Learning Clustered Sub-spaces for Sketch-based Image Retrieval

Koustav Ghosal Ameya Prabhu Riddhiman Dasgupta Anoop M Namboodiri
koustav.ghosal* ameya.prabhu* riddhiman.dasgupta* anoop†
Centre for Visual Information Technology, IIIT-Hyderabad, India

Abstract

Most of the traditional sketch-based image retrieval systems compare sketches and images using morphological features. Since these features belong to two different modalities, they are compared either by reducing the image to a sparse sketch like form or by transforming the sketches to a denser image like representation. However, this cross-modal transformation leads to information loss or adds undesirable noise to the system. We propose a method, in which, instead of comparing the two modalities directly, a cross-modal correspondence is established between the images and sketches. Using an extended version of Canonical Correlation Analysis (CCA), the samples are projected onto a lower dimensional subspace, where the images and sketches of the same class are maximally correlated. We test the efficiency of our method on images from Caltech, PASCAL and sketches from TU-BERLIN dataset. Our results show significant improvement in retrieval performance with the cross-modal correspondence.

1. Introduction

Facebook, Instagram and Flickr recently announced in press reports that users upload and share 350M, 40M, 1.83M images daily to their servers, respectively. Organizing and retrieving this humongous amount of image data is a challenging task but unlike document retrieval, content based image retrieval (CBIR) still dwells at an infantile stage in terms of both usability and performance, which makes it an active and interesting area of research.

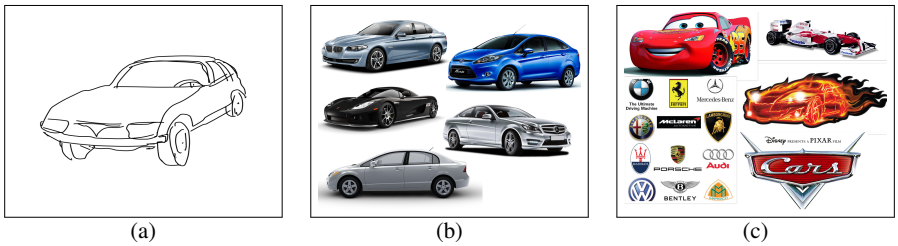
Existing CBIR systems can be broadly categorized into three divisions: text-based, example-based and sketch-based. In text-based query image search, similar keywords from meta data space (tags and annotations) associated with images are searched. But the meta data is generally not reliable as it may not represent the actual content in the image or it could be misleading. Apart from that, pertaining to perceptual variability, different users may use entirely different queries to search for the same images which require NLP techniques for correct interpretation. On the other hand, in case of the *query by example* paradigm, example images are not always available at hand, in fact their absence being the reason for a search. Sketch-based interfaces can be effectively used in such scenarios.

For example, as shown in Figure 1, if a user needs to search for the image of a car from the front-view , a sketch as in Figure 1(a) could be a very convenient way to frame the query. Unlike text, it is more intuitive to the user and contains information regarding the shape, position and orientation of the object, concisely. With all the information embedded in the query itself, images like in Figure 1(b) are more likely to appear as results. On the other hand, "*car*" as a text-query might retrieve random diverse images of cars and their associated entities, as shown in Figure 1(c).

A problem with Sketch Based Image Retrieval (SBIR) is that existing approaches rely on edge and shape based similarity between sketches and images. But this fundamental assumption about the similarity between these two modalities is often violated since most humans are not faithful artists. Instead, people use shared, iconic representations of objects (e.g., stick figures for humans) or they make dramatic simplifications or exaggerations (e.g. pronounced ears on rabbits). According to Li *et al*. [12], a simple sketch is a high level sparse representation of the object/scene being searched for. Yong *et al*. [20] found that because of this sparsity, when a sketch is presented as a query to Clarifi [21], cartoon images, which resemble the sketches markedly, are retrieved. So, instead of a direct comparison, in this work we try to learn a cross-modal correspondence between the two modalities. To retrieve an image based on a sketch, our algorithm tries to understand the sketch and the image independently and then compare them using the learned correspondence.

Our contribution in this paper lies in modelling the correspondence between the images and sketches belonging to the same category using a modified version of Canonical Correlation Analysis (CCA). CCA operates on two vector spaces and maps both of them to a lower dimensional subspace such that the correlation between them is maximized.

*@research.iiit.ac.in
†@iiit.ac.in



(a) (b) (c)

Figure 1. (a) A sketch of a car from a front-view (b) Search results which are more accurate to the query in terms of Orientation, Shape and View. (c)Image Search results based on a text query *"Cars"*. Note the variety of related entities to cars that appear as results.

