Dense Segmentation-aware Descriptors

Eduard Trulls, Iasonas Kokkinos, Alberto Sanfeliu, Francesc Moreno-Noguer

▶ **To cite this version:**

Eduard Trulls, Iasonas Kokkinos, Alberto Sanfeliu, Francesc Moreno-Noguer. Dense Segmentation-aware Descriptors. Computer Vision and Pattern Recognition, Jun 2013, Portland, Oregon, United States. hal-00856023

HAL Id: hal-00856023
https://inria.hal.science/hal-00856023

Submitted on 30 Aug 2013

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Dense Segmentation-aware Descriptors

Eduard Trulls[1], Iasonas Kokkinos[2], Alberto Sanfeliu[1], Francesc Moreno-Noguer[1]
[1]Institut de Robòtica i Informàtica Industrial, Barcelona, Spain
[2]Center for Visual Computing, Ecole Centrale de Paris, France
[2]Galen, INRIA Saclay, France
[1]{etrulls,sanfeliu,fmoreno}@iri.upc.edu, [2]iasonas.kokkinos@ecp.fr

Abstract

In this work we exploit segmentation to construct appearance descriptors that can robustly deal with occlusion and background changes. For this, we downplay measurements coming from areas that are unlikely to belong to the same region as the descriptor's center, as suggested by soft segmentation masks. Our treatment is applicable to any image point, i.e. dense, and its computational overhead is in the order of a few seconds.

We integrate this idea with Dense SIFT, and also with Dense Scale and Rotation Invariant Descriptors (SID), delivering descriptors that are densely computable, invariant to scaling and rotation, and robust to background changes.

We apply our approach to standard benchmarks on large displacement motion estimation using SIFT-flow and wide-baseline stereo, systematically demonstrating that the introduction of segmentation yields clear improvements.

1. Introduction

Ever since the advent of SIFT [19], appearance descriptors have become an indispensable tool in matching, recognition, and retrieval, while a host of recent works such as SURF [1], ORB [26], BRIEF [8] have been developed to facilitate their use in real-time applications. A different thread of works, such as Daisy [32], dense SIFT [34] or the dense Scale-Invariant Descriptors [13] have demonstrated that it is possible to efficiently compute descriptors *densely*, i.e. for every pixel, and use them as a generic low-level image representation on a par with filterbanks.

A problem that emerges when applying dense descriptors is invariance; unlike interest points, which allow for some estimation of local scale and orientation, on arbitrary image locations scale estimation is not obvious. Some recent advances to address this problem include the treatment

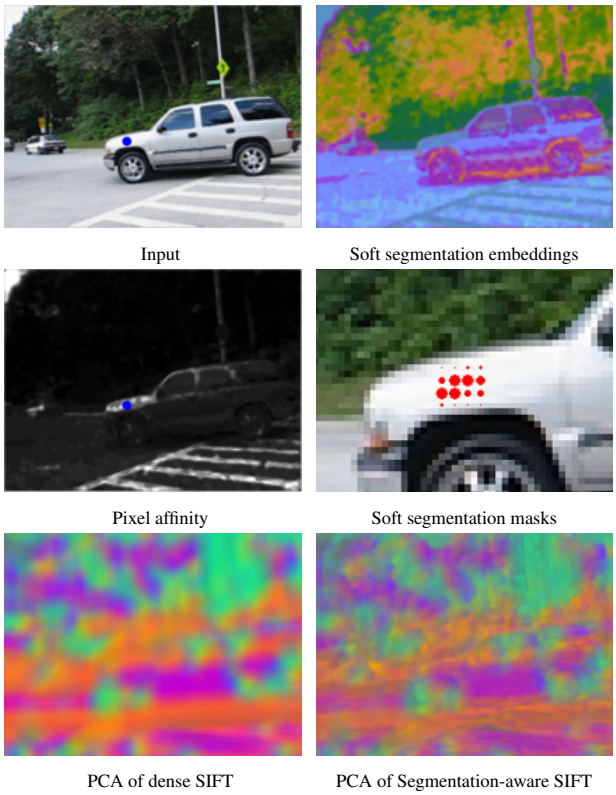

Figure 1. We exploit segmentation to construct appearance descriptors that are robust to background motion and/or occlusions. Left to right, top to bottom: (1) Source image and a feature point $\mathbf{x}$. (2) RGB encoding of the first three soft segmentation masks of [16]. (3) Segmentation-based affinity between $\mathbf{x}$ and the whole image (as per Eq.(2)). (4) Affinity values at the cells of a SIFT descriptor. (5) RGB encoding of first three principal components of dense SIFT. (6) Same as (5), but using the affinity mask in (4). We obtain much sharper features, which are built using only from the object's interior. We obtain similar results by applying this technique to the SID descriptors of [14].

