Discovering Multi-Relational Latent Attributes by Visual Similarity Networks

Fatemeh Shokrollahi Yancheshmeh, Joni-Kristian Kamarainen, and Ke Chen

Department of Signal Processing
Tampere University of Technology, 33101 Tampere, Finland
http://vision.cs.tut.fi

Abstract. The key problems in visual object classification are: learning discriminative feature to distinguish between two or more visually similar categories (e.g. dogs and cats), modeling the variation of visual appearance within instances of the same class (e.g. Dalmatian and Chihuahua in the same category of dogs), and tolerate imaging distortion (3D pose). These account to within and between class variance in machine learning terminology, but in recent works these additional pieces of information, *latent dependency*, have been shown to be beneficial for the learning process. Latent attribute space was recently proposed and verified to capture the latent dependent correlation between classes. Attributes can be annotated manually, but more attempting is to extract them in an unsupervised manner. Clustering is one of the popular unsupervised approaches, and the recent literature introduces similarity measures that help to discover visual attributes by clustering. However, the latent attribute structure in real life is multi-relational, e.g. two different sport cars in different poses vs. a sport car and a family car in the same pose - what attribute can dominate similarity? Instead of clustering, a network (graph) containing multiple connections is a natural way to represent such multi-relational attributes between images. In the light of this, we introduce an unsupervised framework for network construction based on pairwise visual similarities and experimentally demonstrate that the constructed network can be used to automatically discover multiple discrete (e.g. sub-classes) and continuous (pose change) latent attributes. Illustrative examples with publicly benchmarking datasets can verify the effectiveness of capturing multi- relation between images in the unsupervised style by our proposed network.

1 Introduction

Active research on visual object class detection and classification during the last ten years has produced novel approaches and many effective methods. At the same time, the main benchmark has switched from the Caltech-4 dataset of 4 categories and 3k images to the ImageNet [3] LSVRC challenge of 200 classes and 450k images (in ILSVRC 2014). The only change is not the increased number of images and classes but also a more realistic problem setting. That is,

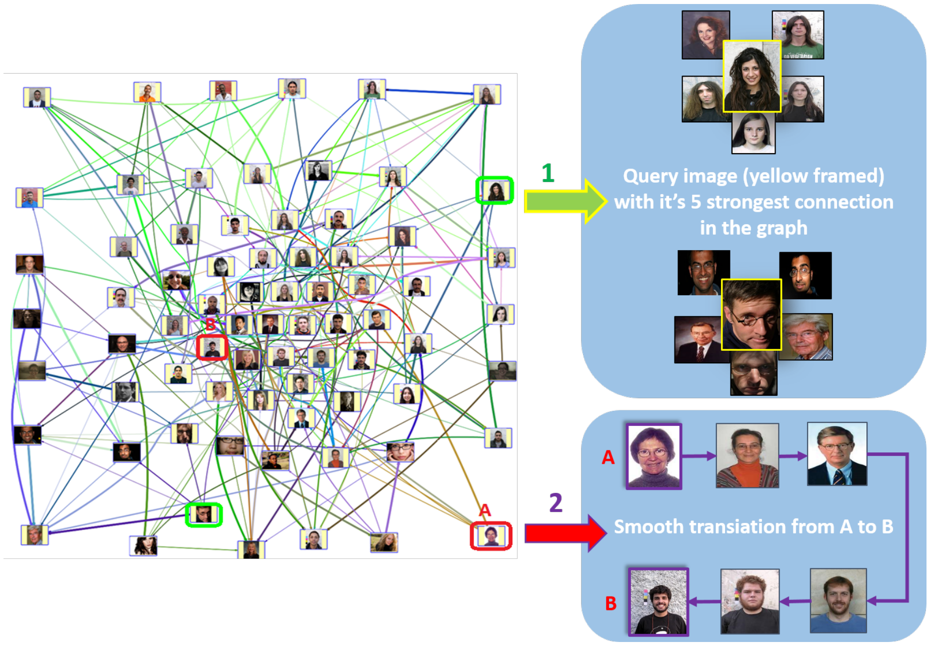
B
A
1
Query image (yellow framed) with it's 5 strongest connection in the graph
2
A
Smooth transiation from A to B
B