We use Cluster-CCA, which is a modified version of the standard CCA, to create a class wise correspondence between the two modalities instead of a point to point correspondence as in standard CCA.

Our work is inspired from and spans across multiple domains. It is closely associated with object and sketch classification, both of which are two classical problems in Computer Vision and one finds a plethora of sophisticated techniques for each of these tasks. We pose this problem as a cross-domain retrieval task which is a comparatively lesser explored domain.

Traditional techniques like SIFT [14] and HOG [2] and Fisher vectors [15] perform well for image classification tasks. Convolutional Neural Networks also have been around for a while, being first introduced by Lecun *et al*. [11] in 1989, but have recently become popular with the success achieved by Krizhevsky *et al*. [10]

On the other hand, in the last two decades, sketch recognition has been mainly limited to understanding gestures, mathematical symbols, alphabets and digits [18, 7] . A more generic sketch recognition framework was proposed by Eitz *et al*. [3], where they extracted SIFT-like features from sketches. Cao *et al*. [1] used a symmetry aware flip invariant descriptor. Li *et al*. [13, 12] suggested similar solutions using star-graphs and multi-kernel feature learning. Rosalia *et al*. [17] encoded sketches as Fisher vectors which performed well. Recently, Yang *et al*. [20] designed a Deep Neural Network architecture on sketches.

Cross-modal retrieval has been an active area for research for quite sometime. CCA was introduced by Hotelling *et al*. [9] to find relation between two sets of variates. Rasiwasia *et al*. [16], modified the standard version of CCA which finds point-to-point correspondence across two modalities, and proposed Cluster-CCA, which finds cluster to cluster correspondence.

2. Proposed Approch

In this section, we formulate our problem as a cross-modal retrieval task . In Section 2.1, we state the problem formally. In Section 2.2, we briefly explain CCA and it's modified version Cluster-CCA.

2.1. Canonical Correlation Analysis (CCA)

In cross-modal retrieval systems, the query space and the search space are disjoint. In our problem, given a set of images, $I = \{I_1^1 \ldots, I_{n_1}^1, I_1^2, \ldots, I_{n_2}^2 \ldots\ldots\ldots I_1^C \ldots, I_{n_C}^C\}$, where I_q^p is q^{th} sample belonging to category p, where there are C categories and each category contains $n_1, n_2, \ldots, n_C$ samples, respectively. Similarly, we have a set of hand-drawn sketches, having the identical number of object categories, $S = \{S_1^1, \ldots, S_{m_1}^1, S_1^2, \ldots, S_{m_2}^2 \ldots\ldots\ldots S_1^C \ldots, S_{m_C}^C\}$. We would like to find a correspondence between the two sets I and S, and project each of them into a different subspace, such that, they are mapped closely. To achieve this we choose CCA [8], which, given two sets A_x and A_y, tries to find two projection matrices $W_x \in \mathbb{R}_x$ and $W_y \in \mathbb{R}_y$ such that the correlation between $P_x = \langle W_x, A_x \rangle$ and $P_y = \langle W_y, A_y \rangle$ is maximized. Mathematically,

$$\begin{aligned} \rho &= \max_{W_x, W_y} corr(P_x, P_y) \\ &= \max_{W_x, W_y} \frac{\langle P_x, P_y \rangle}{\|P_x\| \|P_y\|} \end{aligned} \tag{1}$$

where ρ is the maximum canonical correlation coefficient.

However, this standard form of CCA finds a point to point correspondence between two sets agnostic to class differences and hence not applicable in our case. Instead, we use a modified version, Cluster-CCA [16], which establishes a one to one correspondence between all pairs of data points in a given class. We explain it in detail in Section 2.2.

2.2. Cluster-CCA

Rasiwasia *et al*. [16] introduced and used Cluster-CCA for cross-modal retrieval tasks with image and text as two modalities. As derived in [8], Equation 1, reduces to the following form

$$\rho = \max_{W_x, W_y} \frac{W_x' \text{Cov}_{xy} W_y}{\sqrt{W_x' \text{Cov}_{xx} W_x} \sqrt{W_y' \text{Cov}_{yy} W_y}} \tag{2}$$

and the covariance matrix of (A_x, A_y) given by:

$$Cov = \mathbb{E}\left[\begin{pmatrix} A_x \\ A_y \end{pmatrix} \begin{pmatrix} A_x \\ A_y \end{pmatrix}' \right] = \begin{bmatrix} Cov_{xx} & Cov_{xy} \\ Cov_{yx} & Cov_{yy} \end{bmatrix} \tag{3}$$

