Coarse-to-Fine Region Selection and Matching

Yanchao Yang[1,2,*], Zhaojin Lu[1,3,*], and Ganesh Sundaramoorthi[1]
[1]KAUST, Saudi Arabia [2]University of California, Los Angeles, USA [3]Institute of Automation, Chinese Academy of Sciences
{yanchao.yang, zhaojin.lu, ganesh.sundaramoorthi}@kaust.edu.sa

Abstract

We present a new approach to wide baseline matching. We propose to use a hierarchical decomposition of the image domain and coarse-to-fine selection of regions to match. In contrast to interest point matching methods, which sample salient regions to reduce the cost of comparing all regions in two images, our method eliminates regions systematically to achieve efficiency. One advantage of our approach is that it is not restricted to covariant salient regions, which is too restrictive under large viewpoint and leads to few corresponding regions. Affine invariant matching of regions in the hierarchy is achieved efficiently by a coarse-to-fine search of the affine space. Experiments on two benchmark datasets shows that our method finds more correct correspondence of the image (with fewer false alarms) than other wide baseline methods on large viewpoint change.

1. Introduction

Determining correspondence in images of a scene under wide baseline is a fundamental problem in computer vision. Applications include 3D reconstruction, motion estimation, and recognition. Although much research has been performed, nuisances in image formation such as viewpoint pose difficulties to existing methods [21].

The simplest method to establish correspondence under wide baseline uses many neighborhoods around each pixel in both images. Image data in all neighborhoods in image 1 are matched against image data in all neighborhoods of image 2. Best matches form pixel correspondences. One considers a neighborhood around each pixel since pixel values are not discriminative enough to establish unique correspondence. Multiple neighborhoods around the pixel are chosen since it is unknown a-priori what neighborhood size to choose. Small neighborhoods are not discriminative, establishing multiple matches, while large neighborhoods are subject to occlusions and may not find correspondence. Due to scale changes, one must match multiple neighborhoods of one pixel in image 1 to multiple neighborhoods at every pixel in image 2. While in principle this naive method is foolproof, it is obviously computationally intractable.

Interest point matching methods (see [29] for a survey) address the computational cost of this naive approach. These methods sample *interesting* neighborhoods from both images, reducing the number of comparisons between neighborhoods. Neighborhoods that are salient and not unlikely to match are chosen using *detectors* (e.g., [14, 17, 18, 2]). Corners and blobs are typical choices. Constant neighborhoods, unlikely to find unique correspondence, are eliminated. Neighborhood comparison must be invariant to possible transformations. This is accomplished efficiently by comparing low-dimensional *descriptor* vectors of neighborhoods (e.g., [3, 14, 19, 2, 28, 6]) that are invariant to basic transformations.

Interest point methods have had tremendous success. However, existing methods are unable to find correspondence under large viewpoint change of non-planar objects [21]. One reason for this is that detectors select only salient regions, discarding much of the image. While the detector scale parameter may be increased so that the regions' union covers the entire image, large regions typically fail to find correspondence under viewpoint change of non-planar scenes. This is because existing detectors are at most covariant to affine transformations, and large regions, which likely undergo complex transformations, do not have corresponding detections in image 2.

In this paper, we propose a new approach to wide baseline matching that addresses the limitations of interest point methods and the cost of the naive approach. We use a hierarchical decomposition of the first image's domain. Large regions are arranged at the top of the hierarchy and smaller regions are at lower levels. The union of all regions at each level covers the entire domain. Regions are not limited to salient regions. The hierarchical decomposition enables efficient coarse-to-fine traversal through regions. Our method starts with regions at the top of the hierarchy (likely most discriminative) and proceeds to regions at lower levels of the hierarchy, which are less discriminative but more likely to match. A child region is only matched to the second im-

*Joint first authors

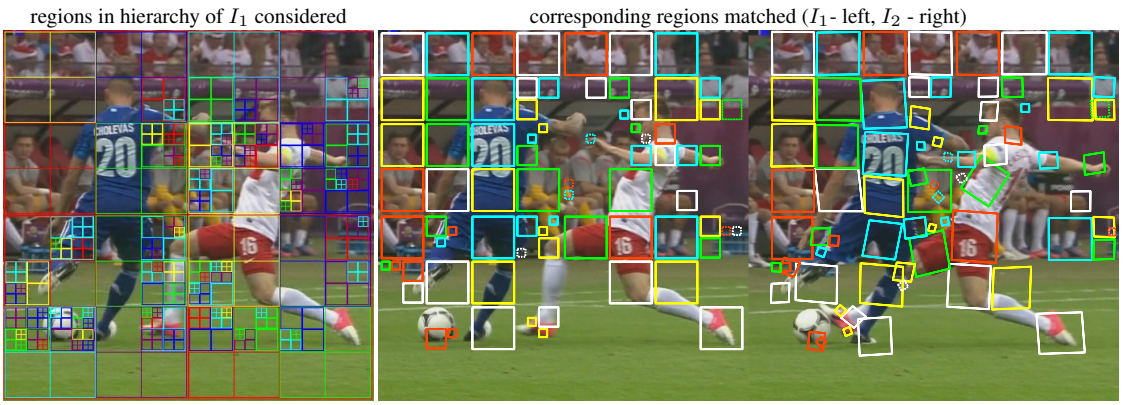


Figure 1. Coarse-to-Fine Region Selection. Our method selects regions online during matching in a hierarchical fashion: large regions are matched and sub-regions (lower levels in the hierarchy) are matched only if the large region does not match with enough fidelity. This avoids searching all neighborhoods. [Left]: regions in the hierarchy that were processed. [Right]: I_1 and I_2 with matching patches (colors show correspondence; solid boxes are correct matches, and dashed boxes are incorrect matches).

age if its parent region failed to match, eliminating the need to search through all regions. See Figure 1.