Fig. 1. A visual similarity network of ImageNet faces constructed using our procedure. The network can be used to find similar examples (strong links) or gradual change path from one example to another via the shortest path.

instead of a single well-captured object in a fixed pose, ILSVRC contains multiple objects with severe 3D pose changes, occlusions and background clutter. These result great problems to the standard monolithic 2D methods and therefore novel learning paradigms, such as attribute learning [8, 23], have recently gained momentum. Visual class detection can benefit from discovered latent visual attributes by learning multiple "fine-grained" classifiers [4].

Many attribute learning works utilise manually annotated attributes [2, 8], which are not suitable for large-scale problems due to the expensive manpower involved in the annotation. In view of this, unsupervised approaches that can automatically discover latent attributes [1, 4, 10, 30] are considered and adopted. The popular unsupervised tool is clustering, but it omits the fact that often latent attributes are multi-relational and thus breaking them to discrete "modes" is not sensible – what is anyway the more dominant attribute, car pose or model?

In this work, we adopt a network structure that can unsupervisely learn and represent multi-relational attributes simultaneously. For network construction, we propose a pairwise similarity measure and in the experimental part demonstrate that the network can be used to automatically discover multiple discrete (e.g. sub-classes) and continuous (pose change) latent attributes (see Fig. 1). Our network establishes a structure which can be used in visual object categorisation to learn and represent multiple complex attribute-interpreted interconnections and benefits from more and more data. Our main contributions are as follows:

- A novel similarity measure, which combines descriptor based local appearance similarity and part-based constellation similarity into a unique similarity score, is proposed for constructing a visual similarity network based on pairwise matches of images.
- Experimental results where multiple multi-relational latent attributes are discovered using a network (sub-categories, gradual change between examples in a same category and continuous attributes such as 3D pose change).

All source codes and data will be made publicly available[1].

2 Related Work

Attribute learning – Explicit learning of visual attributes was first proposed by Ferrari and Zisserman [8]. Their method learned visual models of various attributes via weakly supervised setting where the training set was produced by pre-defined Web searches. Recently, unsupervised attribute discovery has gained more attention owing to its superiority to saving the involvement of manpower. Methods for completely unsupervised visual object classification (no labels or bounding boxes) have been proposed [10, 11, 27], but due to their large accuracy gap to the state-of-the-art supervised methods [5, 12, 25] they have not received enough attention. Attributes may still be beneficial in certain cases, such as zero-shot learning [14], with only a small number of training images [2], fine-grained classification [9] or utilising scene attributes to improve detection [16, 19].

Visual networks – Methods employing a network (graph) structure to represent visual relationships have recently been proposed [4, 10, 19, 22, 29, 30]. Most of these algorithms aim at finding classes or a specific object automatically [10, 22, 29]. The core element of the methods is to introduce a proper similarity measure and tailoring it for a problem-specific goal. In recent works, Aghazadeh et al. [1] and Dong et al. [4] establish their similarity measures using classification scores of the exemplar SVM [17] which forms own classifier for each sample. Another similarity measure was proposed [30] using feature's tree distances in unsupervised random clustering forests. Learning similarity measures can be time consuming since they may change if new images are added and therefore our measure will be based on pairwise structural similarity combining local part appearance and part configuration.

3 Visual Similarity Networks

Given a set of images containing objects from visual classes and with appearance and pose variation and imaging distortions, we aim to create an image network where link strengths represent the visual similarity between two network nodes (images). The network, or termed graph, consists of nodes (images),

[1] The codes can be downloaded from :https://bitbucket.org/kamarainen/imgalign/code

edges between the nodes and weights representing the visual similarity (distance) between two nodes. The graph is directed if the used similarity measure does not commute, or undirected if it does.

In our proposed network-based framework, the assumption is that such network can be constructed from pairwise similarities by forming a similarity matrix that represents a full-connected network. If the matrix is symmetric, the network is undirected. The core elements in this procedure are *i)* a pairwise image visual similarity measure (Sec. 3.1) and *ii)* a network construction strategy (Sec. 3.2).

