21st International Conference on Pattern Recognition (ICPR 2012)
November 11-15, 2012. Tsukuba, Japan

Image Contextual Representation and Matching through Hierarchies and Higher Order Graphs

Jose C. Rubio[1], Joan Serrat[1], Antonio López[1], Nikos Paragios[2,3,4]
[1]*Centre de Visio per Computador. Universitat Autònoma de Barcelona. Bellaterra. Spain.*
[2]*Center for Visual Computing, Ecole Centrale de Paris, FR.,* [3]*Equipe Galen, INRIA Saclay, Ile-de-France, FR*
[4]*Universit Paris-Est, LIGM (UMR CNRS), Center for Visual Computing, Ecole des Ponts ParisTech, FR.*

Abstract

We present a region matching algorithm which establishes correspondences between regions from two segmented images. An abstract graph-based representation conceals the image in a hierarchical graph, exploiting the scene properties at two levels. First, the similarity and spatial consistency of the image semantic objects is encoded in a graph of commute times. Second, the cluttered regions of the semantic objects are represented with a shape descriptor. Many-to-many matching of regions is specially challenging due to the instability of the segmentation under slight image changes, and we explicitly handle it through high order potentials. We demonstrate the matching approach applied to images of world famous buildings, captured under different conditions, showing the robustness of our method to large variations in illumination and viewpoint.

1 Introduction

Image region matching has been widely used for applications such as 3D surface registration [13], object retrieval [4] and place recognition [12]. The main challenge when matching image regions lies in encoding the large variability in both appearance and arrangement of regions obtained from natural images under changing conditions (viewpoint, illumination). The existing segmentation algorithms are often sensitive to small changes, generating image regions with poor repetitivity in terms of shape and size, and fusing and dividing regions inappropriately.

A common approach to tackling these variability problems is the use of graph matching. Generally, the graph nodes represent the features to be matched, while the edges encode information about their spatial structure. Most graph matching methods establish geometric constraints on the image structure by preserving a distance measure between nodes embedded in a Euclidean space [8]. One major drawback of this approach is its restriction to a near-isometric assignment, which results in poor performance when there are large variations in the node arrangements. One way to overcome this limitation is to rely on the statistical properties of the graph. Commute times between graph nodes have been recently used in computer vision applications to characterize the layout of a graph, proving to be stable against structural variations of the scene. In [1, 9], commute-times are successfully applied to describe the structure of the image content.

Typically, only *one-to-one* correspondences are considered in graph-matching [7], which is a very restrictive assumption. A much less explored problem, called *many-to-many* matching, involves simultaneously matching multiple vertices in one graph to multiple vertices in the other. In the case of region matching this is a critical issue, since the inconstant pattern of regions generated by a segmentation algorithm produces frequent region fusions and divisions.

Works like [3] have noted the convenience of combining *many-to-many* matching with an abstract model of the scene structure. Following this idea, we propose a coarse-to-fine hierarchical representation that encodes the image global structure in a loose manner using commute times, while relying on local invariant features to handle photometric changes of the image. The bottom level of the hierarchy deals with *many-to-many* correspondences by comparing subsets of region boundary shapes.

The graph matching is modeled as a high-order discrete energy minimization. Because solving such high-order functions is a difficult problem with existing optimization techniques, we propose very sparse high order potentials which, as shown in [10], can be efficiently optimized with standard message passing inference algorithms.

978-4-9906441-0-9

2 Hierarchical Region Matching

We address the problem of matching two instances of the same object, which can undergo significant changes in illumination, viewpoint, and scale. As a consequence of these changes, regions in two different images that correspond to the same object, when segmented, display poor repetitivity. In addition, occlusions are frequent and fusions and divisions of regions from one image to the next may occur.

Our image representation contains information at two levels of abstraction. The lower level or *region-graph* contains image regions generated by a standard segmentation algorithm such as *mean-shift*. The upper level, or *cluster-graph*, groups regions into potential semantic objects, with similar appearance and location. While the cluster level encodes the layout of the semantic components of the image, the region level improves the correspondences by examining local features within these components.

