Scale-invariant Region-based Hierarchical Image Matching

Sinisa Todorovic and Narendra Ahuja
Beckman Institute, University of Illinois at Urbana-Champaign
{*n-ahuja, sintod*}*@uiuc.edu*

Abstract

This paper presents an approach to scale-invariant image matching. Given two images, the goal is to find correspondences between similar subimages, e.g., representing similar objects, even when the objects are captured under large variations in scale. As in previous work: similarity is defined in terms of geometric, photometric and structural properties of regions, and images are represented by segmentation trees that capture region properties and their recursive embedding. Matching two regions thus amounts to matching their corresponding subtrees. Scale invariance is aimed at overcoming two challenges in matching two images of similar objects. First, the absolute values of many object image properties may change with scale. Second, some of the finest details visible in the high-zoom image may not be visible in the coarser scale image. We normalize the region properties associated with one of the subtrees to the corresponding properties of the root of the other subtree. This makes the scales of objects represented by the two subtrees equal, and also decouples this scale from that of the entire scene. We also weight contributions of subregions to the total similarity of their parent regions by the relative area the subregions occupy within the parents. This reduces the penalty for not being able to match fine-resolution details present within only one of the two regions, since the penalty will be down-weighted by the relatively small area of these details. Our experiments demonstrate invariance of the proposed algorithm to large changes in scale.

1. Introduction

This paper is about matching real-world images to identify all pairs of similar subimages, e.g., for discovering and segmenting out any frequently occurring objects belonging to a visual category in a given set of arbitrary images [7]. The main contribution of this paper over previous work is the ability to perform the matching in a scale invariant manner. Thus, we would like an object to be matched across images even if its size varies – e.g., when images have been acquired at varying distances from the object. Such variations in the scale of capturing an object result in two main differences in the object's images to which matching must be made invariant. First, the absolute values of many object image properties may change with scale. Second, some of the finest details visible in the high-zoom image may not be visible in the coarser scale image. This paper presents an approach in which we achieve both of the above invariances. We extend our earlier work in which we perform image matching invariant to object orientation, illumination, and to a limited extent, to object size.

Before we outline our approach, we first very briefly review the large amount of related work in image matching based on region properties. Finding region correspondences across images is one of the fundamental problems in computer vision. It is frequently encountered in many vision tasks, such as unsupervised learning of object models [7], and extraction of texture elements (or texels) from an image texture [2]. Most approaches use only geometric and photometric properties of regions. Others improve robustness by additionally accounting for structural properties of regions. They represent images as graphs which capture region structure, and formulate image matching as graph matching problem.[4, 9, 5, 3, 6, 8, 7]. Some graph-based methods allow many-to-many region correspondences to handle possible splits or merging of regions, caused, e.g., by differences in illumination across the images[6, 8].

We define the goal of matching as finding for all regions in one image all similar regions in the other image, so the total number of matched regions and their associated similarities are maximized. The similarity measure is defined in terms of the geometric (e.g., shape, area), photometric (e.g., brightness), and structural properties (e.g., embedding of regions within larger ones). The image and object representations that we use here and in [7, 2, 8] have the following major features relevant to achieving the proposed scale invariance. Our image representation captures multiscale image segmentation via a segmentation tree. Segmentation tree captures the recursive embedding of smaller/finer regions inside bigger/more salient re-

gions. This naturally facilitates scale based analysis. To achieve the invariance to object size, in the past we have expressed certain properties of a region, corresponding to a segmentation tree node, relative to the corresponding properties of the parent region in the segmentation tree. However, such definition of the relative properties of a region retains the dependence on the properties of the root node, i.e., on the image size. In this paper, we make two main contributions. First we eliminate the mentioned dependence by normalizing the region properties with respect to the candidate object pair instead of the images. Second, we build into the matching process insensitivity to those fine level regions, which are present in the high-zoom image, but have been lost in the coarser scale image.