3.1 A Pairwise Visual Similarity Measure

Explicit visual measures, such as the simple pixel-wise difference, have the problem that they cannot tolerate well standard imaging distortion and typically behave well only close to a perfect match. Therefore, the recent trend in measuring visual similarity is to use "learning metrics" that establish a computational measure via ad hoc learning [1, 4, 30]. With learning metrics, the problem is that they depend on used training images and a selected objective function, and it is unclear how they generalise beyond images in the training set.

To measure pairwise similarity on object class level, we adopt structural visual similarity based on the part-based models of visual classes that has been particularly successful in object class detection [6, 7, 13]. In the simplest form, we describe $j = 1, 2, \ldots, M$ parts of an image I_i by feature descriptors $F_{i,j}$ (e.g. SIFT). Every descriptor is associated with a spatial location $\boldsymbol{x}_j = (x, y)_j$. The part-based visual similarity of two images I_a and I_b can be defined as

$$s(I_a, I_b) = s\left(\{< F_{a,j}, \boldsymbol{x}_j >_{j=1 \ldots M_a}\}, \{< F_{b,j}, \boldsymbol{x}_j >_{j=1 \ldots M_b}\}\right) \tag{1}$$

The problem is that the two images are related to each other by unknown geometric transformation $\mathbf{T} : \{< F, \boldsymbol{x} >\} \mapsto \{< F', \boldsymbol{x}' >\}$ that aligns the object parts (and affects also to the part descriptors if these are not invariant to the selected transformation type). The similarity measure of two unaligned images must therefore include the transformation term

$$s(I_a, I_b) = \max_{\mathbf{T}} s(I_a, \mathbf{T}(I_b)) \ . \tag{2}$$

While the parts may have false matches due to background clutter, self-similarity or descriptor mismatch, or no matches due to occlusion, practical implementation of the similarity function becomes very complex.

In order to match two images I_a and I_b well, there should be good matches between the descriptors $F_{a,j} \sim F_{b,j'}$ and the matching descriptors $(j \leftrightarrow j')$ should locate spatially close $\boldsymbol{x}_{a,j} \sim \boldsymbol{x}_{b,j'}$ under the transformation $\mathbf{T}$ (e.g., 2D scaling, translation and rotation). To avoid the complex approximation, we construct the similarity matching step-wise: first we construct the part appearance similarity matrix $\mathbf{D}$ and then using the sparse binary $\mathbf{D}$ we sample transformations $\mathbf{T}$ such that the geometric matching of feature locations is maximised. Our visual measure combines part-based appearance similarity and parts' constellation based structural similarity into a single novel measure.

Part-based appearance similarity – As a standard procedure, we compute dense SIFT descriptors using the VLFeat library [28] for every image scaled into the same image resolution (320 × 200, 640 × 400). Scaling makes the method resolution independent such that proportionally same size objects are matched against each other. We can achieve scaling invariance by using multi-resolution pyramid grids scaling invariance, but we did not find it necessary with the standard benchmark datasets (e.g. Caltech-101/256, Pascal VOC, ImageNet). For pairwise similarity of images a and b, the full descriptor distance matrix is computed between all features $\{F_a\}_i$ and $\{F_b\}_j$ forming the descriptor distance matrix $\mathbf{D}_{M_a \times M_b}$. We convert the distance matrix into a sparse binary form by assigning 1 to the five best matches and 0 for the rest. The five best is justified by the class level matching which is much weaker than between two views of a same object. 2-5 best matches were found to improve the matching considerably while beyond 5 improvement saturated quickly.