This work has been partially funded by Spanish Ministry of Economy and Competitiveness under projects PAU+ DPI2011-27510 and TaskCoop DPI2010-17112; by the EU project ARCAS FP7-ICT-2011-28761; by the ERA-Net CHISTERA project VISEN; by grant ANR-10-JCJC-0205; and by the EU Project MOBOT FP7-ICT-2011-600796. E. Trulls is supported by Universitat Politècnica de Catalunya.

of scale- and/or rotation-invariance in [14, 13, 11, 27] as well as invariance to non-rigid deformations in [17, 22].

In this work we push this line of works a step further to also deal with occlusion effects, i.e. cases where the plane of the point is partially hidden by another plane lying closer to the camera. Recent work has demonstrated the merit of incorporating occlusion: [32] reported substantial performance improvements in multi-view stereo by treating occlusion as a latent variable, while [24] demonstrated that exploiting a soft figure/ground segmentation can boost the performance of a sliding-window detector.

Our main contribution in this work is a new approach to suppress background information during descriptor construction. For this we use soft segmentation masks to compute the affinity of a point with its neighbors, and shun the information coming from regions which are likely to belong to other objects.

We extract soft segmentation masks before descriptor construction, using either Normalized Cut eigenvectors [28, 20], or the Global Boundary masks of [16], with the latter coming with a minimal computational overhead. We combine this scheme with dense SIFT, and the dense Scale- and Rotation-Invariant Descriptor (SID) extraction of [13], thereby constructing a descriptor that is dense, invariant to rotations, scaling, and occlusions. We evaluate our approach on large-displacement, multi-layered motion and wide-baseline stereo. We demonstrate increased performance with respect to state-of-the art appearance descriptors: dense SIFT, dense SID, and the dense Scale-Invariant descriptors of [11]. Most importantly, we demonstrate that the introducion of segmentation results in systematically better results over the respective baselines.

Our approach is particularily simple - it involves a single parameter, that we tuned with a few images and then used throughout all experiments. It is also very efficient, introducing an overhead of a few seconds. We make our Matlab-based code publicly available from `http://vision.mas.ecp.fr/Personnel/iasonas/descriptors.html`.

2. Related work

After the seminal works of SIFT [19] and Shape Contexts [3], large strides have been made in improving performance [4, 21, 35, 29] and efficiency [1, 25, 8, 26] and decreasing memory requirements [12, 6, 30].

A complementary research direction that started from Daisy [32] and dense SIFT [34] is to extract dense image descriptors. This is motived both by experimental evidence that dense sampling of descriptors yields better performance in Bag-of-Words classification systems [23], but also from applications such as dense stereo matching which require dense features.

Other problems are less amenable to SIFT-like descriptors, including the treatment of non-rigid deformations, scale, and occlusions. Recent advances have shown that kernels based on heat diffusion geometry can effectively describe local features of surfaces subjected to non-rigid deformations and photometric changes by representing the image as a 2D surface embedded in 3D space, given the pixel coordinates and its intensity [22].

Regarding **scale**, the standard approach to accommodate scale changes is *scale selection* [19], which however is only applicable to singular points where scale can be reliably estimated. An alternative that allows to compute scale-invariant descriptors *densely* is the Scale- and rotation-Invariant Descriptor (SID) of [14], which exploits a combination of logarithmic sampling and multi-scale signal processing to obtain scale- and rotation-invariance. To achieve this the image is sampled over a log-polar grid, which turns image rotation and scaling into descriptor translations. The latter can be discarded by computing the Fourier Transform magnitude, which is unaffected by translations. The principles of Daisy were recently used in [13] to efficiently compute dense SIDs.

A more recent work on scale-invariant descriptors is the Scale-Less SIFT (SLS) of Hassner et al [11]. Their approach is to compute a set of SIFT descriptors at different scales, and then project these into an invariant low-dimensional subspace that ellicits the scale-invariant aspects of these descriptors. This descriptor comes at an increased computational price, and is not rotation-invariant by design, but gives clearly better results than dense SIFT in the presence of scaling transformations. We include also this state-of-the-art descriptor in our multi-layered motion benchmarks.

Regarding **occlusions**, there is little work around appearance descriptors. In [24], an implicit color segmentation of objects into foreground and background was used to augment histograms of gradients for people detection. The Daisy paper of [32] demonstrated clear performance improvements in multi-view stereo from treating occlusion as a latent variable and enforcing spatial consistency with Graph Cuts [5]. To deal with occlusions, a predefined set of binary masks was applied over the Daisy grid coordinates, effectively disabling half the grid at different orientations—the descriptor being a 'half moon' instead of the full circle. These masks are applied iteratively, interleaved with successive rounds of stereo matching, yielding increasingly refined depth estimates.

Our work was largely inspired by the performance improvements demonstrated in [32]. These show that separating foreground and background results in a distinct boost in performance. A marked difference with [32] is that we make this approach applicable also to the case where a single image of the scene is available; furthermore we do not constrain the masks to be of a half-moon shape, and show

how this technique can be combined with the construction of a descriptor that is invariant to scale and rotation, in addition to occlusions and background motion.