Overview of Proposed Approach: We extend in this paper the basic image representation, and many-to-many matching strategy presented in [8]. Our approach consists of four steps. **(1)** As in [8], images are represented by segmentation trees that capture the recursive embedding of regions obtained from a multiscale image segmentation. We associate a vector of *absolute* geometric and photometric properties of a corresponding region to each node in the tree, and thus depart from the specification presented in [8], where only relative region properties were used. In the sequel, we will refer to node and region, and descendant node and subregion, interchangeably. **(2)** Similar to the steps in [8] that allow many-to-many matching, the segmentation tree is modied by inserting new nodes, referred to as mergers. A merger represents the union of two sibling regions, whose boundaries share a part. A merger instantiates the hypothesis that a border between two regions is incorrectly detected (due to, e.g., lighting changes etc.), and therefore their union should be restored as a separate node. Each merger is inserted between its source nodes and their parent, as a new parent of the sources, and as a new child of the sources' original parent. Then, to provide access to all descendants under a node during matching, and thus improve robustness, new edges are established between each node and all its descendants, transforming the tree into a DAG. **(3)** Note that every node in the DAG defines a subgraph rooted at that node. Thus, matching any two nodes can be formulated as finding a bijection between their descendants in the corresponding subgraphs. This bijection can be characterized by a similarity measure, defined in terms of region properties. Since our goal is to identify pairs of regions with similar intrinsic and structural properties, the bijection of their subregions needs to preserve the original node connectivity and maximize the associated similarity. Formally, two nodes are matched by finding the maximum-similarity subgraph isomorphism between their respective subgraphs. In this step of our approach, we modify the algorithm of [8] in two critical ways mentioned at the beginning of this section and aimed at achieving scale invariance. First, when matching two regions, i.e., two subgraphs, we make the assumption that they represent similar object occurrences which should be matched, and normalize the absolute region properties of one of the subgraphs to the corresponding properties of the other. This modification addresses the problem of object properties being measured relative to the entire image as is the case in [8]. Second, when finding similarity of two regions, we weight the contributions of their subregions to the similarity, by the area the subregions occupy within the regions. This modification helps eliminate the direct dependency of similarity of two regions on the total number of their subregions present in the image. In turn, this helps achieve the second desirable scale property mentioned earlier in this section, since the penalty for not being able to match fine-resolution details present within only one of the two regions will be down-weighted by the rather small (relative) area of these details. **(4)** The subgraph isomorphism of a visited node pair is computed as a maximum weighted clique of the association graph constructed from the descendants. The proposed approach is validated on real images captured under large scale variations. The following sections describe the details of our algorithms.

2. Image Representation

This section presents steps (1)–(2) of our approach. An image is represented by the segmentation tree, $T{=}(V, E, \psi)$. Nodes $v{\in}V$ represent regions obtained from the multiscale segmentation algorithm presented in [1]. This algorithm partitions an image into homogeneous regions, so changes in pixel intensity within the region are smaller than those across its boundary, regardless of the absolute degree of variation. Segmentation is performed regardless of the size, shape and location of regions, and their contrasts with the surround. The multiscale regions are then organized in tree T, where the root corresponds to the entire image, and each edge $e{=}(v, u){\in}E$ represents the embedding of region u within v, i.e., their parent-child relationship. Function $\psi : V \rightarrow [0, 1]^d$ associates a d-dimensional vector, $\boldsymbol{\psi}_v$, with every $v{\in}V$, where $\boldsymbol{\psi}_v$ consists of the following absolute region properties: 1) area, 2) outer-ring area not occupied by subregions, 3) mean brightness of the outer-ring area, 4) orientation of the principal axis with respect to the image's x-axis, 5) four standard affine-invariant shape moments, and 6) centroid location. The elements of $\boldsymbol{\psi}_v$ are in $[0, 1]$.