Structural similarity – The part-based similarity (e.g. summing N best descriptor distances) would somehow resemble the *visual bag of words* approach which has been used in graph construction [29]. However, the problem with that measure is that it does not constrain the accepted matches to be spatially consistent. Instead, we propose a similarity scoring procedure similar to used in specific image matching [15, 21, 26], but in our case for multiple candidates and not restricted by some fixed number of inliers. Due to enormous number of potential matches for exhaustive search, we repeat a random sampling procedure that selects two (minimum for similarity transform) features from a, their best candidates within the five best in $D_{M_a \times M_b}$, transforms all features in a to b and counts how many matches were found within the descriptor ellipses [18]. This procedure is repeated R times (100 found sufficient in our experiments for images of size 320 × 200) and the highest number used as the similarity measure between the images from a to b. It is noted that we do not restrict the transformation although it would improve the results with the standard datasets where objects are typically captured in a few standard poses (e.g. pictures of horizontal and vertical guitars in ImageNet). Moreover, the similarity matrix $S_{N \times N}$ is non-symmetric since matching from a to b can be different from b to a.

Similarity matrix – The output of the structural similarity is the number of parts that match both by their appearance and their configuration between the two images a and b. To establish a similarity matrix, the procedure is executed between all images making our method's computational complexity $O(N^2)$ for N images. At this stage, we make one more trick that prevents classes with plenty of salient parts to dominate by their high similarity scores and transform the actual scores to match ranks, i.e. the highest rank N is assigned to the best matching image and 1 to the worst matching:

$$S'\,(i = 1 \ldots N, j = 1 \ldots N) = row\_wise\_rank(S(i,j), S_{N \times N}) \ . \tag{3}$$

3.2 Network Construction

The network construction is straightforward if we change the structural visual similarity scores given as rank numbers to distances by

$$\hat{S}\left(i = 1 \ldots N, j = 1 \ldots N\right) = \frac{N}{S'(i,j)} \quad . \tag{4}$$

Moreover, to make undirected graph algorithms available we convert the distance matrix symmetric by

$$\hat{S}\left(i, j\right) = min\left(\hat{S}(i,j), \hat{S}(j,i)\right) \quad . \tag{5}$$

The visual similarity network of images is constructed as a graph $G = (V, E, W)$, where $V = \{V_1, ..., V_N\}$ denotes the set of vertices (nodes), $E_{N\times N}$ denotes the set of edges or links, and $W_{N\times N}$ is the set of edge weights. We assign each image $I_q \in \{I_1, ..., I_N\}$ to the corresponding vertice V_q and set the computed $\hat{S}_{N\times N}$ as the edge weights such that $W\left(i, j\right) = S\left(i, j\right)$. Therefore the edge weights would reflect the visual similarity such that low weights are assigned between similar images. It should be noted that we set the diagonal of W to 0 to remove self-references. The full network consists of $N \times (N-1)$ links. An example of the constructed network is illustrated in Fig. 1.

4 Examples and Experiments

In the experiments, we used images from various synsets of the ImageNet Large Scale Visual Recognition Challenge 2014 (ILSVRC[2]), EPFL GIMS08 [20], and 10 categories of 3D object [24]. The only input for the network construction was images and therefore only their visual content affects the results.

4.1 Discovering Classes and Sub-classes

At first, it is attempting to use the network structure to automatically discover the most apparent discrete attributes such as visual classes (synsets) [10, 29] or their sub-classes [4]. Whether our network structure can represent inter-class relationships we selected three distinct ImageNet classes: cars, motorbikes and airplanes. Using the Prim's algorithm, we constructed a minimum spanning tree (MST) shown in Fig. 2. From the tree and its closeups, it is evident that the classes have distinct branches that also represent sub-class information (scooters form their own branch in the motorbikes).

