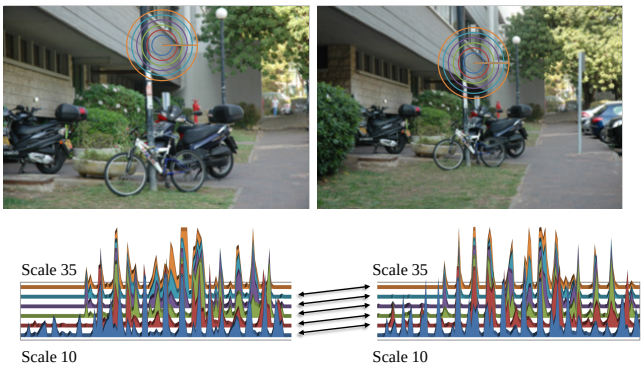

Figure 3. **SIFT behavior through scales. Top:** Two images separated by a $\times 2$ scale factor. SIFT descriptors are extracted at a low contrast area where *no* interest point was detected, at scales ranging from 10 to 35. **Bottom:** SIFT descriptor histograms. These demonstrate that (a) SIFTs from the Left image match those at higher scales in the Right, implying that setting the same scale to all pixels in both images may lead to poor matches. (b) Even in low contrast areas, SIFT values are not uniform. Finally, (c) the values of the SIFT descriptors gradually change through scales.

In other cases, the statistics do change with the scale, however, if we sample the scales densely enough these changes are gradual and monotonic (Fig. 3). In such cases we get $h_{\sigma_i} = \sum_j w_{ij} h_{\sigma_j}$, where $w_{ij} = 0$ when h_{σ_i} does not depend on h_{σ_j} and $w_{ij} = scalar$ otherwise. In other words, each descriptor can be represented as a linear combination of several other descriptors at different scales. This occurs when the regions surrounding the patch are piecewise stationary. Enlarging the window size by small steps maintains similar statistics within each window.

The observations above suggest that the set of descriptors $h_{\sigma_1}, \ldots, h_{\sigma_k}$ approximately lie on a linear subspace:

$$H = [h_{\sigma_1}, \ldots, h_{\sigma_k}] = \left[\hat{h}_1, \ldots, \hat{h}_b\right] W = \hat{H} W \quad (2)$$

where $\hat{h}_1, \ldots, \hat{h}_b$ are basis vectors spanning the space of descriptors and W is a matrix of coefficients.

Combining the two assumptions. According to assumption 1, for two corresponding pixels, if we knew the set of corresponding scales we would have $H = H'$. This implies that the two sets of descriptors share the same spanning basis, i.e., $\hat{H} = \hat{H}'$. While we do not know the scales required to construct H and H', according to assumption 2 this is not crucial. As long as we sample the scale densely enough we can compute the bases $\hat{H}$ and $\hat{H}'$.

The distance between a pair of pixels, p and p', can be measured by the distance between the corresponding subspaces $\mathcal{H}_p$ and $\mathcal{H}_{p'}$, represented as matrices $\hat{H}$ and $\hat{H}'$

with orthonormal columns. There are several possible definitions to the distance $\mathrm{dist}^2(\mathcal{H}_p, \mathcal{H}_{p'})$ between two linear subspaces [9]. Here we use the Projection Frobenius Norm (Projection F-Norm), defined as:

$$\mathrm{dist}^2(\mathcal{H}_p, \mathcal{H}_{p'}) = ||\sin\boldsymbol{\theta}||_2^2 \tag{3}$$

Where $\sin\boldsymbol{\theta}$ is the vector of sines of the principal angles between the two subspaces $\mathcal{H}_p$ and $\mathcal{H}_{p'}$. This may be computed by considering the cosines of the principal angles obtained from $SVD(\hat{H}^T\hat{H}')$ in $O(128 \times d^2)$ operations, where d is the subspace dimension.

3.3. The Scale-Less SIFT (SLS) representation

It is often beneficial to have a *point* representation for each pixel, rather than a subspace. Such is the case when, for example, efficient indexing is required. We therefore employ the subspace-to-point mapping proposed by Basri et al. [3, 4, 5] to produce the Scale-Less SIFT (SLS) descriptor for each such subspace.