where Cov_{xx} and Cov_{yy} are intra-set covariance matrices and Cov_{xy} is the inter-set covariance matrix. But as we previously mentioned in Section 2.1, sets I and S do not have a direct correspondence to each other. Instead we would require a one-to-one correspondence between all pairs of data points in a given class across the two sets I and S. Thus for categorical data, Equation 2 can be modified to the following form,

$$\rho = \max_{W_I, W_S} \frac{W_I' \Sigma_{IS} W_S}{\sqrt{W_I' \Sigma_{II} W_I} \sqrt{W_S' \Sigma_{SS} W_S}} \tag{4}$$

where the new covariance matrices are defined as follows,

$$\Sigma_{IS} = \frac{1}{M} \sum_{c=1}^{C} \sum_{j=1}^{|I^c|} \sum_{k=1}^{|S^c|} I_j^c S_k^{c'} \tag{5}$$

$$\Sigma_{II} = \frac{1}{M} \sum_{c=1}^{C} \sum_{j=1}^{|I^c|} |S^c| I_j^c I_j^{c'} \tag{6}$$

$$\Sigma_{SS} = \frac{1}{M} \sum_{c=1}^{C} \sum_{k=1}^{|S^c|} |I^c| S_k^c S_k^{c'} \tag{7}$$

where $M = \sum_{c=1}^{C} |I^c||S^c|$, is the total number of pairwise correspondences across C classes. Hereafter, the optimization problem in Equation 4 can be formulated and solved as an eigen value problem as in [8].

To summarize, in this section we explained, how a modified version of the standard CCA can be used to create a one-to-one correspondence between samples belonging to the same category but to two different modalities, image and sketch. We projected each modality, having different dimensions, into two lower dimensional subspaces, such that they are maximally correlated.
Please note that this method is different from other state-of-the-art dimensionality reduction techniques like PCA and LDA. Apart from finding basis vectors along the most variant directions, it operates jointly on both of them. It enhances the association between the sets by projecting them into the new sub spaces.

3. Experiments

In this section, we quantitatively evaluate the performance of our proposed approach. We use three datasets, PASCAL VOC 2007 [4] and CALTECH-256 [5] for images and TU-BERLIN [3] dataset for sketches. We divide the TU-BERLIN dataset into training and testing sets and use the training set to create the correspondence with images. As already discussed, CCA projects the sketches and images to two new subspaces, having same dimensions. Once we get W_I and W_S as explained in the previous section, we can project any query from the test set of TU-BERLIN dataset to the new subspace P_S and retrieve k-nearest neighbours from P_I, as illustrated in Figure 2. We use PR curves and MAP values as quantitative measures.

3.1. Datasets

TU-BERLIN is a well known benchmark dataset for evaluating sketch recognition systems with 250 object categories, each containing 80 sketches. This dataset was annotated by humans with an accuracy of 73%. The best recognition accuracies reported till date is 72.2% by Yang *et al.* [20] and 68% by Rosalia *et al.* [17]. **PASCAL VOC 2007** dataset contains 5011 training images and 4952 test images divided into 20 classes with some images containing multiple labels and serving as text annotations. In our experiments, we have used the entire dataset except class *sofa*, for which there was no corresponding sketch category in TU-BERLIN dataset. **CALTECH 256** dataset consists of 256 classes containing $30,607$ images. However, some categories in this dataset did not belong to the TU-BERLIN dataset and vice versa. Hence, we selected a subset of this dataset, which contained 105 classes, containing $14,231$ images.

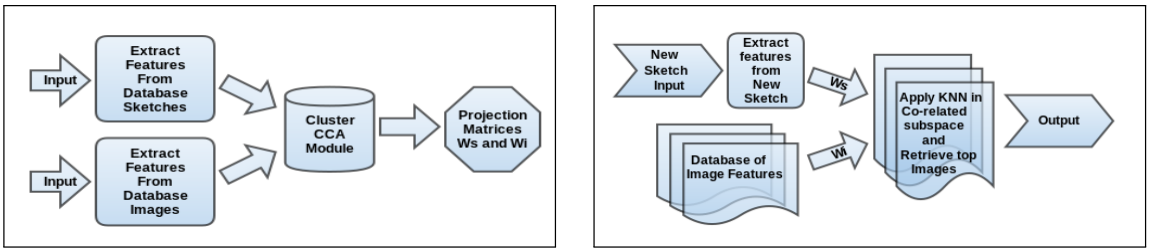
3.2. Features

Given an image I and a sketch S, it is imperative to obtain suitable features which can be used downstream for Cluster-CCA. The rationale behind choosing the features was the assumption that the features which performed well in classification tasks could also perform well in our case. Hence we tried some state of the art features which perform well in recognition and classification. We experiment with local SIFT features, global HOG features, as well as Fisher vectors and features obtained from convolutional networks. We list the set of features used in our experiments in Table 1.
The features are available for download and can be found online on our website*