As mentioned in Sec. 1, T is modified by insert-

ing and appropriately connecting mergers of contiguous, sibling regions, i.e., regions that share boundaries and are embedded within the same parent. The presence of mergers allows addressing the instability of low-level image segmentation under varying imaging parameters. Next, the augmented T is transformed into a DAG by adding new edges between each node and all its descendants. This allows considering matches of all descendants under a node, even when its direct children cannot find a good match. We keep the same notation T for the segmentation tree and its corresponding DAG.

3. Scale-invariant Graph Matching

Given two DAGs, T and T', we match all possible pairs of nodes $(v, v') \in V \times V'$, and estimate their similarity, $S_{vv'}$. Let $f_{vv'}=\{(u, u')\}$ denote a bijection between descendants u of v and u' of v'. The goal of our matching algorithm is to find subgraph isomorphism $f_{vv'}$ that preserves the original connectivity of subgraphs rooted at v and v', and maximizes $S_{vv'}$.

We define $S_{vv'}$ in terms of region properties of all descendants u and u' included in the subgraph isomorphism $f_{vv'}$. Invariance is achieved by normalizing properties of the subgraph rooted at v' to those of v, so as to decouple the (relative) properties of their descendents from those of the entire images, and make them compatible by expressing them relative to the normalized common reference regions. To render $S_{vv'}$ invariant to scale changes, we re-scale the area of v' to be equal to the area of $\widetilde{\boldsymbol{\psi}}_{v'}(\text{area})=\alpha \cdot \boldsymbol{\psi}_{v'}(\text{area})=\boldsymbol{\psi}_{v}(\text{area})$, and use the same scaling factor α to re-size all descendants u' of v', $\widetilde{\boldsymbol{\psi}}_{u'}(\text{area})=\alpha \cdot \boldsymbol{\psi}_{u'}(\text{area})$. This decouples the scales of objects represented by v and v' from the scales of the scenes represented by entire T and T'. To also achieve rotation-in-plane and translation invariance, we rotate and translate all descendants u' under v' by the unique delta angle and displacement which make the orientations and centroids of v and v' equal. Additionally, we make the mean brightness of v and v' equal, and then use the same delta-brightness factor to increase (or decrease) the brightness of descendants u'. This addresses local illumination changes, and decouples the brightness of objects represented by v and v' from the global brightness of the entire images represented by T and T'. There is no need to compute new shape moments of the resized regions, since they are affine invariant. This normalization yields new properties $\widetilde{\boldsymbol{\psi}}_{v'}$ and $\widetilde{\boldsymbol{\psi}}_{u'}$ of region v' and its descendants u'.

While finding the subgraph isomorphism that maximizes $S_{vv'}$, the algorithm should match subregions of v and v' whose differences in region properties are small. At the same time, it is desirable to find good matches between subregions that are perceptually salient. In general, regions that have large intensity contrasts with the surround and occupy large areas within their parent regions are perceptually salient. Thus, we define the saliency of region u, σ_u, as a relative degree of difference between the brightness and area of u from the corresponding properties of parent v, $\sigma_u \triangleq \frac{|\boldsymbol{\psi}_u(\text{brightness})-\boldsymbol{\psi}_v(\text{brightness})|}{255} + \frac{\boldsymbol{\psi}_u(\text{area})}{\boldsymbol{\psi}_v(\text{area})}$. Note that σ_u is invariant to changes in scale and illumination.

Using the above definitions of subgraph isomorphism f, absolute region properties $\boldsymbol{\psi}$, normalized properties $\widetilde{\boldsymbol{\psi}}$, and region saliency σ, we define the similarity of two regions v and v' as

$$S_{vv'} \triangleq \max_{f_{vv'}} \sum_{(u,u')\in f_{vv'}} \rho_{vv'uu'}(\sigma_u + \sigma_{u'} - |\boldsymbol{\psi}_u - \widetilde{\boldsymbol{\psi}}_{u'}|), \tag{1}$$

where the weights $\rho_{vv'uu'}$ make contributions of matches (u, u') in (1) proportional to the relative areas they occupy within v and v'. We define $\rho_{vv'uu'}$ as the total outer-ring area of u and u' that is not occupied by the other descendants of v and v' included in $f_{vv'}$, expressed as a percentage of the total area of v and v', $\rho_{vv'uu'} \triangleq \frac{\boldsymbol{\psi}_u(\text{outer-ring area})+\widetilde{\boldsymbol{\psi}}_{u'}(\text{outer-ring area})}{\boldsymbol{\psi}_v(\text{area})+\widetilde{\boldsymbol{\psi}}_{v'}(\text{area})}$. As mentioned in Sec. 1, these weights help diminish the scale effects, since the weights are small for tiny subregions most affected by scale changes.