[2] http://image-net.org/challenges/LSVRC/2014/

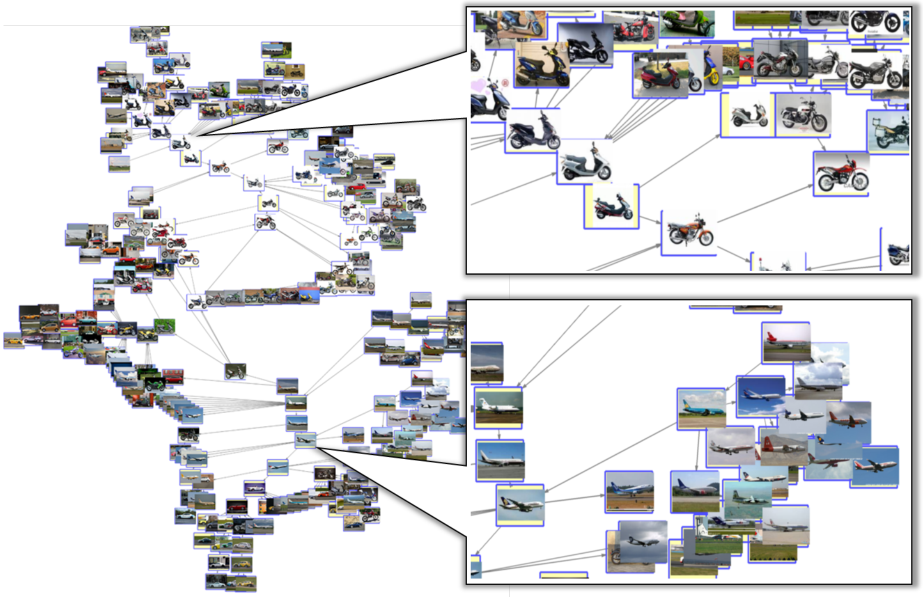

Fig. 2. A minimum spanning tree of a network constructed using our pairwise similarity measure for ImageNet cars, motorbikes and airplanes with closeups demonstrating the class branches.

4.2 From Classes to Objects in Similar Pose

The interpretation that our network can be used to unsupervised discover classes and sub-classes as shown in Fig. 2 is not accurate. It can be used for dedicatedly selected examples, but if we add more ImageNet images that span almost full 3D poses, pairwise similarities do not anymore code class level similarity but combined pose and class similarity, and which dominates depends of a specific image pair. This finding is demonstrated in Fig. 3 where we generated the graphs for a large number of images from multiple ImageNet synsets (guitar, airplane), 3D object dataset (head, bicycle, shoe, stapler). In these examples, the pose dominates the pairwise similarity and not the specific object example. The finding conflicts with the works that try to discover classes and sub-classes in a graph structure [4, 10, 29], but supports works learning multiple classifiers for different poses [1]. Overall, our examples illustrate the fact that pairwise visual similarity graph represents multiple multi-relational latent attributes at the same time.

4.3 Traveling Graph in Latent Continuous Attribute Space

The previous examples illustrated how a single graph can represent multiple “competing” attributes simultaneously and therefore methods based on strict boundaries, such as clustering, are deemed to fail. In a single network, there

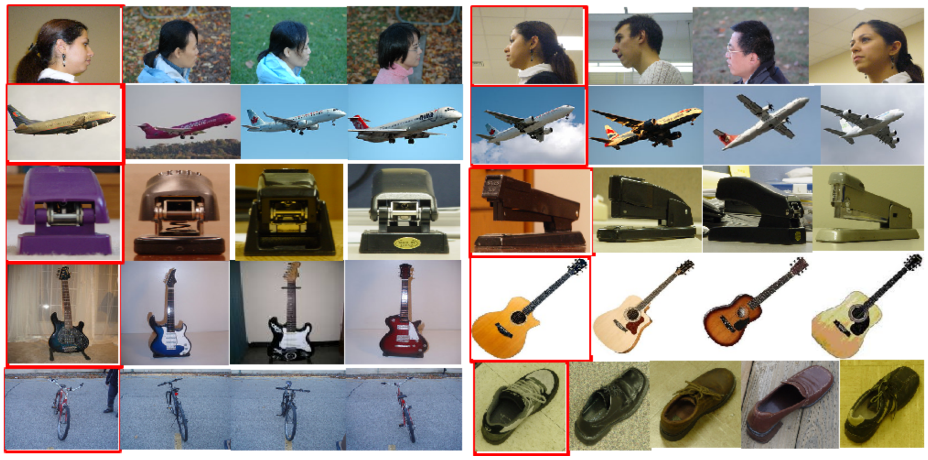

Fig. 3. Single network node images (denoted by the red rectangular) and for each 3-4 neighbour images with the strongest connection links (similarity scores). The results are more dominated by a specific pose rather a specific object.

can be several almost equally good paths between two nodes and the transition “smoothness” depends on how many images there are in a network. A single shortest path can be selected by, for example, MST algorithm. In Fig. 4 we demonstrate continuous attributes by selecting “source” and “sink” images, and traveling the graph from the source to the sink via the MST’s path. The images in nodes within the path represent the active attributes.