*http://cvit.iiit.ac.in/projects/sketchbasedretrieval/

Table 1 : Summary of Features

Feature	Dimension	Source
CALTECH - SIFT	1000	Vl-Feat. [19]
CALTECH - HOG	20000	Vl-Feat [19]
CALTECH - CNN	4096	Krizhevsky *et al.* [10]
PASCAL - SIFT	1000	Guillaumin *et al.* in [6]
PASCAL - HOG	20000	Vl-Feat [19]
PASCAL - CNN	4096	Krizhevsky *et al.* [10]
TU-BERLIN - SIFT-Like	501	Eitz *et al.* in [3]
TU-BERLIN - HOG	20000	Vl-Feat [19]
TU-BERLIN - Fisher	250000	Rosalia *et al.* [17]
TU-BERLIN - CNN	4096	Yang *et al.* [20]



(a) (b)

Figure 2. Proposed pipeline : It involves two stages. (a) In the training stage inputs from two modalities are provided to the system. Features are extracted from both the sketches and images and passed to the Cluster-CCA module which projects the inputs onto a lower subspace in such a way that they are maximally correlated. It returns the projection matrices W_S and W_I. (b) In the testing phase, the projection matrices W_S and W_I, transform a new input sketch and the database of images onto the lower dimensional maximally correlated subspace. Finally, a K-NN search is performed and the top-k results are retrieved.

Table 2: Mean Average Precision (MAP) for Image-Sketch feature combinations

Dataset	SIFT-SIFT	SIFT-HOG	SIFT-Fisher	HOG-SIFT	HOG-HOG	HOG-Fisher	CNN-CNN
Caltech	0.06	0.03	**0.20**	0.14	0.02	0.01	**0.20**
Pascal	0.13	0.12	0.05	**0.18**	0.09	0.06	0.06

3.3. Results

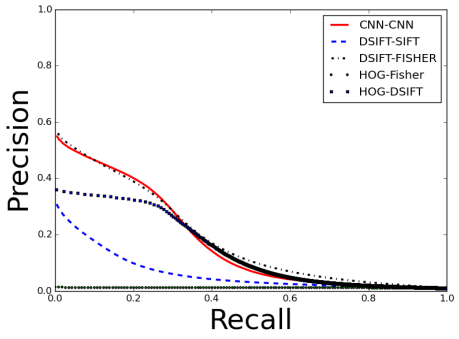
Mean Average Precision (MAP): Table 2 shows the MAP values for all the feature combinations. It can be seen that in case of Caltech dataset, the SIFT features give best results. On the other hand HOG features perform better with PASCAL. Such results can be attributed to the fact that the images in Caltech are of single objects. Dense SIFT features are known to be very good descriptors for single object classification. However, in case of PASCAL, the images are much more complicated and consist of multiple labels. The images are of scenes rather than of single objects. HOG descriptors capture the global information better than the other features. Hence they perform better on the PASCAL dataset. However, the performance of the sketch-features was not very consistent across these two datasets, but we believe more sophisticated feature learning techniques could alleviate this problem.

Table 3: Performance improvement in mAP values

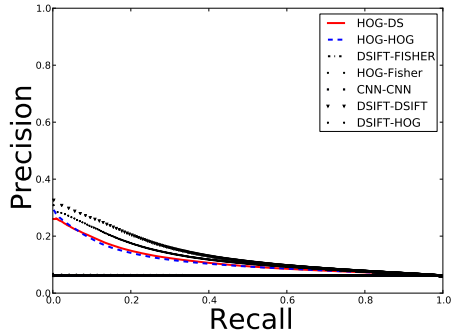
Dataset	Features	Before CCA	After CCA
Caltech	SIFT-Fisher	0.01	**0.20**
Caltech	CNN-CNN	0.01	**0.20**
Pascal	HOG-SIFT	0.01	**0.18**
Pascal	SIFT-SIFT	0.06	**0.13**

In order to validate the impact of Cluster-CCA , we compared the performance of the best four feature combinations, with and without doing Cluster-CCA. The effectiveness of our method can be observed from Table 3 where the performance of all the feature combinations significantly improve.