3. Segmentation-aware descriptors

3.1. Soft segmentations

Our goal is to construct appearance descriptors that are not only local, but also contained within a single surface/object ('region' from now on). In this way changes in the background, e.g. due to layered motion, will not affect the description of a point in the interior of a region. Similarly, when a region is occluded by another region in front of it, even though we cannot recover its missing information, we can at least ignore irrelevant occluders.

Our problem is connected with segmentation, where one wants to extract a partition of the image into homogeneous regions. Despite rapid progress on this area, it is understood that the problem is still far from solved. We therefore turn to algorithms that do not strongly commit to a single segmentation, but rather determine the affinity of a pixel to its neighbors in a soft manner. This soft affinity information is then incorporated into descriptor construction.

We explore two different approaches to extracting such soft segmentations. First, we use the approach of Maire et al [20]; in brief, [20] combines multiple cues to estimate a probability of boundary cue $Pb_\sigma(x, y, \theta)$, which is then used to estimate a boundary-based affinity using the 'intervening contour' technique of [28]. These local affinities are subsequently 'globalized' by finding the eigenvectors of the relaxed Normalized cut criterion. Instead of trying to form a hard segmentation out of the resulting eigenvectors we use them as pixel *embeddings* which bring closer pixels which are likely to belong to the same region and pull apart pixels which do not belong together—we thus stay closer in spirit to the Laplacian eigenmaps works of [2]. The affinity of two pixels is computed as the euclidean distance of their respective embeddings.

We also use the soft segmentation masks of Leordeanu et al [16]—there the authors use local color models constructed around each pixel to construct a large set of figure/ground segmentations. These are then projected to a lower dimensional subspace through PCA, which provides us with a low-dimensional pixel embedding. The main advantage of these features is that they are obtained at a substantially smaller computational cost.

For simplicity, we refer to the eigenvector embeddings of [20] as 'Eigen', and to the soft segmentation masks of [16] as 'SoftMask'. Fig. 2 shows the first three coordinates of the 'Eigen'/'SoftMask' embeddings as an RGB image. Note that the embeddings from Gb have higher granularity, which makes them a bit more noisy, but also better suited to capturing smaller features.

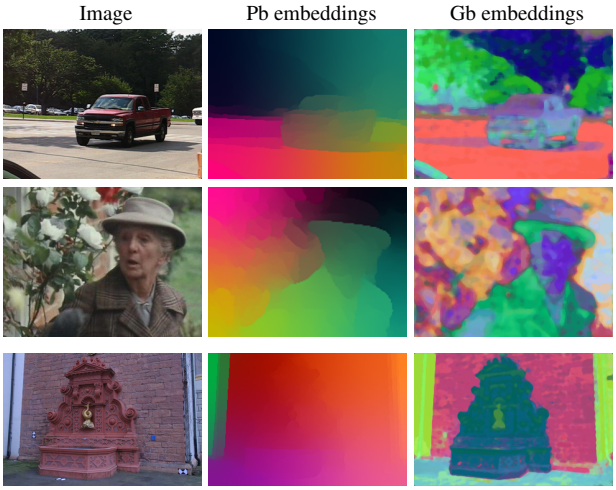

Figure 2. Soft segmentation cues: We show as RGB maps the first three coordinates of the 'Eigen' embeddings (middle column) and the 'SoftMask' embeddings (right column).

3.2. Descriptor construction

We now describe how the pixel embeddings described above can be used to render local descriptors robust to background changes and/or occlusions. Our technique is equally well applicable to dense SIFT [34], Daisy [32] and dense SID descriptors [13]; we focus on SID, as it allows us to also achieve scale- and rotation- invariance, but later on will report results for dense SIFT as well. We will make our code publicly available, and therefore refrain from providing all of the details required to reproduce the results.

We start with a brief introduction of the SID descriptor: the log-polar sampling technique of [14, 13] allows us to densely compute scale- and rotation-invariant features through the Fourier Transform Modulus/Fourier-Melin transform technique. We sample the neighbourhood around a point with a log-polar grid, defined by K rays at angles $\theta_k = 2\pi k/K$, and N rings spaced at increasing intervals. The measurements on those points are obtained after smoothing the image by Gaussian filters whose scale σ_n increases for larger radii, and extract image derivatives at 4 orientations and two polarities, using steering to have measurements aligned with the angle directions.

This results in a $K \times N$ measurement matrix for every feature channel. By design, image scalings and rotations of the image amount to translations over the radial and angular dimensions, respectively, of this descriptor. From the time-shifting property of the Fourier Transform we know that if we have a Fourier Transform pair $h(k, n) \leftrightarrow H(j\omega_k, j\omega_n)$, then

$$h(k-u, n-v) \leftrightarrow H(j\omega_k, j\omega_n)e^{-j(\omega_k u + j\omega_n v)} \quad (1)$$

This means that the Fourier Transform Magnitude (FTM) $|H(j\omega_k, j\omega_n)|$ is unaffected by signal translations; applying this observation to our descriptor matrix we realize that this provides a scale- and rotation-invariant quantity. Alternatively, we can apply the Fourier transform only over scales, to obtain a scale-invariant but rotation-dependent quantity. We will refer to the scale- and rotation-invariant descriptor as **SID** and to the scale-invariant but rotation-sensitive descriptor as **SID-Rot**.