Affine transformations of regions in the hierarchy can approximate arbitrarily well any transformation arising from viewpoint change (see Theorem 1), an advantage over salient regions. Thus, we perform affine invariant matching of regions in the hierarchy. Affine invariance is typically attempted by a region normalization procedure [18], but this is not fully affine invariant as noted by [22]. We achieve affine invariance by comparing affine orbits, which are *maximal* affine invariants (see Theorem 2). Comparison of orbits is performed by efficiently traversing the affine space. This is accomplished by a hierarchical decomposition and a coarse-to-fine search of the affine space. This tailors coarse-to-fine searches introduced in object detection [9, 10] to the affine space. The idea is to construct a hierarchy of equal complexity tests. Top levels of the hierarchy are coarse tests that eliminate large portions of the parameter space. Lower levels of the hierarchy contain tests that are increasingly selective to fewer parameters. A top-down search through the hierarchy avoids an expensive linear search.

Contributions: Our main contribution is using a hierarchical decomposition of the image domain and a coarse-to-fine selection of regions for wide baseline matching. We show that the hierarchical approach finds more correct correspondence of the images (with fewer false matches) under viewpoint change than existing methods based on detectors. Further, we show how to achieve efficient affine invariant matching of regions with a hierarchical decomposition of the affine space, using ideas in [9, 10] for the affine space. Most ideas in this paper appeared in a technical report [27].

1.1. Related Work

Optical flow methods (e.g., [11, 15, 4, 25]) achieve high accuracy correspondence of all pixels. They are designed for small baseline not large displacements, large scale/rotations changes, nor when large parts of the images are occluded or disoccluded. There has been recent interest (e.g., [5]) to generalize these methods to wide baseline. Those methods use descriptor matching results as a prior to compute optical flow. They are limited by descriptor matching, which fails under large viewpoint change.

Block matching methods (see [12, 8] for surveys) are used in video compression to estimate motion. These methods tesselate the image into blocks and estimate translations of the blocks. Fixed block sizes are usually considered, but more recent methods use adaptive block sizes so that motion estimation from larger blocks can bias the motion of smaller patches. The limitation of these methods is the translational model; they are unable to cope with scale changes. Other fixed size patch matching methods include [15, 24], which match under affine but are based on local optimization, and do not work for wide-baseline. Block matching incorporating spatial regularity is used in [1] to obtain correspondence of each pixel. However, fixed size blocks are used, and thus, it is unable to address scale changes.

Recently, a fast method for matching a template under one affine warp to an image is introduced [13]. The method uses branch and bound to search the affine space. However, [13] does not address region selection. Thus, it does not directly apply to matching under viewpoint change of non-planar scenes, which induce piecewise diffeomorphisms.

2. Coarse-to-Fine Matching Algorithm

Our algorithm finds corresponding regions between two images I_1 and I_2. It consists of two hierarchies that enable coarse-to-fine search. The *region hierarchy*, decomposes the image domain into regions. The *affine hierarchy*, decomposes the affine space. The next two sub-sections describe the hierarchies and the coarse-to-fine searches.

2.1. Coarse-to-Fine Region Selection

The key properties of the region hierarchy are 1. regions at the top of the hierarchy consist of large regions and sub-regions are below in the hierarchy, and 2. any transformation arising from viewpoint change can be approximated arbitrarily well with low-dimensional (e.g., affine) transformations of regions in the hierarchy. The first property allows for efficiency in matching: any region that has found correspondence necessarily implies that all sub-regions have also found correspondence. Thus, there is no need to match sub-regions. The second property is necessary so that the transformation is approximated well. It also allows invariant region descriptors to be computed for matching, increasing efficiency.

We use a region hierarchy that contains regions that are formed by successively splitting I_1 into four equal rectangles. Each region is a node in the hierarchy and the four equal sub-rectangles are the children of the node. This hierarchy satisfies Property 2 (see Theorem 1 in Section 3). In practice, a minimum and maximum region size is chosen.

Our region selection algorithm (see Figure 1) starts with regions at the top of the hierarchy. Each region is matched to image I_2 under the hypothesis that the region transforms under the affine group. This is done by searching all neighborhoods and locations in I_2, and comparing these neighborhoods to the region by mean normalized cross-correlation (NCC). NCC is used to achieve invariance to affine contrast change. The procedure for performing this search efficiently requires another hierarchy and is described in the next sub-section. This procedure returns the first and second best affine transforms A_1 and A_2 that match the region with highest fidelity to regions in I_2.

The NCC scores s_1 and s_2 (assume $s_1 \geq s_2$) between the best affine transformed regions of I_1, and the corresponding regions in I_2 are computed. Provided that the highest score (s_1) passes a threshold $0 < T_1 \leq 1$ and the ratio of the second best to the best score is such that $s_2/s_1 < T_2$ ($0 < T_2 \leq 1$), the region is accepted as a match and the sub-regions in the hierarchy are not visited. A region that did not pass the first threshold test with enough fidelity is refined as follows. Any sub-region of the region that does not pass the first threshold test with A_1 (the transform of the region) is matched (see Figure 2). Other sub-regions are accepted.

Remark 1. *Interest point methods eliminate regions of I_1 before matching by selecting only salient regions to reduce region comparisons. In our approach, regions are eliminated online: if a region in the hierarchy has matched, no further sub-regions need to be matched, reducing region comparisons. The fact that a region in I_1 is matched to all regions in I_2 (in an efficient way) is less restrictive than covariant detectors. This implies that more of the image will find correspondence using our approach.*

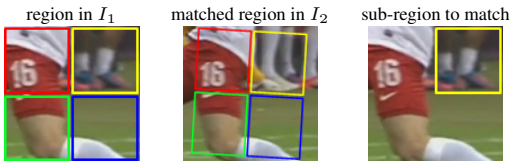


Figure 2. Sub-regions (colored squares, left image) of a region that does not match with enough fidelity according to the first threshold test, but uniquely, are not all matched. Only sub-regions that do not match with enough fidelity according to the first threshold test with the parent region's affine transform are matched.