2.1 Region Layer

From each image we obtain a set of regions using a standard segmentation algorithm (e.g mean-shift). We define them as $r_i = (xy_i, T_i, H_i)$, where xy_i denotes the position of the region centroid, and T_i and H_i its texture and hue descriptors respectively. In order to express the region layout, we define a distance between regions as

$$D_r(r_i, r_j) = \alpha f(xy_i, xy_j) + (1-\alpha) g(T_i, T_j), \quad (1)$$

where f is the Euclidean distance between region centroids, g is the Euclidean distance between their texture descriptors (e.g. local binary pattern), and $\alpha \in \{0, 1\}$ is a weighting factor which adjusts the importance of location or appearance in the graph structure. The local region structure (graph adjacency matrix Ω_r) is defined by connecting nodes which are adjacent to each other in the image, with weight D_r.

2.2 Cluster Layer: graph of commute times

We group the image regions with the LP-based stabilities clustering method introduced in [5], using the distance of equation (1) as the affinity matrix. Each of the clusters acts as a node on the cluster-graph, composing a collapsed version of the region-graph. We define a distance between clusters by averaging all region edge weights between two clusters as

$$D_c(c_i, c_j) = \frac{1}{|c_i||c_j|} \sum_{r_k \in c_i} \sum_{r_l \in c_j} \Omega_r(r_k, r_l). \quad (2)$$

The cluser-graph is a fully connected graph, and its adjacency Ω_c is given by the commute time matrix between clusters. As shown in [2], this can be computed from the spectrum of the normalized Laplacian of the distance matrix of Eq. (2).

3 Formulation

Let R_1 and R_2 be the sets of regions generated from two images, I_1 and I_2 respectively. The set of all possible correspondences between the regions is denoted by $R = R_1 \times R_2$. Analogously, let C_1 and C_2 represent the sets of clusters of regions, and the set of all correspondences between clusters as $C = C_1 \times C_2$. A matching configuration between two images is represented as a binary vector of indicator variables $x \in \{0, 1\}^{C \cup R}$, where $x = 1$ means that the corresponding matching exists, and $x = 0$ otherwise. Our energy function is defined as the weighted sum of 4 energy terms,

$$E = \lambda_c E^c + \lambda_r E^r + \lambda_{ct} E^{ct} + \lambda_{shape} E^{shape}, \quad (3)$$

where the λ coefficients weight the influence of each of the terms. The *c* and *r* terms refer to the appearance cost on the cluster and region layer respectively. The *ct* term favors a coherent cluster structure on pairs of cluster associations, and the *shape* is a high order term with two different roles: First, it encourages subsets of segments to match other subsets with similar shape, and second, it communicates between both layers by constraining the clusters and segments that can be matched simultaneously. Next, we develop the potential function formulation for each of the terms.

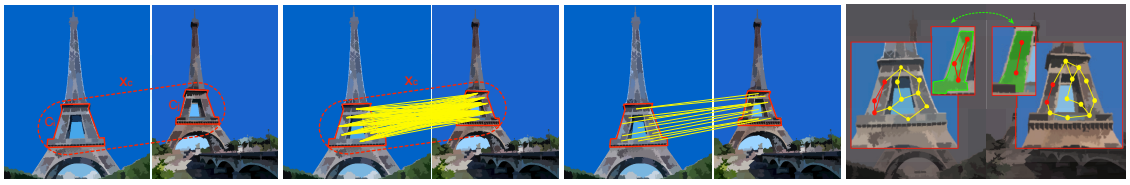
3.1 Potential Functions

The term E^c encourages correspondences between clusters with similar appearance, by assigning a cost to match two clusters as the average distance between the descriptors (color, texture) of the regions contained in the clusters.

We denote the set N of pairs of associations involved in a fusion or division as every pair $x_a = (c_i, c_j), x_b = (c_k, c_l)$ such that $c_i = c_k \vee c_j = c_l$. The set M contains the remaining pairs of cluster associations $x_c = (c_i, c_j), x_d = (c_k, c_l)$ such that $c_i \neq c_k \wedge c_j \neq c_l$. Then, the term E^{ct} is defined as

$$E^{ct}(x) = \sum_{x_a, x_b \in N} \theta^{fd}_{ab} x_a x_b + \sum_{x_c, x_d \in M} \theta^{ct}_{cd} x_c x_d \quad (4)$$

The first cost θ^{fd} encourages fusions or divisions among clusters that are very close to each other, in terms of centroid distances and appearance. The θ^{ct} penalizes



(a) (b)

Figure 1. In (a) left, an example of High Order clique ($\mathbf{x}_c$). In (a) center, all possible associations within the clique. In (a) right, a low cost clique realization. In (b), an example of sub-graph matching of region boundaries. Best Viewed in color.

pairs of associations with dissimilar commute time distances between the source and destination cluster graph edges. The energy term E^r is a unary measure of appearance compatibility between image regions.We have adapted the *self-similarity* descriptor introduced in [11] to describe the compatibility between two regions. Finally, the high order potential E^{shape} binds both graph layers by constraining the cluster and region correspondences that should not be active simultaneously. It also models the compatibility of many-to-many region matching by defining a shape similarity between subsets of regions. We describe it in detail in the next section.

