A Sketch-based Approach for Multimedia Retrieval

MS (by research) Thesis,

August, 2012 - November, 2016

Koustav Ghosal
Supervisor : Dr. Anoop Namboodiri

Centre for Visual Information Technology
IIIT Hyderabad

November 29, 2016

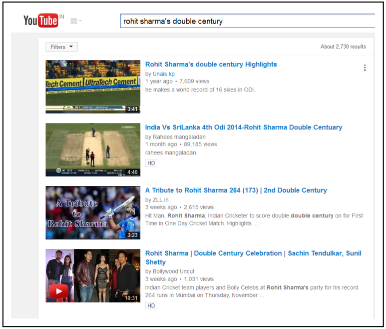
Outline

Outline

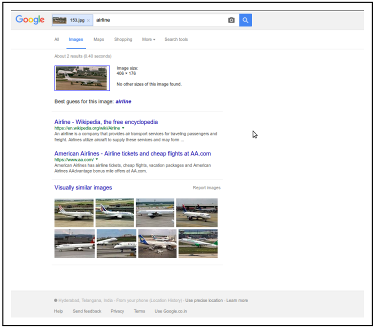
1. Motivation
2. Video Retrieval
3. Image Retrieval
4. Zero-Shot Learning

Content Based Multimedia Retrieval

Textual and Example Based Queries



(a)



(b)

Figure: (a) Query by Text (b) Query by Example

Content Based Multimedia Retrieval

Limitations of Text based Query

Figure: Image Search

Metadata may not represent the original content.

"All those red coloured vehicles which came south and turned left"

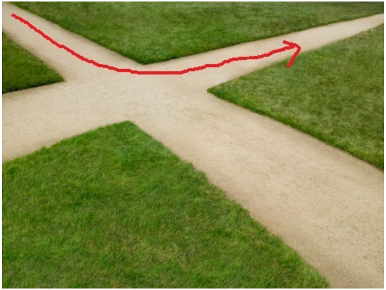


Figure: Tracking

A more complicated event means a more complicated query.

Examples are not always available.
In fact, their absence being the reason for the search.

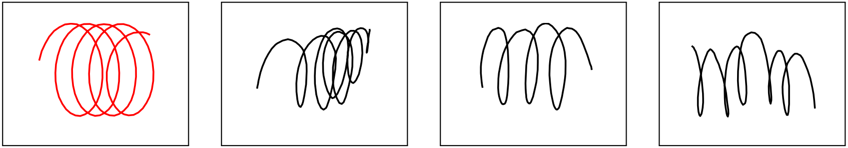
Why Sketch-based queries ?

Advantages

- Efficiently encodes information like shape, pose, colour, size etc. , all at once.
- A free-hand sketch is more convenient to draw than typing lengthy queries.
- A sketch is closer to the *content* of a video as compared to meta-data (tags, comments, captions).

Challenges in Sketch-based systems

Perceptual Variability



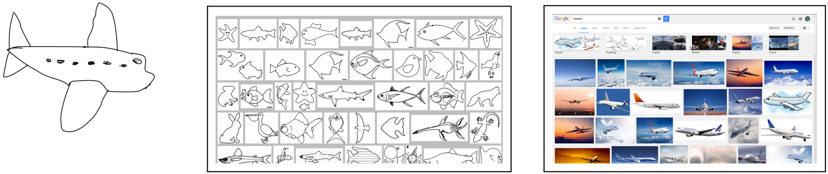
(a) (b) (c) (d)

Figure: Different interpretations of the same trajectory. (a) Original (b),(c),(d) User Inputs

Cognitive variation in motion perception in human beings .

Challenges in Sketch-based systems
Multimodality

An image is a multi-channel dense representation of an object/scene.