Having provided the outline of SID, we now proceed to describe how we combine soft segmentation masks with it. Using the embeddings described in the previous subsection, we have an embedding of every pixel into a space where euclidean distances indicate how likely it is that two pixels will belong to the same region. When constructing a descriptor around a point $\mathbf{x}$ we construct the affinity $\mathbf{w}^{[i]}$ between $\mathbf{x}$ and every other point on its grid, $\mathbf{G}^{[i]}(\mathbf{x}), i = 1 \dots KN$ as follows:

$$\mathbf{w}^{[i]} = \exp\left(-\lambda \cdot d(\mathbf{x}, \mathbf{G}^{[i]}(\mathbf{x}))\right), \quad \text{where} \tag{2}$$

$$d(\mathbf{x}, \mathbf{G}^{[i]}(\mathbf{x})) = \|y(\mathbf{x}) - y(\mathbf{G}^{[i]}(\mathbf{x})\|_2^2 \tag{3}$$

In Eq. 3, $y(\cdot)$ is the embedding of point $\cdot$, and λ is a design parameter that determines the magnitude of the weighting; we experimentally determine good values for λ in Sec. 4. We then multiply these weights $\mathbf{w}^{[i]} \in [0, 1]$ with the measurements extracted around each grid point:

$$\mathbf{D'}^{[i]} = \mathbf{w}^{[i]}\mathbf{D}^{[i]}, i = 1, \dots, KN, \tag{4}$$

where for SID $\mathbf{D}^{[i]}$ is the concatenation of D convolved oriented gaussian derivatives at grid point i, while for SIFT, $\mathbf{D}^{[i]}$ are the entries of the SIFT cell positioned at $[i]$.

Multiplying by these weights effectively shuns measurements which come from the background (occluders, background planes, other objects). As such, the descriptor extracted around a point is affected only by points belonging to the same region with and remains robust to background changes. As our results indicate, this particularly simple modification yields noticeable improvements in performance.

4. Experimental evaluation

We study two different scenarios: video sequences with multi-layered motion, and wide baseline stereo. We explore the use of both of the embeddings described in 3.1, and several dense descriptors (SID, Segmentation-aware SID, dense SIFT, Segmentation-aware dense SIFT, SLS). We use the 'S' prefix to indicate 'Segmentation-aware' so for instance 'SSID' stands for our variant of SID.

For SID construction we use the implementation of [13], which adopts Daisy to compute dense features. We take $N = 28$ rays, $K = 32$ steps and $D' = 4$ derivatives, which are computed with oriented gaussian filters [10]. The derivatives preserve the polarity as in [32], so that the effective number of orientations is $D = 8$. We exploit the symmetry of the FTM to discard two quadrants, as well as the DC component, which is affected by additive lighting changes, and we normalize the resulting descriptor to have unit L_2 norm. The size of the descriptor is 3328 for SID and 3360 for SID-Rot. We refer to the publicly available code for further details.

4.1. Multi-layered Motion

We test our approach on the Berkeley Motion Dataset (Moseg) [7] which is an extension of the Hopkins 155 dataset [33]. This dataset contains 10 sequences of outdoor traffic taken with a handheld camera, three sequences of people in movement, and 13 sequences from the TV series *Miss Marple*. All of these sequences exhibit multi-layered motion. The dataset provides ground truth segmentation masks for a subset of frames in every sequence, roughly in one out of ten frames.

We evaluate *SSID* with *'Eigen'* and *'SoftMask'* embeddings against: Dense SIFT (DSIFT) [34], SLS and SID. We use SLS both in its original form and a PCA variant made publicly available by the authors: we refer to them as **SLS-paper** and **SLS-PCA**—a SLS is size 8256, whereas its PCA variant is size 528. For all the SID-based descriptors we also consider the rotation-sensitive version **SID-Rot**. We use the same parameters for both SID and SSID unless stated otherwise.

We use the 10 traffic sequences, pairing the first frame with all successive frames for which we have ground truth segmentation masks, which yields 31 frame pairs. The images are resized to 33%, in particular to permit comparison with SLS, which has high computational requirements. To take advantage of the segmentation annotations we use SIFT-Flow [18], a variant of optical flow methods which uses densely sampled SIFT descriptors instead of raw pixels to solve the correspondence problem, while preserving spatial discontinuities—this framework is publicly available and can be applied to any SIFT-like feature vector, as shown in [11]. To evaluate each descriptor we use the flow estimates to warp the segmentation mask for the second frame over the first, and compute its overlap with the ground truth using the Dice coefficient [9]. We use this experiment to determine the values for the λ parameter of SSID (Eq. (2)): $\lambda = 0.7$ for 'Eigen' and $\lambda = 37.5$ for 'SoftMask'.