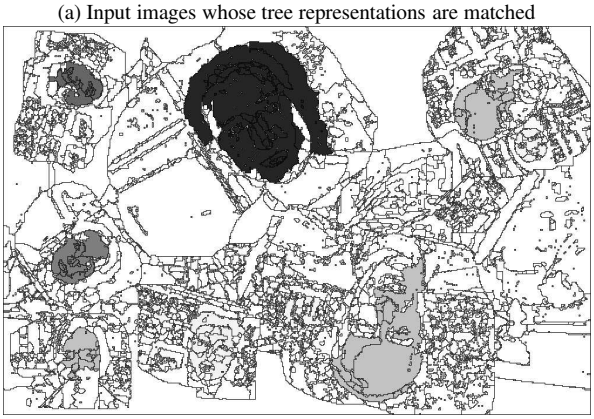
From (1), the algorithm seeks matches among the descendants of v and v' whose saliencies are high, and differences in normalized geometric and photometric properties are small. As shown in [8], the maximum similarity subgraph isomorphism f is equal to the maximum weighted clique of the association graph constructed from all descendant pairs (u, u') of v and v', which can be found by using the replicator dynamics algorithm presented in [4].

Complexity of matching two regions, i.e., subgraphs each containing no more than $|V|$ descendants, is $O(|V|^2)$. Note that the branching factor of our image DAGs varies from node to node, in proportion with the spatial variation of image structure. The DAG is not complete and the total number of subgraphs is image dependent. In experiments presented in the following section, we have observed that the number of subgraphs within an image DAG is typically $|V|/2$ (almost the same as would be in a complete binary tree containing $|V|$ nodes). Therefore, complexity of matching all pairs of regions from two DAGs is $O(|V|^4)$. It takes about 1min in MATLAB on a 2.8GHz, 2GB RAM PC.

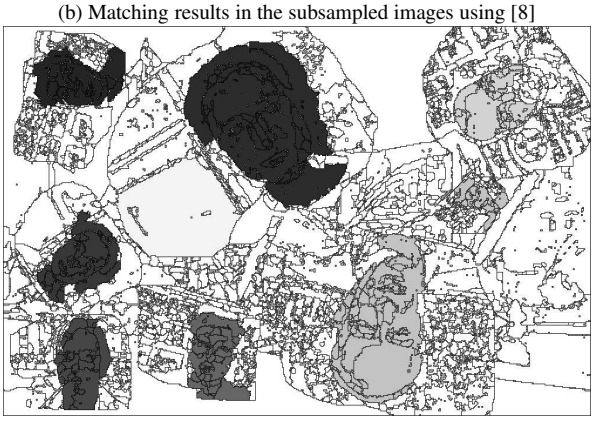
4. Results

The proposed scale-invariant matching (SIM) is evaluated on 435 faces from Caltech-101, and 80 images of UIUC 2.1D natural textures [2]. Caltech-101

(a) Input images whose tree representations are matched

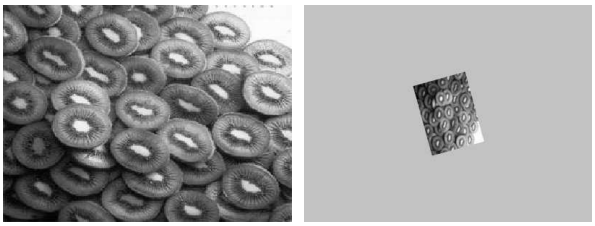


(b) Matching results in the subsampled images using [8]