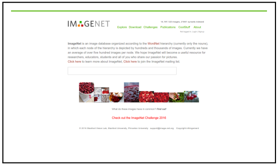
(a) (b) (c)

Figure: (a) Query (b) Results (c) Desired Output

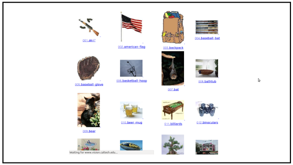
"A simple sketch is a high level sparse representation of the object/scene being searched for." [Li et al., 2015]

Challenges in Sketch-based systems

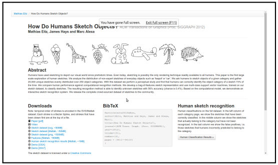
Scarcity of Data



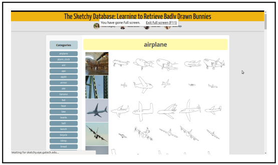
(a) ImageNet



(b) Caltech



(c)TU-Berlin



(d) Sketchy

Figure: Number of samples (a) 14,197,122 (b) 30,607 (c) 20,000 (d) 75,471

The system should be able to generalize to unknown classes.

Challenges in Sketch-based systems

Summary

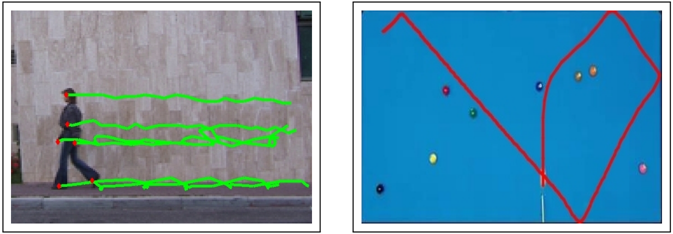
- Perceptual Variability.
- Multimodality.
- Scarcity of Data.

- Perceptual Variability : Video Retrieval
- Multimodality : Image Retrieval
- Scarcity of Data : Zero Shot Learning

Outline

Motion

Videos characterized by motion.



(a) (b)

Figure: (a) Human Tracking (b) Sports Analysis

[0]Image Courtsey : www.google.com

Features
Qualitative

Objective : Features minimize perceptual variability.

Qualitative Spatio Temporal Features

Features based on motion properties which tell us "how" rather than "how much".

- Shape
- Direction
- Scale

Combines 3 different aspects of motion.

Aspects of Motion
Shape

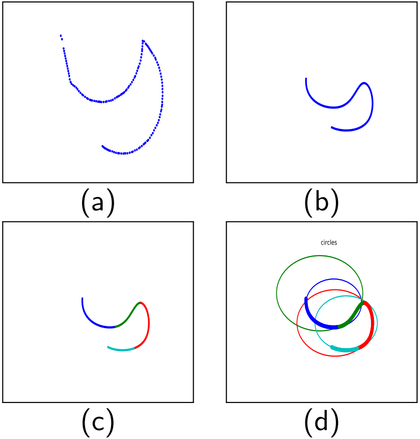
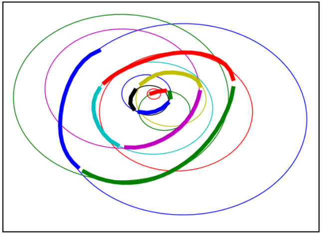


Figure: A sample motion with the corresponding *m*-segments: (a) Original (b) Smooth and Normalized (c) *m*-segments (d) Circle-Based Representation

Aspects of Motion
Shape

$$J = \min_{x_0, y_0, r} \sum_{i}^{n} x_i^2 + y_i^2 - 2x_0x_i - 2y_0y_i + x_0^2 + y_0^2 + r^2$$



$S = (\ x_\mu,\ y_\mu,\ r,\ w,\ s\ )$

$(x_\mu, y_\mu), r =$ center, radius of the circle.

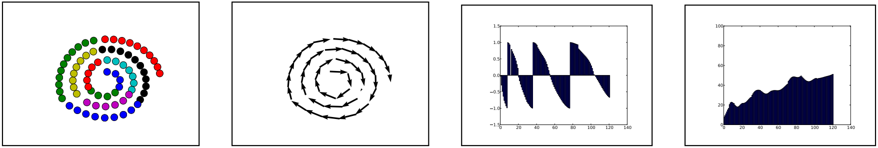
$w =$ Slope of the segment.

$s =$ Normalized length of arc.