5 Conclusions

In this work, we proposed to use pairwise “structural visual similarity” between images to construct a network of images. The structural visual similarity is based on a simple principle that local regions of the two images match and are in similar spatial configuration (constellation). By using the similarity measure in network construction we noticed that the network can represent multi-relational attributes of discrete type (e.g. class/sub-class) and continuous type (e.g. 3D pose) (Fig. 5). However, with a large number of images, such as in the largest ImageNet synsets, certain attributes, such as the pose, may dominate other attributes and therefore any unsupervised graph-based attribute search using tight boundaries, such as clustering or minimum spanning tree (Fig. 6), is deemed to fail. Therefore, we believe that the network structure representing images as nodes and their similarities as connecting edges is a natural presentation of visual data. In this sense, even the attribute classifiers may share some images. Our future work will exploit the network structure further and the immediate interests are two-fold: 1) how to boost our network algorithm of the computational complexity $O(N^2)$ (similarity matrix construction) more efficient to cope with

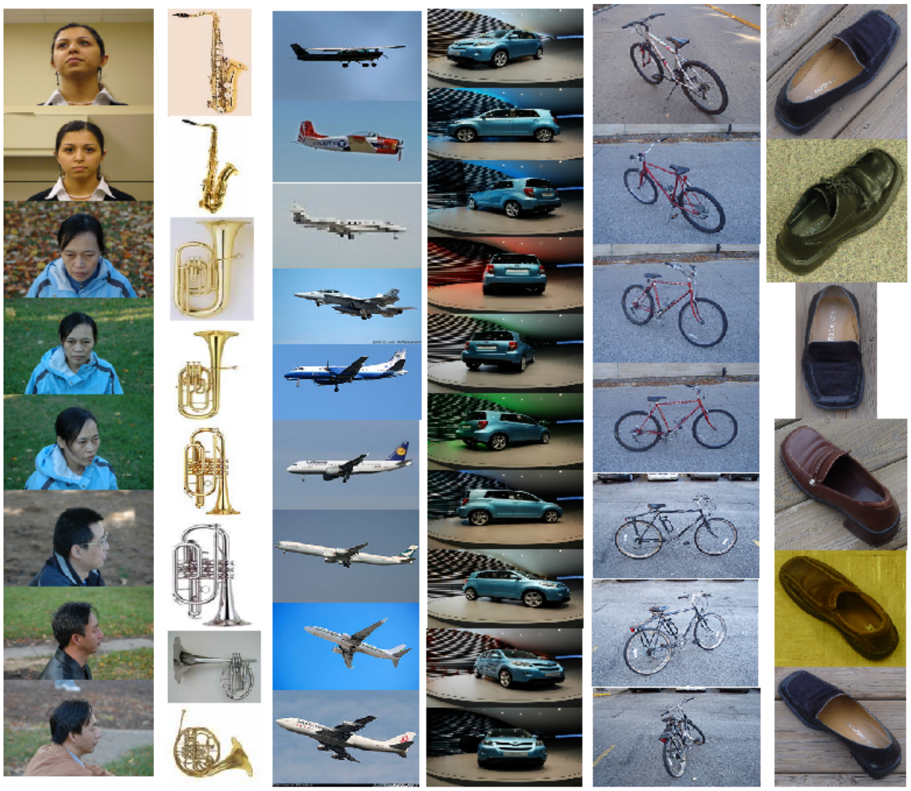

Fig. 4. “Source” images selected from 3D object dataset (head, bicycle, shoe), ImageNet (sax, airplane), EPFL GIMS08 (car) on the top and “sink” images at the bottom. The other images represent nodes, “smooth transition”, between the source and sink within the minimum spanning tree. Note that, the path encodes gradual change in pose (face, car, bike and shoe) or appearance mixed with pose (music instrument, airplane). The shortest path depends on all images and there can be multiple almost equally good paths.

millions of images; and 2) how to unsupervised discover all active multi-relational attributes.