Specifically, consider the subspace $\mathcal{H}_p$ produced at pixel p, represented as a $128 \times d$ matrix $\hat{H}$ with orthonormal columns. We produce the SLS representation by mapping this subspace to a point P by rearranging the elements of the projection matrix $A = \hat{H}\hat{H}^T$ using the following operator:

$$P \triangleq SLS(\hat{H}_p) = \left(\frac{a_{11}}{\sqrt{2}}, a_{12}, ..., a_{1d}, \frac{a_{22}}{\sqrt{2}}, a_{23}, ..., \frac{a_{dd}}{\sqrt{2}}\right)^T \tag{4}$$

Where a_{ij} is the element (i, j) in matrix A. A key property of this mapping is that the distance between two such mapped subspaces, P and P' is monotonic with respect to the Projection F-Norm between the original subspaces $\mathcal{H}_p$ and $\mathcal{H}_{p'}$ [4, 5]. That is:

$$||P - P'||^2 = \mu\,\mathrm{dist}^2(\mathcal{H}_p, \mathcal{H}_{p'}) \tag{5}$$

for a constant μ. Point P thus captures the behavior of SIFT descriptors throughout scale space, at a quadratic cost in the dimension of the descriptors. Here, we employ the SLS descriptor, P, as a surrogate for the subspace $\mathcal{H}_p$ without making further adjustments to the method used to compute correspondences.

4. Experiments

Our evaluation code was written in MATLAB, using the SIFT code of [29] and the SID code of [12]. Flow was estimated using the original SIFT-Flow code [16, 17], with either its original DSIFT, or alternatively using SID, and our own SLS descriptor. Our SLS results were produced using 8D, linear subspaces obtained by standard PCA. We used 20 scales at each pixel, linearly distributed in the range $[0.5, 12]$. Note that the size of the SLS representation and the matching time depends only on the dimension of the underlying SIFT descriptor (Sec. 3.3).

Quantitative results on Middlebury data [1]. We compare our SLS with both SID and DSIFT, on the Middlebury optical flow set. Since this data does not include significant scale changes, we modify it by rescaling the left and right images by factors of 0.7 and 0.2, respectively. The quality of an estimated match was measured using both angular and endpoint errors ($\pm$ SD) [1]. Table 1 shows that both multi-scale approaches outperform the single-scale DSIFT significantly. Furthermore, our SLS descriptors lead to lower errors when compared to the descriptors of [12].

Qualitative results. We present a visual comparison of the quality of the estimated flows, using each of the three alternatives: DSIFT, SID and our SLS descriptor. Our results present a Left-image (source) warped onto the Right-image (target) according to the estimated flows. SLS results in Fig. 4 and 5 are further cropped to show areas of high confidence matches (see below).

We ran tests on image pairs with independent scene motion (Fig. 4) and images of different scenes with similar appearances (Fig. 5). All these images include scale differences, often extreme. We know of no previous method which successfully presents dense correspondences on such challenging image pairs. Our results show that the SLS enables accurate dense correspondences even under extreme changes in scale.

In Fig. 4 DSIFT typically manages to lock onto a single scale quite well, while missing other scale changes in the scene. The SLS descriptor better captures the scale-varying behavior at each pixel and so manages to better match pixels at different scales with only local misalignments.

Fig. 1 and 5 present matches estimated between images of *different* scenes. A good result would have the appearance of the Left (source) images, in the scales and poses of the Right (target) images. As can be seen, the DSIFT and SID descriptors either leave the source in its original scale, unchanged, or else completely fail to produce coherent matches. Although some artifacts are visible in the SLS results (right column) the results present coherent scenes in the target image scales.

The influence of set-size and subspace dimension. We next evaluate the influence of various parameters on feature matching accuracy and run-time. Here, we use images from the Berkeley set [19]. Images were rescaled by a randomly determined scale factor, uniformly distributed in the range $[1.5 \ldots 4]$. We report the mean$\pm$SD accuracy and

Data	Angular error			Endpoint error		
	DISFT	SID	SLS	DISFT	SID	SLS
Dimetrodon	26.6±36.8	**0.16±0.3**	0.17±0.5	108.9±42.1	**0.70±0.3**	0.80±0.4
Grove2	7.06±5.7	0.66±4.4	**0.15±0.3**	59.07±40.9	1.50±5.0	**0.77±0.4**
Grove3	5.23±4.2	1.62±6.9	**0.15±0.4**	108.95±76.5	4.48±10.5	**0.87±0.4**
Hydrangea	4.24±4.5	0.32±0.6	**0.22±0.8**	33.80±32.2	1.59±2.8	**0.91±1.1**
RubberWhale	24.63±26.9	0.16±0.3	**0.15±0.3**	116.83±57.7	**0.73±1.1**	0.80±0.4
Urban2	6.24±7.6	0.37±2.7	**0.32±1.3**	54.8±54.0	**1.33±3.8**	1.51±5.4
Urban3	10.26±14.0	**0.27±0.6**	0.35±0.9	91.81±66.1	**1.55±3.7**	9.41±24.6
Venus	4.30±4.9	0.24±0.6	**0.23±0.5**	31.52±34.0	1.16±3.8	**0.74±0.3**