2.2. Coarse-to-Fine Affine Search

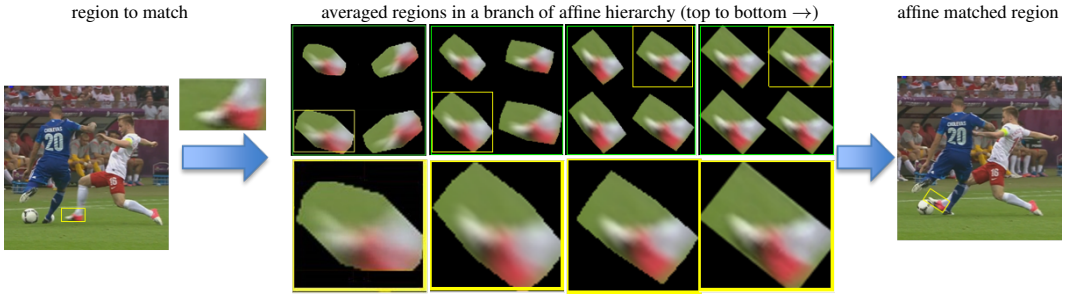
Let R be a subset of the image domain, and let $I_1|R$ denote the restriction of I_1 to R. We describe our approach to matching $I_1|R$ to a corresponding affine transformed region $I_2|A(R)$ in I_2 (see Figure 3). This requires that each location in I_2 is compared against all transformations of $I_1|R$ by the group $\mathbb{GL}(2)$, where $\mathbb{GL}(2)$ is the general linear group of non-singular 2×2 matrices. Direct search over all of $\mathbb{GL}(2)$ is expensive, and so we introduce a hierarchy to search over $\mathbb{GL}(2)$ efficiently.

In practice, we assume a finite sampling of $\mathbb{GL}(2)$. We denote by p, q parameters of a two-dimensional subset of $\mathbb{GL}(2)$, and we denote by $g_{p,q}$ the element of $\mathbb{GL}(2)$ indexed by p, q. Let $N \geq 1$, $\mathbb{P} = \{p_1, p_2, \ldots, p_{2^N}\}$ and $\mathbb{Q} = \{q_1, q_2, \ldots, q_{2^N}\}$ be the sets of the parameters p, q. The hierarchy is constructed as follows. Let L denote the number of levels in the hierarchy. $\mathbb{P}$ and $\mathbb{Q}$ are split into 2^{N-L+l} subsets of size 2^{L-l} at level l of the hierarchy. Let $\mathbb{P}_i^l, \mathbb{Q}_j^l$ denote these subsets, where $i, j \in \{1, \ldots, 2^{N-L+l}\}$. Each subset at level $l+1$ is defined to be a subset of a set at level l, i.e., $\mathbb{P}_i^{l+1} \subset \mathbb{P}_{\lfloor (i-1)/2 \rfloor + 1}^l$ and $\mathbb{Q}_j^{l+1} \subset \mathbb{Q}_{\lfloor (j-1)/2 \rfloor + 1}^l$ where $\lfloor \cdot \rfloor$ indicates the floor function. See Figure 4.

Given a node in the hierarchy at level $l-1$, we denote $\{\mathbb{P}_i^l, \mathbb{Q}_j^l\}$ the set of (four) subsets of the node in level l. We would like to construct a method to choose the subset that contains parameters of $\mathbb{GL}(2)$ which transforms $I_1|R$ to a matching region in I_2. To have the speed advantage of the hierarchy, we must do so without having to match $I_1|R$ under each individual element of each subset to I_2. To this end, we define B_{ij}^l for each subset in $\{\mathbb{P}_i^l, \mathbb{Q}_j^l\}$ as the average of affine transforms of $I_1|R$, where the affine transforms are the group elements arising from 2^{n_l} ($n_l \leq L - l$) samples in $\mathbb{P}_i^l, \mathbb{Q}_j^l$. Let $I_1|R \circ g^{-1}$ denote an affine transformed region defined on $g(R)$. Then $B_{ij}^l : \bigcap_{p \in P, q \in Q} g_{p,q}(R) \to \mathbb{R}^+$ and

$$B_{ij}^l = \frac{1}{2^{2n_l}} \sum_{p \in P, q \in Q} I_1|R \circ g_{p,q}^{-1}, \tag{1}$$

where $P \subset \mathbb{P}_i^l$ and $Q \subset \mathbb{Q}_j^l$ contain 2^{n_l} elements. The idea is that B_{ij}^l should correlate with a matching region in I_2 provided that $I_1|R \circ g_{p,q}^{-1}$ matches to I_2 for some $p, q \in \mathbb{P}_i^l, \mathbb{Q}_j^l$.



[top row]: yellow boxes are averaged regions selected; [bottom row]: the selected averaged regions (zoomed)

Figure 3. Matching of a region in I_1 is done by using a coarse-to-fine hierarchical search over the affine space. Region descriptions at the top levels of the hierarchy are highly invariant and thus respond to a wide range of affine transformed regions. Region descriptions at the bottom are less invariant but more discriminative and thus respond to only a region oriented with respect to a specific affine transform. For illustrative purposes, the hierarchy is shown for scale (vertical direction) and rotation (horizontal direction).

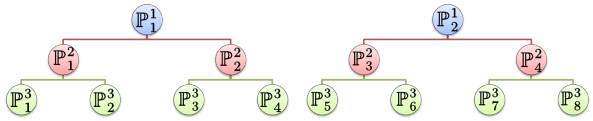


Figure 4. Illustration of the affine hierarchy ($2^N = 8$, $L = 3$). Each set $\mathbb{P}^l_{ij}$ contains the parameters of the sets below.

Further justification is given in Section 3. Note that only 2^{n_l} samples are used for computational speed. Ideally, one would average over all samples, but experimental performance shows that that is not necessary.

Now let the *response* $\mathbf{R}^l_{ij}(x)$ for each pixel x in I_2 denote the NCC between B^l_{ij} and the region of I_2 centered at pixel x. Define $i'j'$ as the indices of the subset from $\{\mathbb{P}^l_i, \mathbb{Q}^l_j\}$ that contains the highest value of the response, and x'_{ij} as a pixel of highest response, i.e.,

$$i'j' = \arg\max_{ij} \mathbf{R}^l_{ij}(x'_{ij}), \quad x'_{ij} = \arg\max_{x} \mathbf{R}^l_{ij}(x). \quad (2)$$

The subsets $\mathbb{P}^l_i, \mathbb{Q}^l_j$ where $ij \neq i'j'$ are eliminated. This narrows the search for the parameters to $\mathbb{P}^l_{i'}, \mathbb{Q}^l_{j'}$, and the procedure is repeated. This gives Algorithm 1.

