Depiction Invariant Object Matching

Anupriya Balikai
anupriyabalikai@gmail.com
Peter Hall
pmh@cs.bath.ac.uk

Department of Computer Science,
University of Bath
Bath,
UK

Matching objects no matter how they are depicted (photo, painting, drawing, etc.) is an important open problem. We propose that the way in which images are described is a key to matching performance. To test this we use a hierarchical descriptor with regions as node to encode structure. The nodes are labelled with photometric descriptors in one case, and with non-photometric descriptors in the other. Measuring performance across a photos-only database yields comparable results, but we see a marked improvement for the non-photometric descriptor when either an art-only or a mixed database is used. We further improve performance using an MRF based matcher of our own design.

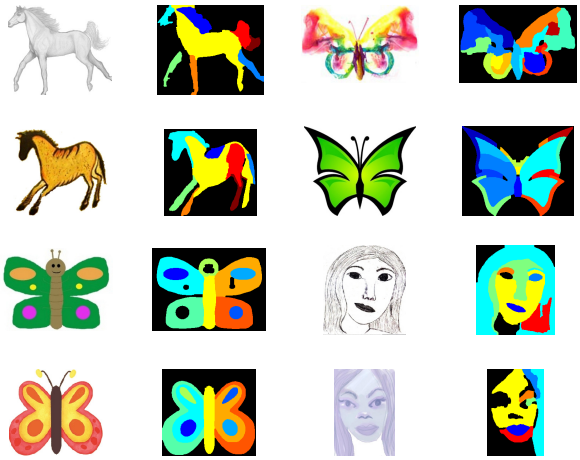


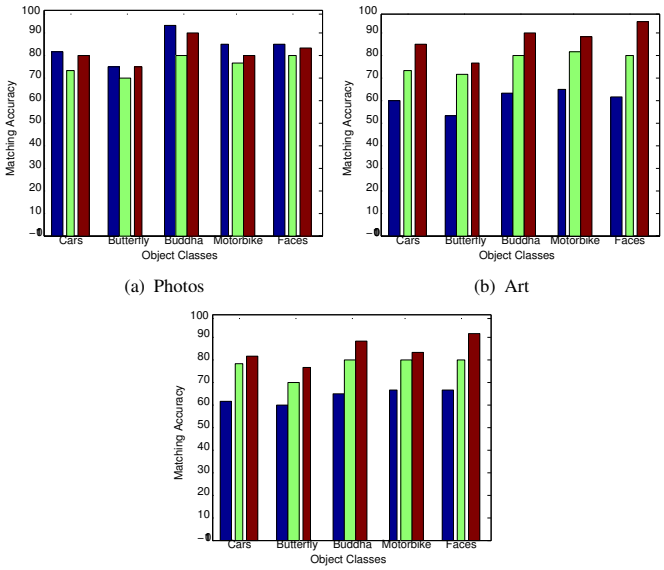
Figure 1: Matching across depictions using structure and depiction-invariant features. Each pair of images is colour-coded to show regions that have been matched using the proposed method.

We are motivated by the human ability to recognise across depictive boundaries. Object matching is an area that has received continuous and consistent attention within Computer Vision. The state of the art is now robust to challenges such as occlusions, viewpoint invariance, pose invariance and so on. However, little attention has been given to matching objects across depictive styles.

There has been little work on matching across depictive styles [2, 3]. Our approach to object matching differs to these methods by finding a generic representation for an object, without the need for any learning. The main contribution of this paper lies in the introduction of an object descriptor that combines global structure and local non-photometric features.

Our object description is based on a labelled graph. An object's structure is encapsulated by a graph constructed using the output of any hierarchical segmentor. In our experiments we find that the Berkeley segmentor [1] output the best segmentations. Since the number of levels in the hierarchy are quite large, we reduce the levels in the hierarchy by using the Laplacian Energy of the graph as explained in [4]. Every region in the segmentation tree is assigned as a vertex in the graph. Vertices at the same hierarchical level are connected by an edge if they share a common boundary. Vertices across consecutive levels are connected by an edge if they are related by containment. To remove noise from the graph we have come up with novel yet simple technique based on discarding regions that lie below a certain area threshold. A recursive algorithm is introduced to ensure that meaningful regions are not discarded. Each region in the hierarchy is described using self similarity descriptors [2] augmented by geometric terms relating to shape and orientation. We now have full description for the entire graph, and hence the object.

Two methods are presented for matching across pairs of images. The first method, based on Feature Correspondence Graph Matching [5], is used to compute a mapping between the segmentation graphs of the two objects. However, inconsistent segmentation of different instances



(a) Photos (b) Art

(c) Mixed

Figure 2: Matching accuracies obtained for graph matching with photometric features (blue), graph matching with SSD features (green) and sliding window search with SSD features (red), for 5 object categories over the three datasets.

of an object brings about failure cases with this method. To counter this problem, a second new approach to object matching is introduced based on a sliding window search that finds best matches for the segmented regions in the first image, across the second image. The overall match is then computed using max-sum on a Markov Random Field [6].

Experiments were conducted on three datasets of images. called *photos only*, *art only* and *mixed*. The first being a subset of the Caltech-101 object categories dataset, while the next two have been introduced in this paper. The accuracy of the matcher is measured against human labelled ground truth. In order to provide a comparison between matching photos to photos, and depiction- invariant matching, experiments included the use of photometric and SSD features using the proposed matchers.

Figure 2 shows that matching performance is contingent upon representation: our descriptor is comparable to the state of the art photometric methods for the dataset of photographs alone, but outperforms the state of the art for the other two datasets.

We conclude that description is important to the problem of cross-domain matching, and that learning is not necessary to achieve that task.

[1] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik. From contours to regions: An empirical evaluation. In *Computer Vision and Pattern Recognition (CVPR)*, 2009.

[2] E. Shechtman and M. Irani. Matching local self-similarities across images and videos. In *Computer Vision and Pattern Recognition (CVPR)*, 2007.

[3] Abhinav Shrivastava, Tomasz Malisiewicz, Abhinav Gupta, and Alexei A. Efros. Data-driven visual similarity for cross-domain image matching. *SIGGRAPH ASIA*, 2011.

[4] Y.Z. Song, P. Arbelaez, P. Hall, C. Li, and A. Balikai. Finding semantic structures in image hierarchies using laplacian graph energy. *European Conference on Computer Vision (ECCV)*, 2010.

[5] L. Torresani, V. Kolmogorov, and C. Rother. Feature correspondence via graph matching: Models and global optimization. *European Conference on Computer Vision (ECCV)*, 2008.

[6] T. Werner. A linear programming approach to max-sum problem: A review. *Pattern Analysis and Machine Intelligence (PAMI)*, 2007.