Fig. 3 shows the results for all the SID and SSID variants. We observe that the rotation-sensitive variants do better, which is to be expected since the foreground elements do not contain many rotations, and discarding rotations implies a loss of information. SSID outperforms SID dramatically, in particular for large frame displacements. Fig. 4 shows the best results obtained from our approach against

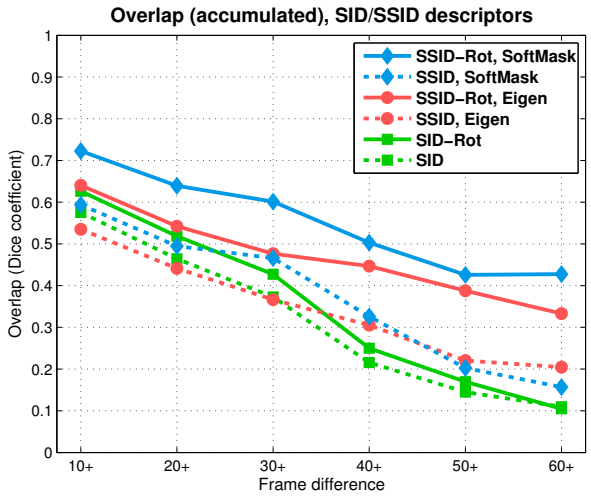

Figure 3. Overlap results over the Moseg Dataset for SID and SSID with 'Eigen' and 'SoftMask' embeddings. The results are accumulated, so the first bin includes all frame pairs, and the second bin includes frame pairs with a displacement of 20 or more frames. Each bin shows the average overlap between all the frame pairs under consideration. The following figures follow the same protocol.

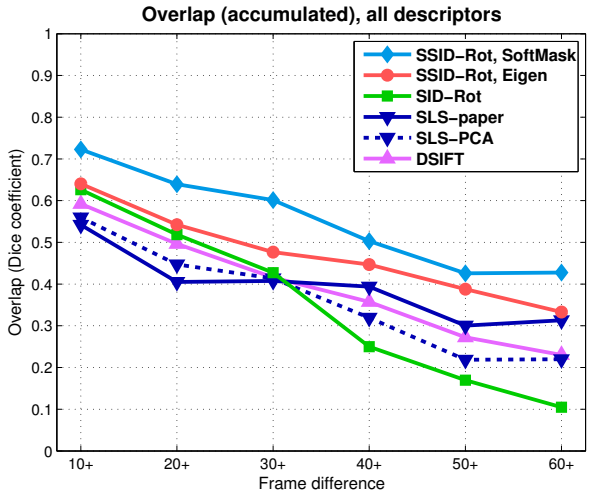

Figure 4. Overlap results over the Moseg dataset for all the dense descriptors considered. For DISFT we show the results corresponding to the best scale.

the other dense descriptors. The best overall results are obtained by SSID-Rot with 'SoftMask' embeddings, followed by the same descriptor with 'Eigen' embeddings—note that the 'SoftMask' variant does better despite its drastically reduced computational cost. Additionally, we use the flow to warp the image. Some large displacement results are shown in Fig. 5—again, SSID outperforms the other descriptors.

4.2. Segmentation-aware SIFT

The application of soft segmentation masks over SID is of particular interest because it alleviates its main shortcoming—the requirement of a large patch. But its success suggests that this approach can be applied to other standard grid-based descriptors—namely SIFT. We extend the formulation to SIFT's 4×4 grid, using the 'SoftMask' embeddings which give us consistently better results with SSID, and repeat the experiments over the Moseg dataset. Fig. 6 shows the increase in performance over three different scales. The gains are systematic, but as expected the optimal λ is strongly correlated to the descriptor size. Fig. 7 displays the performance gains. Note that this variability could be potentially accounted by the low number of samples—31 image pairs.

4.3. Wide-baseline Stereo

For stereo, we use the wide baseline dataset of [31], which contains two multi-view sets of high-resolution images with ground truth depth maps. We use the 'fountain' set, since it contains much wider baselines in terms of angular variation than the 'herzjesu' set, which exhibits mostly fronto-parallel displacements. As in the Daisy paper, we use a much smaller resolution, in our case of 460×308. For this experiment we use a set-up similar to that of [32]. We discretize 3D space into $k = 50$ bins, and use epipolar constraints and the range of the scene to restrict the candidate matches. We store the cost for the best match at every depth layer, and feed this data to a global regularization algorithm to enforce piecewise smoothness. We use Tree-Reweighted Message Passing [15] instead of Graph Cuts [5].

For a first experiment we want to evaluate the accuracy of each descriptor. We compute depth maps using this stereo algorithm and evaluate the error on every visible pixel using the ground truth visibility maps from [31]—note that this does not account for occlusion. We use the fully invariant versions of SID and SSID, as well as DSIFT, Daisy and SLS. Note that *for descriptors other than SID we align the descriptors with the epipolar lines, to enforce rotation invariance* [32]. For SLS we use only the PCA version, which has much lower dimensionality and is thus cheaper to match. The results are shown on Fig. 8. Our segmentation-aware descriptors outperform the others except for SLS—but again we do not need to rotate the patch.