Remark 2. *Each B^l_{ij} for $l = 1$ is an invariant or robust descriptor selective to a wide range of transformed regions. Lower levels of the hierarchy have descriptors that are less invariant and more selective. These descriptors are computed online in contrast to interest point methods where invariance to a predefined range of transformations (determined by the descriptor) is computed prior to matching.*

We now discuss the parameterization of $\mathbb{GL}(2)$. First, we decompose a matrix in $\mathbb{GL}(2)$ by using the QR decomposition. A matrix in $\mathbb{GL}(2)$ is a product of a non-uniform scaling, a shear in the x-direction, and a rotation:

$$\begin{pmatrix} s & 0 \\ 0 & s\lambda \end{pmatrix} \begin{pmatrix} 1 & h \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}. \quad (3)$$

Algorithm 1 Hierarchical search of the affine space.

procedure AFFINESEARCH($\mathbb{P}$, $\mathbb{Q}$, L, $I_1|R$, I_2. T_1)
  Downsample $I_1|R$ and I_2 by d for speed
  return HIERAFFSEARCH(1, $\{\mathbb{P}^1_i, \mathbb{Q}^1_j\}$)
end procedure
procedure HIERAFFSEARCH(l, $\{\mathbb{P}^l_i, \mathbb{Q}^l_j\}$)
  Compute B^l_{ij} and $\mathbf{R}^l_{ij}$ for each subset in $\{\mathbb{P}^l_i, \mathbb{Q}^l_j\}$
  Compute $i'j'$ and $x'_{i'j'}$ using (2)
  if $\mathbf{R}^l_{i'j'}(x'_{i'j'}) < T_1$ **then**
    return no match found
  else if $l \neq L$ **then**
    $\{\mathbb{P}^{l+1}_i, \mathbb{Q}^{l+1}_j\}$:= subsets below $\mathbb{P}^l_{i'}, \mathbb{Q}^l_{j'}$
    return HIERAFFSEARCH($l+1$, $\{\mathbb{P}^{l+1}_i, \mathbb{Q}^{l+1}_j\}$)
  else
    x_s := second highest local max of $\mathbf{R}^l_{i'j'}$
    $\mathbb{P}^l_{i'}, \mathbb{Q}^l_{j'}$ each have one element, p, q
    return affine transforms $(g_{p,q}, x'_{i'j'})$, $(g_{p,q}, x_s)$
  end if
end procedure

We assume $s, \lambda > 0$. We parameterize s and λ by a parameter p that goes around the s, λ plane. We similarly parameterize h and θ with a parameter q. This is done since using uniform sampling of four parameters is costly. We use a 2048 total sample size and $L = 5$ for experiments.

2.3. Cost Savings of the Affine Hierarchy

We examine the tradeoff between cost savings of the coarse-to-fine search in the affine space and the accuracy of method. Clearly, the benefit of the affine hierarchy is the reduction of a linear search to a logarithmic one. More levels in the hierarchy lead to greater computational savings. However, more levels in the hierarchy means that the regions at the top of the hierarchy are robust to a wider degree of transformations, but less discriminative. This could

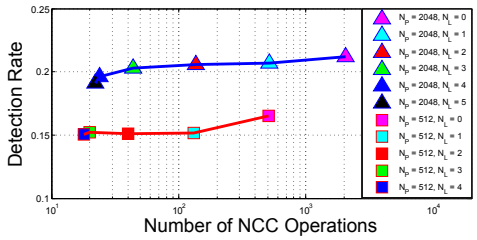


Figure 5. Accuracy vs. Computation of the Hierarchical Affine Search. The detection rate (with the false alarm rate fixed at 0.1) is plotted against the number of NCC operations used in the matching of a single region. More levels in the hierarchy (N_L) imply fewer NCC operations (proportional to computational cost) are performed. N_P indicates the sampling size of the affine space. Results are reported on the Turntable dataset [21].

lead to false matches. We investigate the tradeoff experimentally on the Turntable Dataset [21] (see Section 4). We use a range of parameters $s \in [0.8, 1.2]$, $\theta \in [-60°, 60°]$, $h \in [-0.2, 0.2]$, and $\lambda \in [0.8, 1.2]$. We explore sample sizes of 512 and 2048. We run our entire algorithm (including the search through the region hierarchy) and compare it to the region search using a direct linear search of the affine space. We choose thresholds T_1 and T_2 for the maximum detection rate (accuracy) at a false alarm rate of 0.1. We plot the detection rate versus the number of levels in the affine hierarchy, which is proportional to the number of response or NCC computations for each region. Results are shown in Figure 5. Results indicate that for a sacrifice in the detection rate of 0.01-0.02, there is a savings of nearly a factor 200 NCC operations per region. The CPU runtime on a single processor for matching the entire image using MatLab code is shown in Figure 6. The computational time of the entire algorithm is reduced by a factor of 10 in using a hierarchy of five levels rather than a direct linear search of 2048 parameters. Images are 800×600.

There are a number of speed-ups that are possible. For example, using a hierarchy in the location (translation parameter). One could then localize the NCC computation around the maximum response location (or a few highest local maximum locations) from the previous level of the hierarchy, instead of recomputing $\mathbf{R}^l_{i'j'}$ on all of I_2. Also, many parts of the algorithm can be parallelized.

3. Theoretical Justification

This section outlines the theoretical justification for using the hierarchical image decomposition, and the invariance properties of our affine matching scheme.