Precision Recall: The PR curves in Figure 3 are suggestive of the greater complexity of the PASCAL dataset in comparison with Caltech. The poor PR curves indicate that the correspondence doesn't work well with complex images. In Figure 4, we provide results from two example queries for airplane and backpack.



(a) Caltech



(b)Pascal VOC 2007

Figure 3. Precison Recall Curves

(a) Airplane

(b) Backpack

Figure 4. (a) Success : We observe that airplanes of various shapes and orientations are retrieved which shows that our model learns about objects instead of doing a simple shape based comparison. Interestingly, the last image, which is of a camel, resembles an airplane because of the background. (b) Failure : We observe that it was able to retrieve two backpacks and other random objects. However, a closer look reveals structural similarity between the results, and explains the cause of the failure.

4. Conclusion

In this paper we have proposed a system which performs cross-modal image retrieval, where the query is given in the form of a sketch. We try different state-of-the-art feature combinations for sketches and images and compare the results in an exhaustive manner. Our method learns a projection from a higher dimensional subspace to a lower one using a modified version of Canonical Correlation Analysis. We show that the mAP values increase significantly after the features are correlated using CCA.

Our approach is limited by the fact that it is trained on single objects. In real world scenarios, we look for a scene or a collection of objects. An efficient SBIR system should be able to capture the semantics of a sketch and encode the same in the query. Moreover in our experiments we found that Cluster-CCA cannot be generalized to unknown objects.

However, to the best of our knowledge, our proposed system is the only one till date which deals with sketches coming from a wide range of classes. Most of the existing SBIR systems, uses edge and color based features from sketches and then match them directly with the features extracted from images. In our approach, we have used a very simple query format, where each sketch is a single channel sparse image. Then, instead of a direct comparison, we learn lower dimensional subspace where associated points are much closer to each other than in the original space.

This is an interesting problem and there are a lot of areas which can be explored in future. One such immediate direction might be extending this idea to complex scenes. Another interesting area might be finding representations which are generic enough to retrieve unknown classes.

References

[1] X. Cao, H. Zhang, S. Liu, X. Guo, and L. Lin. Sym-fish: A symmetry-aware flip invariant sketch histogram shape descriptor. In *ICCV*, 2013. 2

[2] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005. 2

[3] M. Eitz, J. Hays, and M. Alexa. How do humans sketch objects? *ACM Trans. Graph.*, 2012. 2, 3

[4] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. 3

[5] G. Griffin, A. Holub, and P. Perona. Caltech-256 object category dataset. 2007. 3

[6] M. Guillaumin, J. Verbeek, and C. Schmid. Multimodal semi-supervised learning for image classification. In *CVPR*, 2010. 3

[7] T. Hammond and R. Davis. Ladder, a sketching language for user interface developers. *Computers & Graphics*, 2005. 2

[8] D. Hardoon, S. Szedmak, and J. Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural computation*, 2004. 2, 3

[9] H. Hotelling. Relations between two sets of variates. *Biometrika*, 1936. 2

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012. 2, 3

[11] B. B. Le Cun, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Handwritten digit recognition with a back-propagation network. In *NIPS*, 1990. 2

[12] Y. Li, T. M. Hospedales, Y.-Z. Song, and S. Gong. Free-hand sketch recognition by multi-kernel feature learning. *CVIU*, 2015. 1, 2

[13] Y. Li, Y.-Z. Song, and S. Gong. Sketch recognition by ensemble matching of structured features. In *BMVC*, 2013. 2

[14] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 2004. 2

[15] F. Perronnin, J. Sánchez, and T. Mensink. Improving the fisher kernel for large-scale image classification. In *ECCV*. 2010. 2

[16] N. Rasiwasia, D. Mahajan, V. Mahadevan, and G. Aggarwal. Cluster canonical correlation analysis. In *AI Statistics*, 2014. 2

[17] R. G. Schneider and T. Tuytelaars. Sketch classification and classification-driven analysis using fisher vectors. *TOG*, 2014. 2, 3

[18] T. M. Sezgin, T. Stahovich, and R. Davis. Sketch based interfaces: early processing for sketch understanding. In *ACM SIGGRAPH*, 2006. 2

[19] A. Vedaldi and B. Fulkerson. VLFeat: An open and portable library of computer vision algorithms, 2008. 3

[20] Y. Yang and T. M. Hospedales. Deep neural networks for sketch recognition. *arXiv preprint arXiv:1501.07873*, 2015. 1, 2, 3

[21] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *Computer Vision–ECCV 2014*. 2014. 1