Most of Daisy's performance gains on wide-baseline stereo stem from its handling of occlusions, which are not taken into account in the previous experiment. The Daisy stereo algorithm introduces an additional depth layer with a fixed cost, to account for occlusions, and exploits binary masks in an iterative process to refine the depth estimate (see 2). Note that the occlusion cost is a nuisance parameter: it can vary from one image set to another, or across

Figure 5. Large displacement image matching using SIFT flow, for different considered in this paper. We warp image 2 to image 1 using SIFT-Flow with different descriptors. The ground truth segmentation masks of image 1 are overlaid in red (a good registration should bring the object in alignment with the segmentation mask). We observe that segmentation-aware variant SSID-Rot does best, and is better than SID-Rot (please zoom in for details). Similar improvements were observed for SDSIFT over DSIFT.

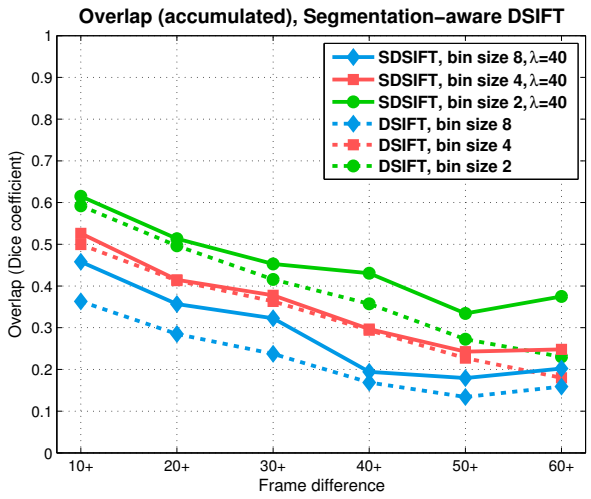

Figure 6. Overlap results over the Moseg dataset for segmentation-aware DSIFT at different scales.

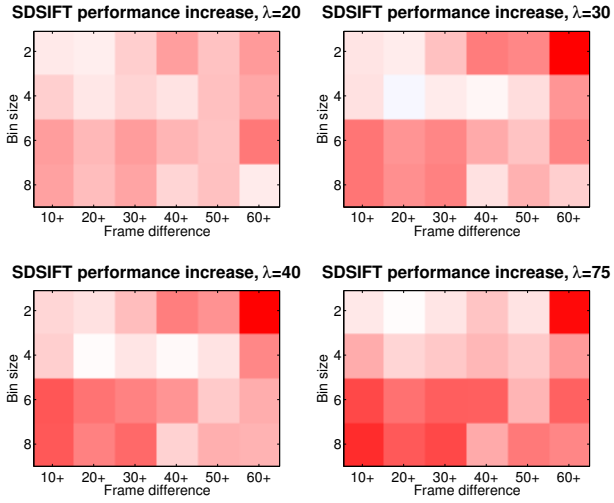

Figure 7. Increase in average overlap using our approach on DSIFT, from white (no difference in overlap) to red (largest increase in overlap, which is 0.14). For clarification, note the correspondence between the bottom left picture and Fig. 6 ($\lambda = 40$).

different baselines of the same set, and it has a drastic effect on the number of pixels marked as occluded.

We thus perform a second experiment to pitch this state-of-the-art iterative approach against our segmentation-based, single-shot approach. We run the Daisy stereo algorithm for 5 iterations, and plot the results on Fig. 9. The performance of SSID with 'Eigen' embeddings is comparable of superior to that of Daisy on most baselines—we achieve this on a single step, and without relying on the calibration data to rotate the patch. Additionally, note that we set the λ

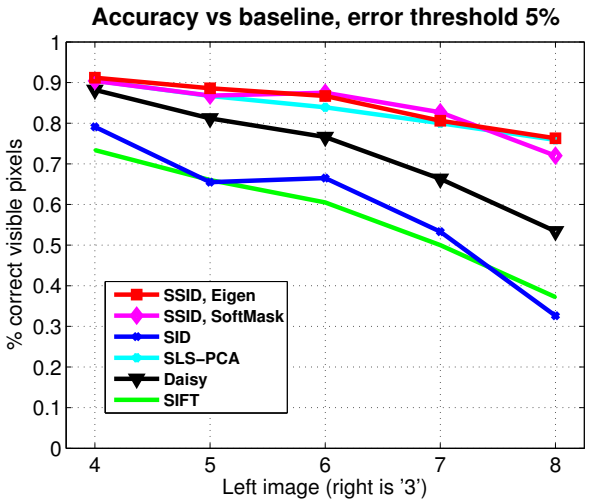

Figure 8. Accuracy at different baselines, for visible pixels only. For this figure in particular we do not consider an occlusion layer, and do not use masks for Daisy.