3.1. Justification of the Region Hierarchy

Transformations that are induced on the image plane from viewpoint change are *piecewise diffeomorphisms*:

Definition 1. *A* **piecewise diffeomorphism** ϕ *on* Ω *is*

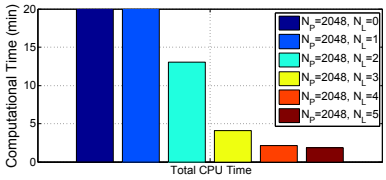


Figure 6. Computational Time vs. Number of Levels in the Affine Hierarchy. Total CPU time for full matching of two images on the Turntable dataset for various number of levels in the affine hierarchy and 2048 samples of the affine space.

1. *a partitioning of the domain* $\{R_i\}_{i=1}^n$ *(the mapped sets) and* O *the occluded set* ($R_i, O \subset \Omega$) *such that*

$$\cup_{i=1}^n R_i \cup O = \Omega,\ R_i \cap R_j = \emptyset\ (i \neq j),\ R_i \cap O = \emptyset$$

where $n \geq 1$ *is the number of regions.*

2. *and maps* $\phi_i : R_i \to \phi_i(R_i) \subset \Omega$ *such that* ϕ_i *is a diffeomorphism*

3. $\phi : \Omega \backslash O \to \Omega$ *is one-to-one*

We denote the set of all such ϕ *as PDiff*(Ω).

Although transformations relating two images under viewpoint are piecewise diffeomorphisms, within *local* regions of an image, the transformations are simpler:

Theorem 1. *Suppose that* $\phi \in$ *PDiff*(Ω) *and* $\varepsilon > 0$*, then there exists* $\{P_i\}$*, a sub-partition of* $\{R_i\}$ *and affine transformations* $A_i \in \mathbb{A}(2)$ *such that* ϕ *is approximated up to error* ε *in* C^1*-norm in each of the sets* P_i*, that is*

$$\|\phi - A_i\|_{C^1} = \sup_{x \in P_i} |\phi(x) - A_i(x)| + |D\phi(x) - DA_i(x)| < \varepsilon, \tag{4}$$

where D *denotes the Jacobian.*

Proof. We assume compactness of each R_i. Since $\phi | R_i$ is a diffeomorphism, each point in the interior of R_i has an affine transform and a neighborhood such that (4) is satisfied by Taylor's Theorem. Each point on the boundary of R_i has a neighborhood inside R_i and an affine transform such that (4) is satisfied using Whitney's Extension Theorem. By compactness, there exists a finite covering of R_i by these neighborhoods. All these neighborhoods for each R_i form a finite set $\{P_i\}$ with (4) satisfied. □

The partition $\{P_i\}$ can be approximated arbitrarily well with regions $\{P'_i\}$ that are formed by splitting Ω successively into four equal rectangles. This justifies the choice of our region hierarchy. Any piecewise diffeomorphism may be approximated arbitrarily well by a finite collection of affine transforms defined on regions obtained by splitting the image domain successively into four parts.

While existing interest point matching schemes make use of local affinity to design descriptors and detectors,

affine transforms of regions detected by detectors are not sufficient to approximate a piecewise diffeomorphism.

3.2. Invariants and the Affine Search

We show how our affine search relates to invariant descriptors. To do so, we formalize the notion of invariance following [26] (see also [30, 23]). Let $\mathcal{I}$ denote the set of images (i.e., functions of the form $I : \Omega \subset \mathbb{R}^2 \to \mathbb{R}$ for all subsets Ω). A *descriptor* is a function $F : \mathcal{I} \to \mathcal{F}$ where $\mathcal{F}$ is the description set. We are interested in descriptors that are *invariant* to nuisances in image formation. In some cases, nuisances may form a *group* (e.g., translations, rotations, and affine transformations of the domain). Characterizing certain invariants to groups can be accomplished, as we show. We denote a group by G, and an element of G by g. The action of g on image I as $I \circ g$. We formalize descriptor invariance to a group:

Definition 2 (Invariance to a Group)**.** *Let G be a group. A descriptor $F : \mathcal{I} \to \mathcal{F}$ is* **invariant** *to G if for all $I \in \mathcal{I}$ and $g \in G$, $F(I \circ g) = F(I)$.*

A constant function is an invariant descriptor, but not useful in matching. *Maximal invariants* are more useful:

Definition 3 (Maximal Invariant to a Group)**.** *A descriptor $F : \mathcal{I} \to \mathcal{F}$ that is invariant to a group G is a* **maximal invariant** *if for all $I_0, I_1 \in \mathcal{I}$, $F(I_0) = F(I_1)$ is equivalent to the existence of $g \in G$ that satisfies $I_0 \circ g = I_1$.*

Maximal invariants are important descriptors since they remove only the effect of G. Further, all other invariants are a function of the maximal invariant. Maximal invariants are related to *orbits*, which are defined as:

Definition 4 (Orbits)**.** *Let G be a group and $I \in \mathcal{I}$ be an image. The* **orbit** *of I is denoted $[I]$ and is $[I] = \{I \circ g : g \in G\}$. The set of all orbits in $\mathcal{I}$ is the* **orbit space***, and is denoted by $\mathcal{I}/G$, i.e., $\mathcal{I}/G = \{[I] : I \in \mathcal{I}\}$.*

Maximal invariants are characterized as:

Theorem 2. *Let G be a group. Define a descriptor F as $F : \mathcal{I} \to \mathcal{I}/G$ and $F(I) = [I]$. Then F is the maximal invariant with respect to G.*

Proof. Clearly, $[I]$ is invariant to G: let $g' \in G$ then $[I \circ g'] = \{(I \circ g) \circ g' : g \in G\} = \{I \circ g'' : g'' = gg', g \in G\} = \{I \circ g'' : g'' \in G\} = [I]$, where the second to last equality is obtained since multiplication by a group element is an isomorphism. Also, if $[I_1] = [I_2]$ then for each $g_1 \in G$ there exists a $g_2 \in G$ such that $I_1 \circ g_1 = I_2 \circ g_2$. Setting g_1 to the identity element yields that there exists g_2 such that $I_1 = I_2 \circ g_2$, and so the orbit is a maximal invariant. □

To determine whether $[I_1] = [I_2]$, it is enough to verify that $I_2 \in [I_1]$. This property is used in our affine search algorithm to test the equality of orbits. In fact, matching $[I_1|R]$ to $[I_2|R']$ for a region R' is done by testing whether a element of the orbit $[I_1|R]$ matches $I_2|R'$ through NCC.

