Modelling Visual Objects Invariant to Depictive Style

Qi Wu
http://www.cs.bath.ac.uk/~qw219
Peter Hall
http://www.cs.bath.ac.uk/~pmh

Media Technology Research Centre
Department of Computer Science
University of Bath,
Bath, UK

Representing visual objects is an interesting open question of relevance to many important problems in Computer Vision. State of the art allows thousands of visual objects to be learned and recognised, under a wide range of variations. Only a small fraction of the literature addresses the problem of variation in depictive style (photographs, drawings, paintings *etc.*), yet considering photographs and artwork on equal footing is philosophically appealing and of true practical significance. This paper describes a model for visual object classes that is learnable and which is able to classify over a broad range of depictive styles. When compared to a collection of state-of-art classifiers, our results show a significant increase in robustness to variance in depictive style.

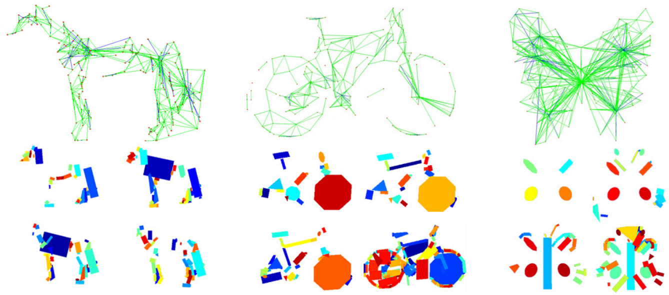

Figure 1: Examples of three graph models generated from 3 categories of objects, which are horses, bicycles, and butterflies. The visualization shows of selected levels below the corresponding model, with the simple shapes fitted. Child-parent arcs are in blue, adjacencies between the nodes in the same level are green.

We argue that models of visual objects should not be premised, even tacitly, on photo-real appearance or indeed on any particular depictive style at all. Rather, visual object models should be based on quasi-invariant properties of the objects in a class. A similar argument is made by those who advocate part-based representations for image. We go further by saying that such models should generalise across depictive styles. This means that if a model is constructed using images in one style, the same object should also be classifiable even when depicted using a different style. In this paper, we investigate a method for modelling visual objects classes in a manner that is invariant to depictive style. The assumption we make is that an object class is characterised by the qualitative shape of object parts and their structural arrangement. Hence we use a graph of nodes and arcs in which qualitative shapes such as triangle, square, and circle to label the nodes. More exactly our model is a hierarchy of levels, yielding a coarse-to-fine representation. Each level contains an undirected graph of nodes and arcs. Nodes between levels are connected via parent-child arcs, which are directed. Child nodes are nested inside their parent.

We learn visual class models from input images, each labelled with the object they contain. There are three major steps. Fig 2 presents a framework of the proposed method.

(i) Build Image Graphs, one for each images: Our modeller uses Berkeley segmentation [1] that automatically yields a hierarchical description of an input image. And then we build a graph from this in which regions label nodes. A novelty here is that we label nodes using qualitative shapes from a collection of shapes. Because we assume that abstracting region shape brings greater robustness to model different depictive styles. These primitive shapes are chosen because they have been shown to explain around 80% of regions in photographs *up to an affine transform*[3].

(ii) Compute an Initial Visual Class Model: The second step is to compute an initial visual class model, by using an approximate median graph generation method [2].

(iii) Refine the Visual Class Model: The initial model contains nodes and arcs that derive from visual clutter in back ground of images in the training set, so we developed a cleaning algorithm to remove such elements.Fig 1 shows some final results.

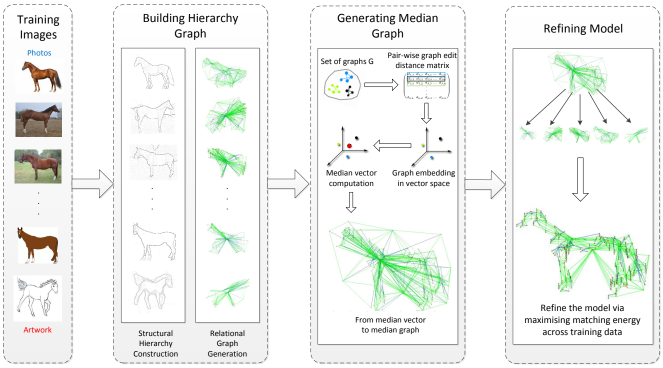
Training Images
Photos
Artwork
Building Hierarchy Graph
Structural Hierarchy Construction
Relational Graph Generation
Generating Median Graph
Set of graphs G
Pair-wise graph edit distance matrix
Median vector computation
Graph embedding in vector space
From median vector to median graph
Refining Model
Refine the model via maximising matching energy across training data

Figure 2: Constructing a class model, from left to right. (a): An input collection (possibly different depictions) used for training. (b): Probability maps for each input image, and graph models for each map. (c): The median graph model for the whole class. (d): The refined median graph as the final class model.

Our *vcm* has the potential to be used in many applications, here we use cross depiction classification. Using our expanded version of CalTech 256, we compare with three state of art classification methods, one BoW method which uses PHOW features, one shape model-based method, another one uses the first few eigenvalues of Laplacian matrix of the object structure as feature vector. In total we test 800 images and Table 1 shows that our method outperforms the shape and structure only method on all cases. We outperform BoW in all cases except case 1i, when photographs are used in both training and testing.

case 1: Training	5p	5a
case 1: Testing	15p	15a
Dense SIFT	70%	59%
Shape Model	25%	33%
Structure Only	16%	19%
Proposed Method	61%	62%

case 2: Training	8p	10p	8a	10a
case2 : Testing	15a	15a	15p	15p
Dense SIFT	43%	47%	49%	51%
Shape Model	33%	35%	34%	34%
Structure Only	19%	23%	22%	25%
Proposed Method	63%	64%	64%	67%

case 3: Training	3a	5a	3p	5p
case 3: Testing	30m	30m	30m	30m
Dense SIFT	46%	50%	50%	54%
Shape Model	27%	30%	24%	27%
Structure Only	13%	16%	14%	16%
Proposed Method	58%	61%	56%	61%

case 4: Training	6m	10m
case 4: Testing	30m	30m
Dense SIFT	60%	61%
Shape Model	32%	34%
Structure Only	21%	24%
Proposed Method	62%	65%

Table 1: Classification accuracy for different cases. From top to bottom, left to right: (a) single domain task, (b) single cross depiction task, and (c) single to mixture depiction task, (d) mixture cross depiction task. The character 'p' is 'photos', 'a' is 'art' and 'm' is 'mixture'. More detailed results for each single experiment can be found in supplementary material.

The ability to generalise to new depictive styles is important, not least because the number of depictive styles is seemingly unbounded. No training procedure can capture them all and so a class model that is able to generalise to unseen depictive styles is of value. Experiments show that our proposal method performs better than the traditional visual appearance based method in cross-depiction problems, in mixed problems, and in art-only problems.

[1] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik. Contour detection and hierarchical image segmentation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(5):898–916, 2011. ISSN 0162-8828.

[2] Miquel Ferrer, Ernest Valveny, Francesc Serratosa, Kaspar Riesen, and Horst Bunke. An approximate algorithm for median graph computation using graph embedding. In *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, pages 1–4. IEEE, 2008.

[3] Qi Wu and Peter Hall. Prime shapes in natural images. In *Proceedings of the British Machine Vision Conference*, pages 45.1–45.12. BMVA Press, 2012. ISBN 1-901725-46-4.