References

1. Aghazadeh, O., Azizpour, H., Sullivan, J., and Carlsson, S. Mixture component identification and learning for visual recognition. In *ECCV* (2012).
2. Akata, Z., Perronnin, F., Harchaoui, Z., and Schmid, C. Label-embedding for attribute-based classification. In *CVPR* (2013).
3. Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. ImageNet: A large-scale hierarchical image database. In *CVPR* (2009).

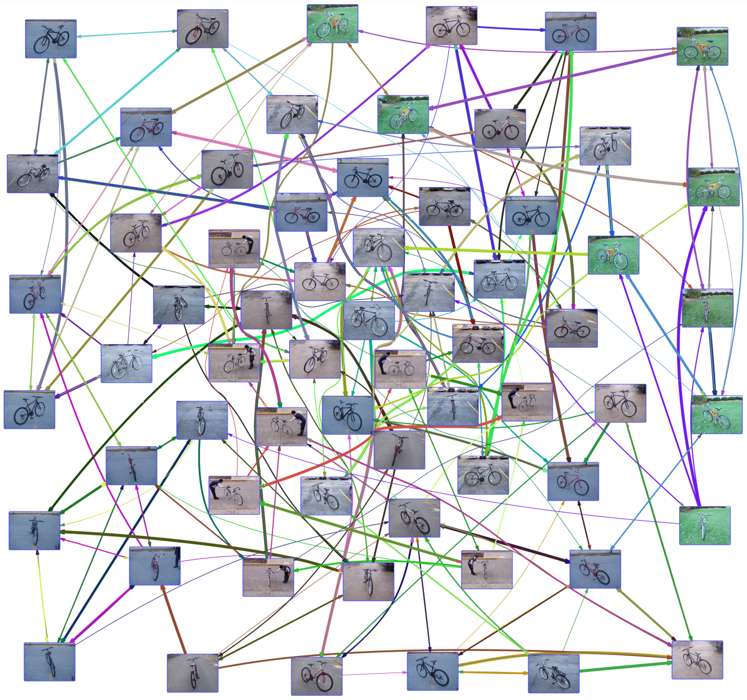

Fig. 5. A pairwise visual similarity network of bike images.

4. Dong, J., Xia, W., Chen, Q., Feng, J., Huang, Z., and Yan, S. Subcategory-aware object classification. In *CVPR* (2013).
5. Felzenszwalb, P. F., Girshick, R. B., McAllester, D., and Ramanan, D. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence 32*, 9 (2010), 1627–1645.
6. Felzenszwalb, P. F., Girshick, R. B., McAllester, D., and Ramanan, D. Object detection with discriminatively trained part based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence 32*, 9 (2010), 1627–1645.
7. Felzenszwalb, P. F., and Huttenlocher, D. P. Pictorial structures for object recognition. *Int J Comput Vis 61*, 1 (2005), 55–79.
8. Ferrari, V., and Zisserman, A. Learning visual attributes. In *Advances in Neural Information Processing Systems (NIPS)* (2007).
9. Gavves, E., Fernando, B., Snoek, C. G. M., Smeulders, A. W. M., and Tuytelaars, T. Fine-grained categorization by alignments. In *ICCV* (2013).
10. Kim, G., Faloutsos, C., and Hebert, M. Unsupervised modeling of object categories using link analysis techniques. In *CVPR* (2008).
11. Kinnunen, T., Kamarainen, J.-K., Lensu, L., and Kälviäinen, H. Unsupervised object discovery via self-organisation. *Pattern Recognition Letters 33*, 16 (2012), 2102–2112.
12. Krizhevsky, A., Sutskever, I., and Hinton, G. ImageNet classification with deep convolutional neural networks. In *NIPS* (2012).