3.2 High Order Potential

For every cluster association $x_c = (c_i, c_j) \in C$, we generate a high order clique $\mathbf{x}_{ij}$, composed of all region correspondences within the two clusters c_i and c_j, and the cluster correspondence itself x_c. Let Z be the set of all high-order cliques. The high order energy term is defined as

$$E^{shape}(x) = \sum_{\mathbf{x}_{ij} \in Z} \Psi(\mathbf{x}_{ij}). \tag{5}$$

The optimization of the high order terms is a difficult problem to solve [6]. In order to make the optimization feasible it is very convenient to model potential functions which are sufficiently sparse [10], so that only a few configurations contribute with a low cost, while the remaining configurations have a high constant penalty. We can take advantage of our hierarchical model to make our terms very sparse, and at the same time unify the matching of both layers, by assigning large costs to incompatible layer configurations. Given a clique $\mathbf{x}_{ij} = \{x_c, x_k, x_l, ..., x_n\}$, where x_c is the cluster association indicator variable, and $x_k, x_l, ..., x_n$ represent the correspondences between the regions, we define the high order potential as

$$\Psi(\mathbf{x}_{ij}) = \begin{cases} \theta_\infty & \text{if } x_c = 0 \wedge \sum_k^n x_k > 0 \\ \theta_0 & \text{if } x_c = 0 \wedge \sum_k^n x_k = 0 \\ s(\mathbf{x}_c) & \text{otherwise} \end{cases} \tag{6}$$

The term θ_∞ is a high penalty that forbids *incoherent* inter-layer configurations. The term θ_0 is a constant cost when no region is matched in the clique. The function $s(\mathbf{x}_c)$ defines a cost for each *compatible* configuration of region correspondences $\mathbf{x}_c = \mathbf{x}_{ij} - \{x_c\}$, within the two clusters associated by the variable $x_c = 1$. This is shown in Figure 1 (a).

Let G_i be the set of all possible subgraphs of Ω_r, obtained by combining the regions $r_k, r_l, ..., r_n \in c_i$, in p-tuples, from $p = 1$ to $p = n$. Each of these combinations is then characterized by a shape descriptor $SC(g_i)$, of the points in the boundary of the image area defined by the subgraph combination $g_i \in G_i$. See Figure 1 (b).

Given a variable set $\mathbf{x}_c = \{x_a, x_b, ..., x_n\}$ related to the cluster correspondence $c = (c_i, c_j)$, and their respective sets of combinations G_i, G_j, we define the *shape* cost as

$$s(\mathbf{x}_c) = \delta(SC(g_i), SC(g_j)) + \rho. \tag{7}$$

The cost s is defined for every p-tuple $g_i = \{r_1, ..., r_p\} \in G_i$, and every q-tuple $g_j = \{r'_1, ..., r'_q\} \in G_j$, with $a = (r_1, r'_1), b = (r_1, r'_2), ..., n = (r_p, r'_q)$. The function δ is a similarity measure between shapes (*shape context*). The term ρ is an extra penalty for unmatched regions, defined as $\rho = 1 - (1/min(|c_i|, |c_j|) \sum_{a \in \mathbf{x}_c} x_a)$.

4 Experiments and Results

We evaluate our method on pairs of images of world-famous monuments downloaded from Flickr. The monument classes selected are: *Eiffel Tower, Liberty Statue, Notre Dame cathedral* and *Taj Mahal*. Figure 2 shows examples for each of the classes. The image dataset is composed by 20 images per class. We consider a pair of regions to be a correct correspondence if the regions share a minimum of 50% of the common pixel area of two objects. We compute the *region mismatch error* as the ratio of failed correspondences to the total number of correct region pairs.