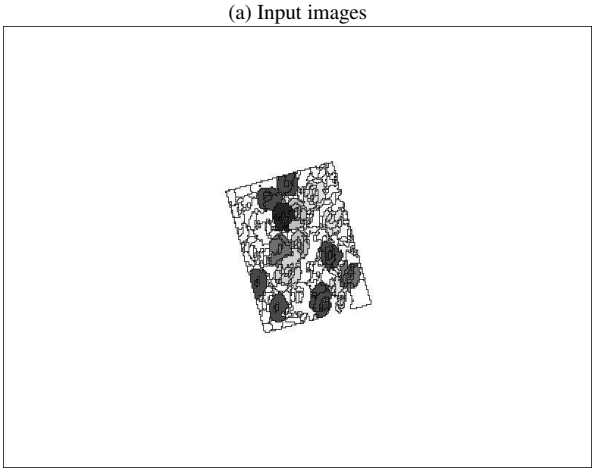


(c) Matching results in the subsampled images using SIM

Figure 1. Unsupervised image-to-image matching: (b)-(c) Sample segmentation of the image on the right in (a). Darker shades of gray indicate higher similarity of matched regions. SIM yields darker shades of faces in (c) than those in (b).



(a) Input images



(b) Matching results in the subsampled image using our algorithm

Figure 2. Model-to-image matching: (a) The texel model, learned using the image on the left, is matched with the DAG of the subsampled image on the right. (b) Darker shades of gray indicate higher similarity between model and image matches, i.e., higher confidence in texel extraction.

images contain only a single, frontally viewed face, against cluttered background, under varying illumination. The texture images present additional challenges: texels are only statistically similar to each other, they may be partially occluded, and their placement is random. When matching Caltech-101 images, our algorithm is unaware that the images contain any common objects (i.e., faces). Thus, Caltech-101 faces are used for unsupervised image-to-image matching. The texture images are used to evaluate our algorithm in a different setting, that of model-to-image matching. Since the recurrence of texels in the image is guaranteed by the assumption that the image shows texture, we first learn the hierarchical texel model, as in [2], and then match the texel model with the segmentation tree of another image showing the same texture. We report comparison only with the approach of [8], for brevity, because the algorithm of [8] has already been demonstrated to outperform the state-of-the-art matching methods on challenging benchmark datasets, like Caltech-101. To evaluate the two proposed modifications with respect to the approach of [8] – namely, normalization of absolute region properties, and similarity weighting – we additionally present results of SIM without one of these modifications, referred to as SIM$_{-N}$ and SIM$_{-W}$.

In experiments with Caltech-101 images, we randomly select a total of 10 images, where the size of one image is kept intact, while the remaining nine images are all equally subsampled. In each experiment, the amount of image-size reduction is increased in increments of 5%, until 80% of the original size. Fig. 1

demonstrates a set of these experiments, where the unaltered image (top) is matched to rotated and subsampled images, which are all placed on a random background (bottom), for brevity. As can be seen, SIM is invariant to changes in scale and illumination, in-plane rotation and translation. It improves the matching results of [8], since similarity values $S_{vv'}$ produced by SIM over regions representing common objects (i.e., faces) are larger than the similarities obtained using the approach of [8], while remaining low over background clutter. This, in turn, allows us to use SIM in solving higher-level vision problems, such as, for example, object detection. Specifically, Fig. 1 suggests that it is possible to select a suitable threshold of $S_{vv'}$ values to detect occurrences of common objects (i.e., faces) present at a wide-range of scales. For quantitative evaluation, we use the threshold that yields equal recall-precision rate, where matches are declared as true positives if the ratio of intersection and union of the matched area and the ground-truth area are greater than 0.5. Detection error includes false positives and negatives. Fig. 3 compares face detection results, averaged over 10 experiments, obtained using SIM, SIM$_{-N}$, SIM$_{-W}$ and the approach of [8]. The plots show that SIM is relatively scale invariant up until the faces become one half of the original size, and that the slope of increase in error of SIM is less than that of [8]. Also, SIM$_{-N}$ yields the worst detection results, since matching uses directly the absolute properties of regions without normalization.