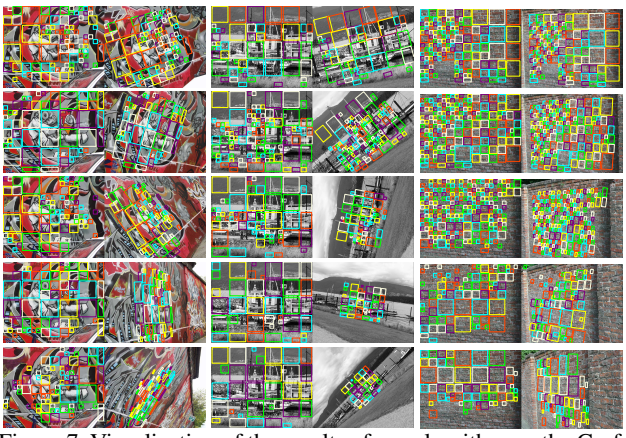


Figure 7. Visualization of the results of our algorithm on the Graffiti, Boat, and Wall datasets in the Oxford dataset. From top to bottom in each column: transformations with larger distortion. Corresponding colors indicate corresponding regions.

We now relate the averaging of $I_1|R \circ g_{p,q}$ over a subset of $\mathbb{P}, \mathbb{Q}$ in (1) to invariance and the orbit. As stated above, any function of the orbit is also an invariant though not necessarily a maximal invariant. Clearly, the integration over the orbit $\int_G I_1|R \circ g \, \mathrm{d}g$ where $\mathrm{d}g$ is the Haar measure is a function of the orbit and is thus invariant. This property is also noted in [3, 16]. Since the average over the orbit may not be discriminative enough, our algorithm averages over a limited subset G' of G as in [3]. This makes the resulting descriptor robust to small perturbations of G'. This enables matching of the descriptor to the corresponding region in I_2, provided that $I_1|R \circ g \approx I_2|R'$ where $g \in G'$.

4. Experiments

We test the performance of our algorithm on the Oxford dataset [20] and CalTech Turntable dataset [21]. These datasets are used to test the performance of wide baseline matching algorithms under large viewpoint change. The Oxford dataset concerns viewpoint change of flat scenes and in-plane transformations of non-flat scenes. Thus, we also test on the Turntable dataset, which concerns viewpoint change of non-planar objects. Code will be available[1].

We compare our algorithm to interest point methods, which are designed for wide baseline. We test many detectors including MSER, Harris-Affine, Hessian-Affine, and the SIFT DOG. We found that the performance of the SIFT descriptor to be best among many descriptors, and thus we used SIFT descriptors. We also compare to ASIFT, which is a fully affine invariant version of SIFT matching. Matching is done using the ratio test introduced in [14], where a

[1] https://site.kaust.edu.sa/ac/frg/vision

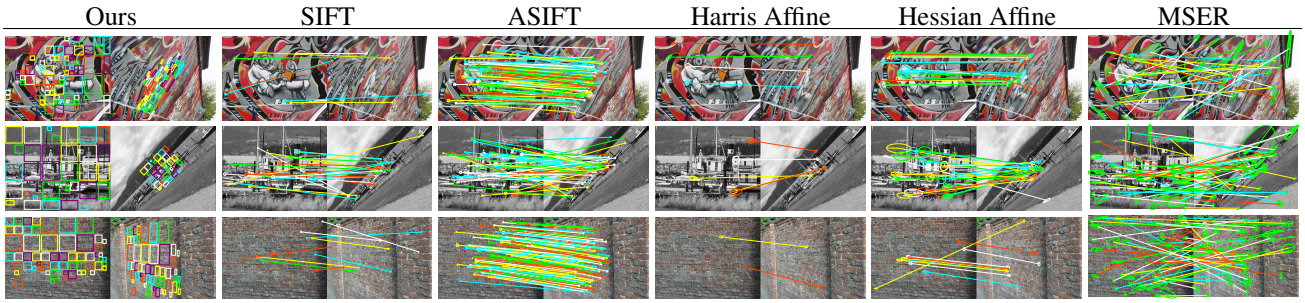


Figure 8. Visualization of matching results for methods tested. Results are shown for the Graffiti (top), Boat (middle), and Wall (bottom) datasets. Only the results on image pairs with the greatest distortion are shown. Corresponding colors indicate corresponding regions.

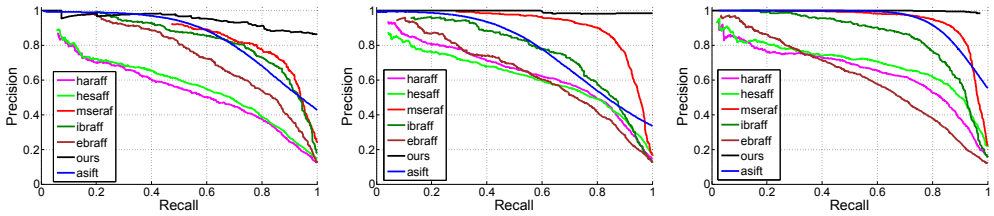


Figure 9. Precision-Recall (PR) curves for methods tested on Graffiti, Boat and Wall datasets from the Oxford dataset.

threshold T_r is used to reject matches that are not significantly better than second best matches. Note that T_r plays a similar role as the threshold T_2 in our method.

4.1. Oxford Dataset Results

We test on the Graffiti, Wall, and Boat datasets. The former two are used to test viewpoint change of flat scenes and the latter is used to test scale changes and in-plane rotations. To compare performance, we used the evaluation protocol in [19]. Methods are compared with precision-recall (PR) curves generated by varying the decision threshold. See Supplementary material for other evaluation metrics. For our method, the threshold is T_2 and for other methods the threshold is T_r. Precision measures the number of regions matched correctly versus the number of matching patches, and recall measures the regions matched correctly versus the actual number of corresponding regions. Parameters of the detectors are chosen according to the code from [19]. Ranges of parameters in our algorithm are chosen as $s \in [0.3,\ 0.9]$, $\lambda \in [0.8,\ 1.3]$, $h \in [-0.3,\ 0.3]$, and $\theta \in [-80°, 80°]$.