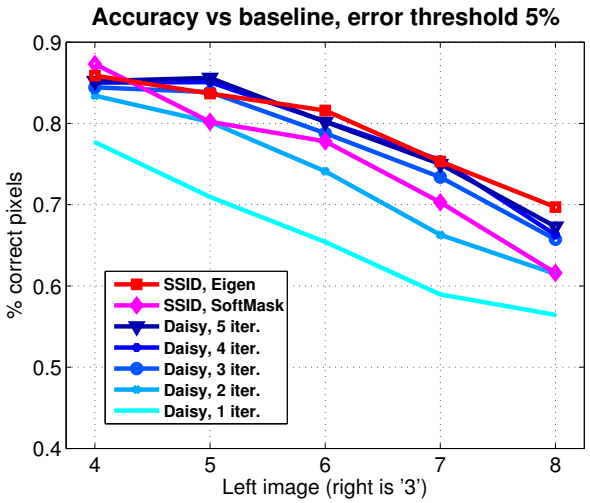

Figure 9. We compare the iterative process of Daisy to our single-shot approach. The plot shows the accuracy at different baselines (including visible pixels and occluded pixels).

parameter of Eq. (2) on the motion experiments and do not retune it for the stereo experiments. Figure 10 displays the depth estimates at two different baselines (image pairs 5-3 and 7-3)—the reference frame (3) is that on the last row of Fig. 2.

4.4. Computational requirements

The cost of computing DSIFT descriptors [34] for an image of size 320×240 is under 1 second (MATLAB/C++ code). SLS (MATLAB) requires ~21 minutes. SID (a highly optimizable MATLAB/C hybrid) takes ~71 seconds. SSID requires ~81 seconds, in addition to the extraction of the masks. Note that for all the experiments in this paper we compute the 'Eigen'/'SoftMask' embeddings at the original resolution (e.g. 640×480) before downscaling the images—the 'SoftMask' embeddings (MATLAB) require ~7 seconds, and the 'Eigen' embeddings (MATLAB/C hybrid) ~280 seconds. The computational cost of matching two images with the SIFT-flow framework depends on the size of the descriptors, varying from ~14 seconds for SIFT (the smallest) to ~80 seconds for SID/SSID, and ~10 minutes for SLS-paper (the largest).

5. Conclusions and future work

This paper presents a novel strategy to dealing with background motion and occlusions by incorporating soft segmentations into the construction of appearance descriptors. We have applied this idea to different dense descriptors, and with different methods of computing the soft segmentations, demonstrating clear improvements in all cases. We have shown improvements on the distinct tasks of wide-displacement, multi-layer optical flow, and stereo. In particular for stereo, we obtain a performance comparable to the state-of-the-art attained by the iterative version of Daisy [32], but (1) without relying on calibration data to obtain rotation-invariance, and (2) in a single step.

We believe that one of the most attractive aspects of our work is its simplicity. Our technique involves a single parameter, λ; we have set that parameter on one application with a small set of images, and ascertained its validity on a remarkably different application.

Regarding future work, an obvious first application of our work is that of object detection or classification. Our approach should be amenable to scenarios such as those of Fig. 1, but the effect of inter-class variability on our soft segmentations remains in question. Second, we are investigating the applicability of the metric learning techniques of [6, 30] to our descriptor. We believe this may not only reduce its high dimensionality, but may increase its discriminative power as well.

References

[1] H. Bay, A. Ess, T. Tuytelaars, and L. V. Gool. Surf: Speeded up robust features. *CVIU*, 2008.

[2] M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6), 2003.

[3] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *T.PAMI*, 2002.

[4] A. Berg and J. Malik. Geometric blur for template matching. In *CVPR*, 2001.

[5] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *T.PAMI*, 23(11):1222–1239, 2001.

[6] M. Brown, G. Hua, and S. Winder. Discriminative learning of local image descriptors. *T.PAMI*, 33(1):43–57, 2011.