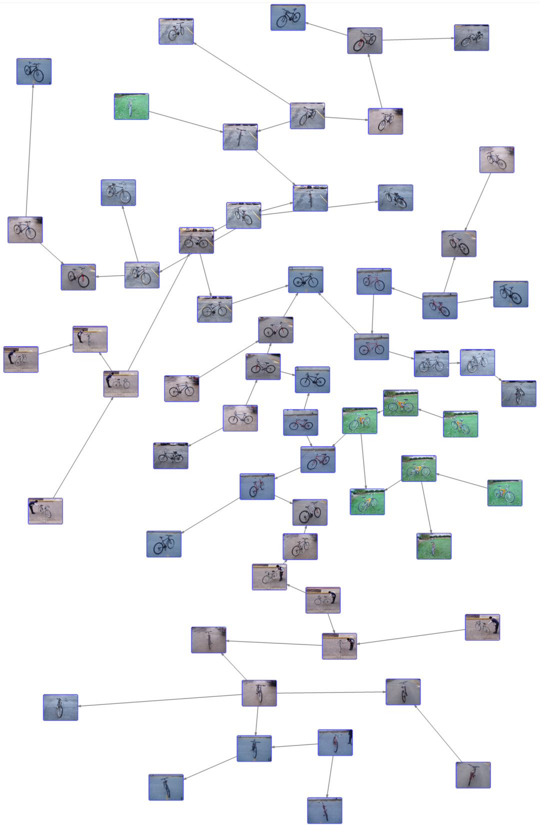

Fig. 6. A minimum spanning tree of the network in Fig. 5.

13. Kumar, M., Zisserman, A., and Torr, P. Efficient discriminative learning of parts-based models. In *ICCV* (2009).
14. Lampert, C. H., Nickisch, H., and Harmeling, S. Attribute-based classification for zero-shot visual object categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence 36*, 3 (2014), 453–465.
15. Lankinen, J., and Kamarainen, J.-K. Local feature based unsupervised alignment of object class images. In *BMVC* (2011).
16. Malisiewicz, T., and Efors, A. Beyond categories: The visual memex model for reasoning about object relationships. In *NIPS* (2009).
17. Malisiewicz, T., Gupta, A., and Efors, A. Ensemble of exemplar-SVMs for object detection and beyond. In *ICCV* (2011).
18. Mikolajczyk, K., and Schmid, C. A performance evaluation of local descriptors. *IEEE PAMI 27*, 10 (2005).
19. Myeong, H., Chang, J. Y., and Lee, K. M. Learning object relationships via graph-based context model. In *CVPR* (2012).
20. Ozuysal, M., Lepetit, V., and P.Fua. Pose estimation for category specific multiview object localization. In *CVPR* (2009).

21. Philbin, J., Chum, O., Isard, M., Sivic, J., and Zisserman, A. Object retrieval with large vocabularies and fast spatial matching. In *CVPR* (2007).
22. Philbin, J., Sivic, J., and Zisserman, A. Geometric latent dirichlet allocation on a matching graph for large-scale image datasets. *Int J Comput Vis 95*, 2 (2011), 138–153.
23. Russakovsky, O., and Fei-Fei, L. Attribute learning in large-scale datasets. In *ECCV), Workshop* (2010).
24. Savarese, S., and Li, F.-F. 3d generic object categorization, localization and pose estimation. In *ICCV* (2007), pp. 1–8.
25. Simonyan, K., Vedaldi, A., and Zisserman, A. Deep Fisher networks for large-scale image classification. In *NIPS* (2013).
26. Tuytelaars, T., and Gool, L. V. Wide baseline stereo matching based on local, affinely invariant regions. In *BMVC* (2000).
27. Tuytelaars, T., Lampert, C., Blaschko, M., and Buntine, W. Unsupervised object discovery: A comparison. *Int J Comput Vis 88*, 2 (2010).
28. Vedaldi, A., and Fulkerson, B. VLFeat: An open and portable library of computer vision algorithms. http://www.vlfeat.org/, (2008).
29. Xia, S., and Hancock, E. R. Incrementally discovering object classes using similarity propagation and graph clustering. In *ACCV* (2009).
30. Zhu, X., Loy, C., and Gong, S. Constructing robust affinity graph for spectral clustering. In *CVPR* (2014).