In experiments with UIUC image textures, the texel model is first learned from the unaltered image texture, as in [2]. The texel model is a hierarchical graph whose nodes encode the statistical properties of corresponding texel parts, and edges capture their spatial relations. The texel model is matched to the segmentation tree of another rotated and subsampled image showing the same texture. Fig. 2 shows improvements of our approach over that of [8] in that $S_{vv'}$ values produced by SIM are larger over regions occupied by the texels than $S_{vv'}$ values obtained using the approach of [8]. By using the same detection strategy as before, we threshold matches of the texel model with subsampled images, and thus achieve texel extraction over various scales. Fig. 3 presents the detection results averaged over the 80 UIUC image textures. As the degree of subsampling increases, SIM is nearly scale invariant until the size of image texture is reduced to one half of the original, after which the slope of increase in texel-extraction error is less than that of [8]. Again, SIM$_{-N}$ yields the worst texel detection.

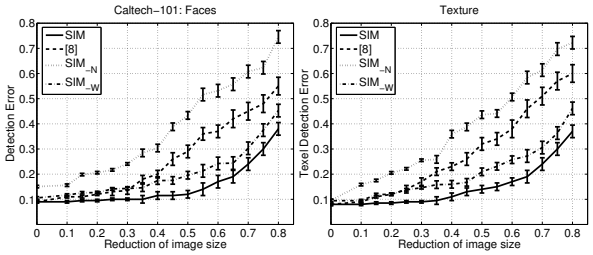


Figure 3. Face and texel detection using SIM, SIM$_{-N}$, SIM$_{-W}$ and approach [8].

5. Conclusion

We have presented an approach to region-based, hierarchical image matching that explicitly accounts for changes in region structure, including their disappearances, due to scale variations. Our experiments demonstrate that the proposed matching is indeed invariant to a wide range of scales, changes in illumination, in-plane rotation, and translation of similar objects. Experiments also suggest that our algorithm facilitates unsupervised detection of texels in a given image texture, or object-category occurrences in an arbitrary set of images.[1]

References

[1] N. Ahuja. A transform for multiscale image segmentation by integrated edge and region detection. *IEEE TPAMI*, 18(12):1211–1235, 1996.

[2] N. Ahuja and S. Todorovic. Extracting texels in 2.1D natural textures. In *ICCV*, 2007.

[3] R. Glantz, M. Pelillo, and W. G. Kropatsch. Matching segmentation hierarchies. *Int. J. Pattern Rec. Artificial Intelligence*, 18(3):397–424, 2004.

[4] M. Pelillo, K. Siddiqi, and S. W. Zucker. Matching hierarchical structures using association graphs. *IEEE TPAMI*, 21(11):1105–1120, 1999.

[5] T. B. Sebastian, P. N. Klein, and B. B. Kimia. Recognition of shapes by editing their shock graphs. *IEEE TPAMI*, 26(5):550–571, 2004.

[6] A. Shokoufandeh, L. Bretzner, D. Macrini, M. Demirci, C. Jonsson, and S. Dickinson. The representation and matching of categorical shape. *Computer Vision Image Understanding*, 103(2):139–154, 2006.

[7] S. Todorovic and N. Ahuja. Extracting subimages of an unknown category from a set of images. In *CVPR*, volume 1, pages 927–934, 2006.

[8] S. Todorovic and N. Ahuja. Region-based hierarchical image matching. *IJCV*, 78(1):47–66, 2008.

[9] A. Torsello and E. R. Hancock. Computing approximate tree edit distance using relaxation labeling. *Pattern Recogn. Lett.*, 24(8):1089–1097, 2003.

[1] **Acknowledgement:** The support of the National Science Foundation under grant NSF IIS 07-43014 is gratefully acknowledged.