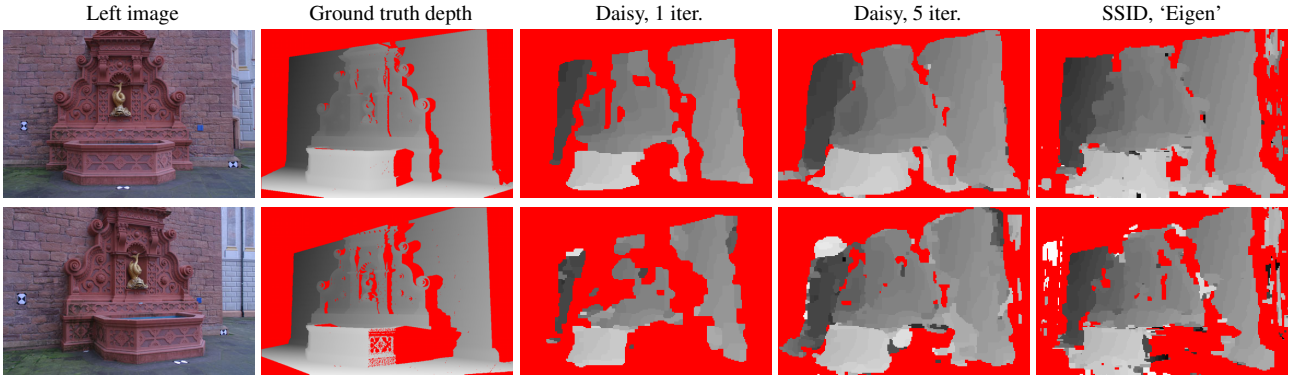

Figure 10. **First column:** The images on the left (Images 5 and 7 of [31]) are matched against Image 3 of [31], shown is Fig. 2, which serves as the the 'Right image'), for an increasing baseline. **Second column:** ground truth depth maps of [31]. **Third and fourth columns:** first and fifth iteration of the Daisy stereo algorithm. **Fifth column:** single shot depth estimation with SSID and 'Eigen' embeddings. The occlusion estimates for the first Daisy iteration may seem aggressive, but allow the algorithm to converge. Higher occlusion costs induce errors in the initial estimate and degrade the final accuracy.

[7] T. Brox and J. Malik. Object segmentation by long term analysis of point trajectories. In *ECCV*, 2010.

[8] M. Calonder, V. Lepetit, C. Strecha, and P. Fua. Brief: Binary robust independent elementary features. In *ECCV*, Heraklion, Greece, 2010.

[9] L. R. Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3), 1945.

[10] J. Geusebroek, A. Smeulders, and J. van de Weijer. Fast anisotropic gauss filtering. *Trans. Image Processing*, 12(8):938–943, 2003.

[11] T. Hassner, V. Mayzels, and L. Zelnik-Manor. On SIFTS and their scales. In *CVPR*, 2012.

[12] Y. Ke and R. Sukthankar. Pca-sift: A more distinctive representation for local image descriptors. In *CVPR*, pages 511–517, Washington, USA, 2004.

[13] I. Kokkinos, M. Bronstein, and A. Yuille. Dense Scale-Invariant Descriptors for Images and Surfaces. Technical report, Ecole Centrale Paris, Tech Report, 2012.

[14] I. Kokkinos and A. Yuille. Scale invariance without scale selection. In *CVPR*, 2008.

[15] V. Kolmogorov. Convergent tree-reweighted message passing for energy minimization. *T.PAMI*, 2006.

[16] M. Leordeanu, R. Sukthankar, and C. Sminchisescu. Efficient closed-form solution to generalized boundary detection. In *ECCV*, 2012.

[17] H. Ling and D. W. Jacobs. Deformation invariant image matching. In *ICCV*, 2005.

[18] C. Liu, J. Yuen, and A. Torralba. Sift flow: dense correspondence across difference scenes. *T.PAMI*, 33(5), 2011.

[19] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 2004.

[20] M. Maire, P. Arbelaez, C. Fowlkes, and J. Malik. Using contours to detect and localize junctions in natural images. In *CVPR*, 2008.

[21] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *T.PAMI*, 27(10), 2005.

[22] F. Moreno-Noguer. Deformation and illumination invariant feature point descriptor. In *CVPR*, 2011.

[23] E. Nowak, F. Jurie, and B. Triggs. Sampling strategies for bag-of-features image classification. In *ECCV*, 2006.

[24] P. Ott and M. Everingham. Implicit color segmentation features for pedestrian and object detection. In *ICCV*, 2009.

[25] M. Ozuysal, M. Calonder, V. Lepetit, and P. Fua. Fast keypoint recognition using random ferns. *T.PAMI*, 2010.

[26] E. Rublee, V. Rabaud, K. Konolige, and G. R. Bradski. Orb: An efficient alternative to sift or surf. In *ICCV*, 2011.

[27] U. Schmidt and S. Roth. Learning rotation-aware features: From invariant priors to equivariant descriptors. In *CVPR*, 2012.

[28] J. Shi and J. Malik. Normalized cuts and image segmentation. *T.PAMI*, 1997.

[29] K. Simonyan, A. Vedaldi, and A. Zisserman. Descriptor learning using convex optimisation. In *ECCV*, 2012.

[30] C. Strecha, A. M. Bronstein, M. M. Bronstein, and P. Fua. LDA-hash: Improved matching with smaller descriptors. *T.PAMI*, 34(1), 2012.

[31] C. Strecha, W. von Hansen, L. V. Gool, P. Fua, and U. Thoennessen. On benchmarking camera calibration and multi-view stereo for high resolution imagery. In *CVPR*, 2008.

[32] E. Tola, V. Lepetit, and P. Fua. Daisy: An efficient dense descriptor applied to wide-baseline stereo. *T.PAMI*, 32(5), 2010.

[33] R. Tron and R. Vidal. A benchmark for the comparison of 3-d motion segmentation algorithms. In *CVPR*, 2007.

[34] A. Vedaldi and B. Fulkerson. VLFeat: An open and portable library of computer vision algorithms. `http://www.vlfeat.org`, 2008.

[35] S. Winder, G. Hua, and M. Brown. Picking the best daisy. In *CVPR*, 2009.