Table 1. **Quantitative results on rescaled Middlebury data [1].** Both angular and endpoint errors (± SD) show that multiple scales (SID [12] & SLS) are always advantageous over a single scale (DISFT [29]) with SLS outperforming SID on most data-sets.

Figure 4. **Dense flow with scene motion**. Image pairs presenting different scale changes in different parts of the scene, due to camera and scene motion. Correspondences from Left to Right images estimated using [17], comparing DSIFT, SID and our SLS, shown here cropped to the area of high confidence matches. See text for details.

Figure 5. **Dense flow between different scenes in different scales**. Objects in the results should have the same scale and position as the target, Right images and *different* from the source, Left images. See text for details.

run-time, for estimating the correspondences of pixels on regular grids, between each such image pair. Accuracy is measured as the ratio of the times a pixel's nearest neighbor is its ground truth matching pixel, to the total number of pixels. Runtime measures the time required for matching.

Fig. 6 presents the following results. **(1) Point-to-point with scale selection**: A single scale is selected for each pixel and used to extract a DSIFT descriptor. Scale selection follows [18], by choosing the extremum DoG scale, but ignoring any additional filtering. **(2) Set-to-set, variable number of scales**: Using the min-dist measure (Eq. 1) to compute pixel similarities. The number of scales sampled was varied, sampling one to ten DSIFT descriptors from scales distributed linearly in the range of $[0.5, 12]$ using the MATLAB expression `linspace(0.5,12,num_sigma);`. **(3) Subspace-to-subspace, variable number of scales**: Using the same sets as in (2) to fit a linear subspace for each pixel (using PCA). Subspace dimensions equal the number of scales sampled.

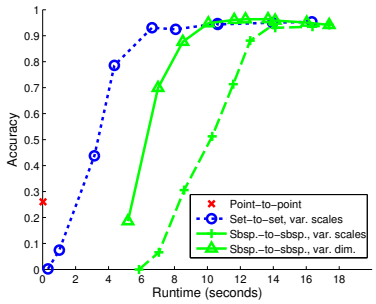

Figure 6. **Accuracy vs. runtime**. Please see text for more details.

The distance between two subspaces was computed using Eq. (3). **(4) Subspace-to-subspace, variable dimension**: Same as (3), but here 10 DSIFT descriptors were used to fit subspaces varying in dimension from 1 to 10.

From Fig. 6 it can be seen that when few scales are sampled, a single, carefully selected scale provides better performance than an arbitrarily selected scale. This advantage disappears at 3 scales; accuracy increasing rapidly with more scales sampled. By 5 scales, matching quality is near perfect for the multi-scale representations. The accuracy of the subspace-to-subspace method testifies that these SIFT sets indeed lie close to a low dimensional linear subspace. In fact, it seems that a 4D linear subspace manages to accurately capture scale varying SIFT values. We note that when a single scale is considered, the set-to-set similarity is equivalent to comparing DSIFT descriptors at an arbitrary scale and the subspace-to-subspace distance reduces to a sine similarity of these two DSIFT descriptors. Both are far worse than choosing the single scale at each pixel.

Run-times for the set-based methods are higher than comparing single points. We made no attempt to optimize our code, using built-in MATLAB functions for all our processing, and so better performance may likely be obtained. The complexity of directly comparing two sets (Sec. 3.1) or two subspaces (Sec. 3.2), however, limits the effectiveness of such optimizations. Yet although the set based methods are more computationally expensive, their significantly higher accuracy makes them an alternative worth considering.

Cropping the result to its ROI. When matching views of significantly different scales, warping one image to the other introduces the problem of cropping the image to its region of interest (ROI). In [25] this problem is avoided by assuming that the high resolution image is neatly cropped. Without this knowledge, the warped high resolution image would include noisy, "smeared" areas where it does not overlap the low resolution image (see Fig. 7).