PR curves are shown in Figure 9. They indicate that our method is more precise for roughly any recall level than any of the other methods. Results indicate that the threshold T_2 in our method can be chosen such that the precision and recall are both close to 1, indicating that regions nearly all correctly match with few errors. The other methods' precision drop considerably at high recall. Figure 7 shows visualization of the matching results of our method. Figure 8 shows visualization of the results of all methods for the most challenging images from each of the datasets. Almost all methods except ours find almost no correct correspondence for high viewpoint change. ASIFT finds a lot of correct correspondence, but at the expense of many incorrect matches, which are filtered out by epipolar constraints. Another advantage of our method is seen visually: much of the image is covered by correctly matching regions (see Supplementary for quantification), whereas other methods only have sparse matches. Detectors detect large regions, but these regions fail to match because they are only affine covariant.

4.2. CalTech Turntable Dataset Results

Next we test the performance of algorithms on the Turntable dataset [21], which contains objects rotated on

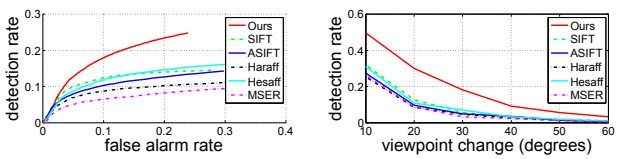


Figure 10. Quantitative results for methods tested on the Turntable dataset. [Left]: ROC curves, and [Right]: Detection rates versus angle of rotation on the turntable for various methods tested.

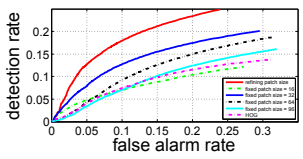


Figure 11. This experiment demonstrates the importance of our region hierarchy. ROC curves for our method and various other fixed sized region schemes are shown. Red curve is our result.

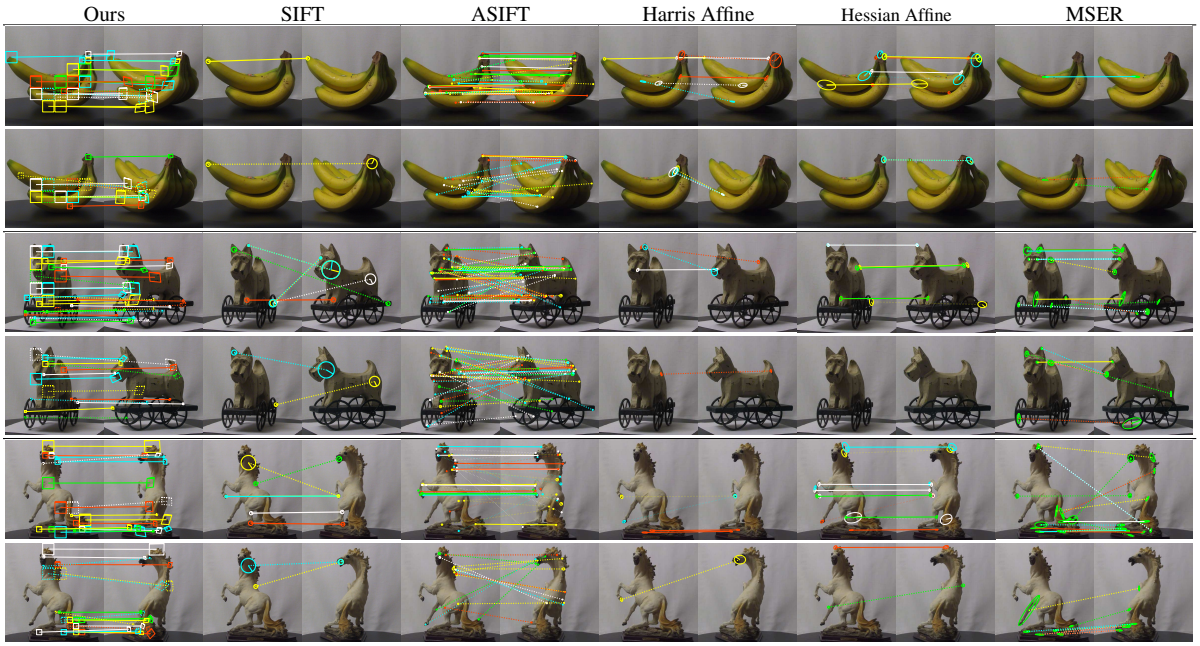


Figure 12. Sample results on the Turntable dataset. The top row of each object group are images related by viewpoint changes of 30 degrees and the bottom row is related by viewpoint changes of 60 degrees. Corresponding regions indicate corresponding regions.

a turntable every 5 degrees up to 60 degrees. The ROC framework is used to show the tradeoff between the region detection rate (rate of correctly matched regions) and false alarm rate. Matching regions are verified using epipolar constraints. See [21] for details.

The ROC curves are generated by varying the ratio threshold T_r. Detectors each have a peak threshold T_p measuring the saliency of regions detected (and controlling the number of regions detected). Each ROC curve for each method is optimized over T_p at a false alarm rate of 0.1. See [21]. We similarly construct ROC curves for our method using T_2 (analogous to T_r) and T_1. Optimal ROC curves are shown in Figure 10. This shows that our method has higher detection rate at nearly all false alarm rates. Further, the plot on the right of Figure 10 shows that our method also has higher detection rates for all viewpoints at a false alarm rate of 0.1. Sample results are visualized in Figure 12.

In the last experiment, we show that our region hierarchy, composed of regions of various sizes, is essential to obtaining the superior performance in the previous experiments. To this end, we compare our method with and without the region hierarchy. Without the hierarchy, regions are chosen with fixed sizes varying from 16-128 (tessellating the image as in a single level of our region hierarchy). These regions are matched using our affine region matching scheme. Further, to show that it is not just the affine orbit that leads to superior results, we also compare to matching with HOG descriptors [7] which are defined in fixed size regions. Figure 11 shows the quantitative results, and that our region hierarchy (with varying region sizes) leads to higher detection rates at all false alarm rates than the other schemes.