To perform these experiments we set the energy component weights to $\lambda_c = 3, \lambda_r = 5, \lambda_{ct} = 4.5$ and

Figure 2. Sample results of three pairs of images. Regions in each pair sharing the same color are in correspondence. Areas in black are not matched. Best viewed in color.

Table 1. Error ratio and optimization avg. time.

mismatch error ratio	Eiffel	Liberty	T.Mahal	N.Damme
spectral pairwise [7]	0.34	0.45	0.51	0.35
ours pairwise	0.46	0.58	0.67	0.53
ours H.O	0.25	0.17	0.30	0.21
optimization time (seconds)				
ours pairwise	11.2	04.8	17.9	23.5
ours H.O	02.3	00.8	06.1	05.9

$\lambda_{shape} = 2.5$. The cost θ_0 is set to 1.25. Figure 2 shows examples of region matching results.

Table 1 shows a comparison of the average error ratio per class. We compare our algorithm with and without high order terms, to the pairwise spectral matching presented in [7]. Removing the high order terms results in a less expressive configuration of region correspondences, which tends to produce matchings between large *patches* of regions, usually belonging to the same cluster. The main handicap of [7] is the one-to-one correspondence constraint, which is very restrictive in the context of region fusions and divisions.

The homogeneous and repetitive texture and color in some monuments misleads the appearance costs, resulting in an aliasing effect where regions in one image are matched to similar incorrect regions in the other image. Adding high order terms solves this problem and, surprisingly, the average computation time is reduced. The explanation is found in these same potentials, which add a constant cost to all the energy variable space, with the exception of the allowed configurations. The energy then tends to fall into the local minima of these low cost variable realizations, and the optimization quickly converges within this restrictive variable space.

We have observed that some mismatched regions are due to a symmetric image structure. This is true for instance, for the bottom left and bottom right vegetation areas surrounding the Eiffel tower. The effect is mainly caused by the commute time graph layer, which imposes equal penalties to clusters that are symmetrically arranged.

5 Conclusions

In this work we have presented a novel approach for image region matching. Our method first encodes the image information with a coarse-to-fine hierarchy graph, whose levels embed different abstractions of the image contents. Our method explicitly handles *many-to-many* correspondences by scoring the similarity of subsets of region boundaries, which are encoded in high order energy terms. The commute times metric solves the structural noise problem of graph-based representations. We demonstrate the effectiveness of our method performing matching on images of monuments. The results prove the matching to be robust against large variations on illumination, and viewpoint.

References

[1] R. Behmo, N. Paragios, and V. Prinet. Graph commute times for image representation. *IEEE CVPR*, 2008.

[2] F. Chung and S.-T. Yau. Discrete green's functions. *J. Comb. Theory Ser. A*, 2000.

[3] S. J. Dickinson, A. Shokoufandeh, Y. Keselman, M. F. Demirci, and D. Macrini. Object categorization and the need for many-to-many matching. 2005.

[4] J. Kim and K. Grauman. Asymmetric region-to-image matching for comparing images with generic object categories. *IEEE CVPR*, 2010.

[5] N. Komodakis, N. Paragios, and G. Tziritas. Clustering via lp-based stabilities. In *NIPS*, 2008.

[6] X. Lan, S. Roth, D. P. Huttenlocher, and M. J. Black. Efficient belief propagation with learned higher-order markov random fields. In *ECCV*.

[7] M. Leordeanu and M. Hebert. A spectral technique for correspondence problems using pairwise constraints. In *In ICCV*, 2005.

[8] J. J. McAuley, T. S. Caetano, and A. J. Smola. Robust near-isometric matching via structured learning of graphical models. In *NIPS*, 2008.

[9] H. Qiu and E. R. Hancock. Graph simplification and matching using commute times. *Pattern Recognition*, 2007.

[10] C. Rother, P. Kohli, W. Feng, and J. Jia. Minimizing sparse higher order energy functions of discrete variables. *IEEE CVPR*, 2009.

[11] E. Shechtman and M. Irani. Matching local self-similarities across images and videos. In *IEEE CVPR*, 2007.

[12] A. Toshev, J. Shi, and K. Daniilidis. Image matching via saliency region correspondences. *IEEE CVPR*, 2007.

[13] Y. Zeng, C. Wang, Y. Wang, X. Gu, D. Samaras, and N. Paragios. Dense non-rigid surface registration using high-order graph matching. *IEEE CVPR*, 2010.