Here we automatically select the region of high confidence matches, as follows. Given images I and I', we compute

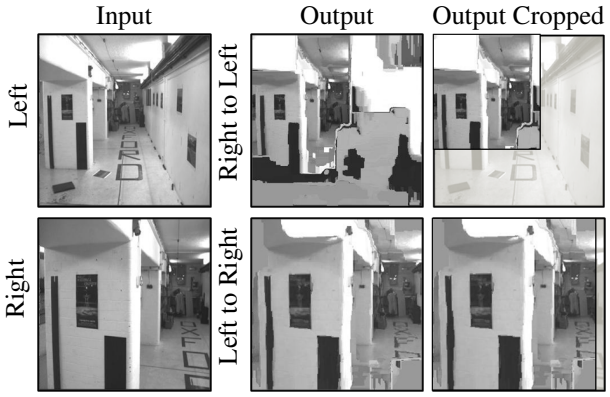

Figure 7. **Auto-crop to the ROI.** Dense matches directly formed, without estimating Epipolar Geometry, between the first and last images of the Oxford Corridor sequence [10] (left column). On the right, Notice the large areas where no information is available in the Right image to correspond with parts of the Left image. These areas are automatically cropped to include only the area onto which pixels from the second image were warped.

the two dense flows, from I to I' and then back, from I' to I. In both cases we count for each pixel in the target image, the number of source image pixels which were mapped onto it. We threshold the pixels by these numbers and then apply morphological operators to remove small clusters of target pixels. Finally, the ROI of image I is selected as the bounding box of the remaining target pixels obtained by warping image I', and vice versa. This is demonstrated in Fig. 7. No optimization was performed on this process and it is applied without modification to all our images.

5. Conclusions

The scale selection methods developed since the early 90s were largely motivated by a need to reduce computational cost as well as the assumption that few scales can be reliably matched [14]. In this paper we show that images contain valuable information in *multiple* scales. Thus, scale selection may be detrimental to the quality of the results when dense correspondences are required. The alternative, extracting SIFT descriptors at multiple scales, significantly improves results but at a computational price. We examine how such multiple scales may be compared, representing them as sets or low-dimensional, linear subspaces. In both cases multiple SIFTs outperform single descriptors in pixel matching tests by wide margins. Finally, we present a point representation for these subspaces, the SLS descriptor, which we use as a stand-in for DSIFT in the SIFT-Flow method, improving correspondences on a wide range of challenging viewing conditions.

We focus on the SIFT descriptor because of its popularity and its convenient property of changing gradually through scales. It remains to be seen how well the same approach carries over to other successful descriptors, includ-

ing DAISY [27], SURF [6], GLOH [22], and others. Extensions to affine invariance also require study. Lastly, examining the impact of this approach in other Computer Vision problems, chiefly, Object Recognition, must be explored.

Acknowledgments

Lihi Zelnik-Manor was supported in part by the Ollendorf foundation, the Israel Ministry of Science, and by the Israel Science Foundation under Grant 1179/11.

References

[1] S. Baker, D. Scharstein, J. Lewis, S. Roth, M. Black, and R. Szeliski. A database and evaluation methodology for optical flow. *Int. J. Comput. Vision*, 92(1):1–31, 2001. 5, 6

[2] C. Barnes, E. Shechtman, D. B. Goldman, and A. Finkelstein. The generalized PatchMatch correspondence algorithm. In *European Conf. Comput. Vision*, Sept. 2010. 3

[3] R. Basri, T. Hassner, and L. Zelnik-Manor. Approximate nearest subspace search with applications to pattern recognition. In *Proc. Conf. Comput. Vision Pattern Recognition*, June 2007. 5

[4] R. Basri, T. Hassner, and L. Zelnik-Manor. A general framework for approximate nearest subspace search. In *Proc. Int. Conf. Comput. Vision Workshop*, pages 109–116. IEEE, 2009. 2, 4, 5

[5] R. Basri, T. Hassner, and L. Zelnik-Manor. Approximate nearest subspace search. *Trans. Pattern Anal. Mach. Intell.*, 33(2):266–278, 2010. 2, 4, 5

[6] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool. Speeded-up robust features (surf). *Comput. Vision Image Understanding*, 110(3):346–359, 2008. 8

[7] T. Brox, C. Bregler, and J. Malik. Large displacement optical flow. In *Proc. Conf. Comput. Vision Pattern Recognition*, pages 41–48, 2009. 2