5. Conclusion

We have introduced a new approach to wide baseline matching to address large viewpoint change. Interest point methods, which are the best methods suited for wide baseline, were shown to have limited performance under large viewpoint. Interest point methods make feasible the task of comparing all regions between images to establish correspondence by sampling regions based on saliency. This eliminates regions before matching. This fundamentally restricts performance since covariant salient regions are sparse in the image. In contrast, our method increases the amount of correct correspondence found, while reducing the task of comparing all regions against regions in a different way. Specifically, our method uses a hierarchy of regions and a coarse-to-fine search to eliminate regions systematically during matching. Our method was shown to achieve affine invariant matching of regions in the hierarchy by using another hierarchical search over the affine space. Our approach was shown to out-perform existing interest point methods on large viewpoint change on two benchmark datasets, while achieving reasonable computational time and considerable speed-up compared to comparing all regions between images. There are also a number of obvious speed-ups that can be done (mentioned in Section 2.3). Future work includes exploring other region hierarchies tailored to image features, and exploiting spatial regularity between regions.

Acknowledgements

This research was partially supported by KAUST baseline funding and AFOSR FA9550-12-1-0364. We thank Moamen Mokhtar for the experiments in the Supplementary material.

References

[1] C. Barnes, E. Shechtman, A. Finkelstein, and D. Goldman. Patchmatch: A randomized correspondence algorithm for structural image editing. *ACM Transactions on Graphics-TOG*, 28(3):24, 2009. 2

[2] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool. Speeded-up robust features (surf). *Computer vision and image understanding*, 110(3):346–359, 2008. 1

[3] A. C. Berg and J. Malik. Geometric blur for template matching. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, volume 1, pages I–607. IEEE, 2001. 1, 6

[4] M. Black and P. Anandan. The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields. *Comp. vision & img. understanding*, 63(1):75–104, 1996. 2

[5] T. Brox and J. Malik. Large displacement optical flow: descriptor matching in variational motion estimation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(3):500–513, 2011. 2

[6] J. Cheng, C. Leng, J. Wu, H. Cui, and H. Lu. Fast and accurate image matching with cascade hashing for 3d reconstruction. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 1–8. IEEE, 2014. 1

[7] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE, 2005. 8

[8] G. Facciolo, N. Limare, and E. Meinhardt-Llopis. Integral images for block matching. *Image Processing On Line*, 4:344–369, 2014. 2

[9] F. Fleuret and D. Geman. Coarse-to-fine face detection. *International Journal of computer vision*, 41(1-2):85–107, 2001. 2

[10] F. Fleuret and D. Geman. Stationary features and cat detection. *Journal of Machine Learning Research*, 9(2549-2578):1437, 2008. 2

[11] B. Horn and B. Schunck. Determining optical flow. *Artificial intelligence*, 17(1):185–203, 1981. 2

[12] Y.-W. Huang, C.-Y. Chen, C.-H. Tsai, C.-F. Shen, and L.-G. Chen. Survey on block matching motion estimation algorithms and architectures with new results. *Journal of VLSI signal processing systems for signal, image and video technology*, 42(3):297–320, 2006. 2

[13] S. Korman, D. Reichman, G. Tsur, and S. Avidan. Fast-match: Fast affine template matching. In *CVPR*, pages 1940–1947. IEEE, 2013. 2

[14] D. Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004. 1, 6

[15] B. Lucas, T. Kanade, et al. An iterative image registration technique with an application to stereo vision. In *Proceedings of the 7th international joint conference on Artificial intelligence*, 1981. 2

[16] S. Mallat. Group invariant scattering. *Communications on Pure and Applied Mathematics*, 65(10):1331–1398, 2012. 6

[17] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide-baseline stereo from maximally stable extremal regions. *Image and Vision Computing*, 22(10):761–767, 2004. 1

[18] K. Mikolajczyk and C. Schmid. Scale & affine invariant interest point detectors. *International journal of computer vision*, 60(1):63–86, 2004. 1, 2

[19] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(10):1615–1630, 2005. 1, 7

[20] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Van Gool. A comparison of affine region detectors. *International journal of computer vision*, 65(1-2):43–72, 2005. 6

[21] P. Moreels and P. Perona. Evaluation of features detectors and descriptors based on 3d objects. *International Journal of Computer Vision*, 73(3):263–284, 2007. 1, 5, 6, 7, 8

[22] J. Morel and G. Yu. Asift: A new framework for fully affine invariant image comparison. *SIAM Journal on Imaging Sciences*, 2(2):438–469, 2009. 2

[23] T. Poggio. The computational magic of the ventral stream. 2011. 6

[24] J. Shi and C. Tomasi. Good features to track. In *CVPR*, pages 593–600. IEEE, 1994. 2

[25] D. Sun, S. Roth, and M. J. Black. Secrets of optical flow estimation and their principles. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2432–2439. IEEE, 2010. 2

[26] G. Sundaramoorthi, P. Petersen, V. Varadarajan, and S. Soatto. On the set of images modulo viewpoint and contrast changes. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 832–839. IEEE, 2009. 6

[27] G. Sundaramoorthi and Y. Yang. Matching through features and features through matching. *arXiv preprint arXiv:1211.4771*, 2012. 2

[28] E. Tola, V. Lepetit, and P. Fua. Daisy: An efficient dense descriptor applied to wide-baseline stereo. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(5):815–830, 2010. 1

[29] T. Tuytelaars and K. Mikolajczyk. Local invariant feature detectors: a survey. *Foundations and Trends® in Computer Graphics and Vision*, 3(3):177–280, 2008. 1

[30] A. Vedaldi and S. Soatto. Features for recognition: Viewpoint invariance for non-planar scenes. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, volume 2, pages 1474–1481. IEEE, 2005. 6