[8] A. Bruhn, J. Weickert, and C. Schnörr. Lucas/kanade meets horn/schunck: Combining local and global optic flow methods. *Int. J. Comput. Vision*, 61(3):211–231, 2005. 3

[9] A. Edelman, T. Arias, and S. Smith. The geometry of algorithms with orthogonality constraints. *SIAM Journal on Matrix Analysis and Applications*, 20:303—353, 1998. 5

[10] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521540518, second edition, 2004. 7

[11] H. Hirschmüller and D. Scharstein. Evaluation of stereo matching costs on images with radiometric differences. *Trans. Pattern Anal. Mach. Intell.*, pages 1582–1599, 2008. 1

[12] I. Kokkinos and A. Yuille. Scale invariance without scale selection. In *Proc. Conf. Comput. Vision Pattern Recognition*, pages 1–8, 2008. Available: vision.mas.ecp.fr/Personnel/iasonas/code/distribution.zip. 2, 5, 6

[13] T. Lindeberg. Scale-space theory: A basic tool for analysing structures at different scales,". *J. of App. stat.*, 21(2):225–270, 1994. 3

[14] T. Lindeberg. Feature detection with automatic scale selection. *Int. J. Comput. Vision*, 30(2):79–116, 1998. 2, 7

[15] T. Lindeberg. Principles for automatic scale selection. *Handbook on Computer Vision and Applications*, 2:239–274, 1999. 2

[16] C. Liu, J. Yuen, and A. Torralba. Sift flow: Dense correspondence across scenes and its applications. *Trans. Pattern Anal. Mach. Intell.*, 33(5):978–994, 2011. 1, 3, 5

[17] C. Liu, J. Yuen, A. Torralba, J. Sivic, and W. Freeman. Sift flow: dense correspondence across different scenes. In *European Conf. Comput. Vision*, pages 28–42, 2008. people.csail.mit.edu/celiu/ECCV2008/. 1, 3, 5, 6

[18] D. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2):91–110, 2004. 1, 2, 3, 6

[19] D. Martin, C. Fowlkes, D. Tal, and J. Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proc. Int. Conf. Comput. Vision*, volume 2, pages 416–423, July 2001. 5

[20] K. Mikolajczyk. *Detection of local features invariant to affine transformations*. PhD thesis, Institut National Polytechnique de Grenoble, France, 2002. 2

[21] K. Mikolajczyk and C. Schmid. Scale & affine invariant interest point detectors. *Int. J. Comput. Vision*, 60(1):63–86, 2004. 1, 2

[22] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *Trans. Pattern Anal. Mach. Intell.*, 27(10):1615–1630, 2005. 8

[23] J. Morel and G. Yu. Is sift scale invariant? *Inverse Problems and Imaging (IPI)*, 5(1):115–136, 2011. 1

[24] E. Nowak, F. Jurie, and B. Triggs. Sampling strategies for bag-of-features image classification. In *European Conf. Comput. Vision*, pages 490–503, 2006. 3

[25] I. Simon and S. Seitz. A probabilistic model for object recognition, segmentation, and non-rigid correspondence. In *Proc. Conf. Comput. Vision Pattern Recognition*, pages 1–7, 2007. 2, 7

[26] C. Strecha, T. Tuytelaars, and L. Gool. Dense matching of multiple wide-baseline views. In *Proc. Int. Conf. Comput. Vision*, 2003. 2

[27] E. Tola, V. Lepetit, and P. Fua. Daisy: An efficient dense descriptor applied to wide-baseline stereo. *Trans. Pattern Anal. Mach. Intell.*, pages 815–830, 2009. 1, 3, 8

[28] M. Varma and R. Garg. Locally invariant fractal features for statistical texture classification. In *Proc. Int. Conf. Comput. Vision*, pages 1–8, 2007. 2

[29] A. Vedaldi and B. Fulkerson. Vlfeat: An open and portable library of computer vision algorithms. In *Proc. int. conf. on Multimedia*, pages 1469–1472, 2010. Available: www.vlfeat.org/. 1, 3, 5, 6

[30] L. Wolf, T. Hassner, and I. Maoz. Face recognition in unconstrained videos with matched background similarity. In *Proc. Conf. Comput. Vision Pattern Recognition*, pages 529–534, 2011. 4

[31] J. Yao and W. Cham. 3D modeling and rendering from multiple wide-baseline images by match propagation. *Signal processing. Image communication*, 21(6):506–518, 2006. 2