K-Means, Bag of Motion

Aspects of Motion

Scale, Direction



(a) (b) (c) (d)

Figure: A Spiral Motion from our synthetic dataset (a) Points sampled equidistantly in each segment (b) Directions tracked for each equipoint segment (c) Temporal Change of Direction (d) Temporal Change of scale

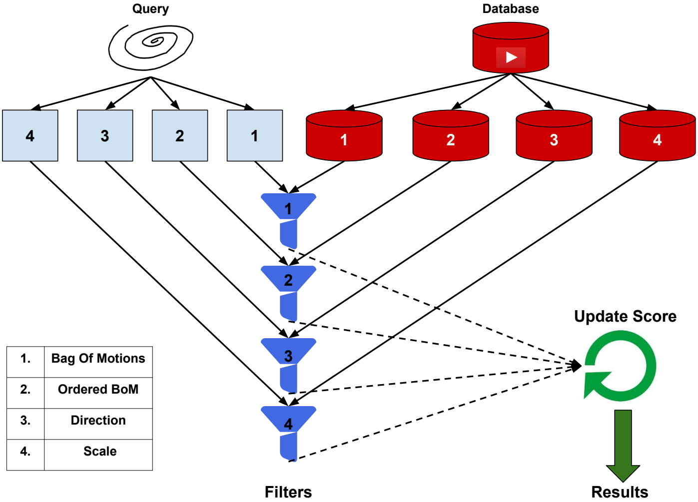
$$\text{Trajectory Direction} = (\alpha_1, \alpha_2, \ldots, \alpha_n)$$
$$\alpha_k = \sin\ \theta_k$$
$$\text{Trajectory Scale} = (d_1, d_2, \ldots, d_n)$$
$$d_k = \text{distance of the current segment from the mean.}$$

Summary of Features

4 types

- **Bag of Motion :** Trajectory = Histogram.
- **Ordered Bag of Motion :** *Trajectory* $= (s_1, s_2, \ldots, s_m)$, where $s_k = (\ x_\mu,\ y_\mu,\ r,\ w,\ s\ )_k$
- **Direction :** $= (\alpha_1, \alpha_2, \ldots, \alpha_n)$
- **Scale :** $= (d_1, d_2, \ldots, d_n)$

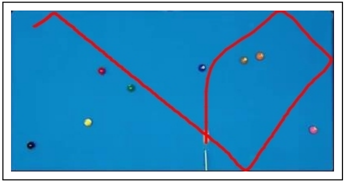
Pipeline



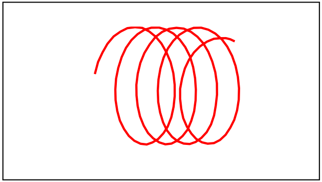
1.	Bag Of Motions
2.	Ordered BoM
3.	Direction
4.	Scale

Figure: Cascaded Retrieval

Datasets



(a) Pool Dataset



(b) Synthetic Dataset.

Figure: (a) Five classes each containing 20 videos each. (b) Five classes containing 20 videos each. Thus the dataset had 200 videos

A Sample Retrieval

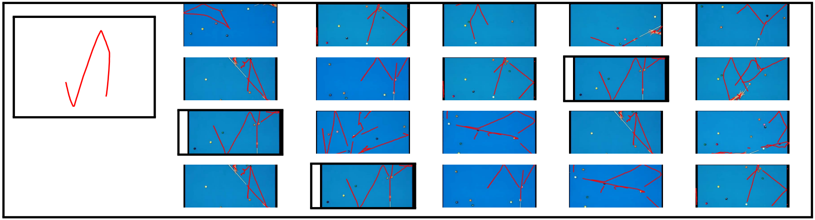
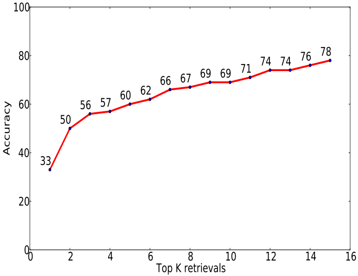


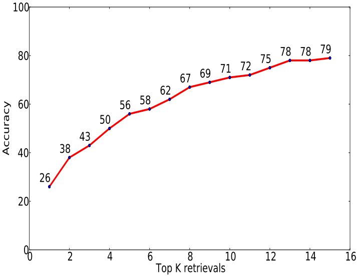
Figure: Qualitative Results

Results

Quantitative Results : Accuracy



(a) Pool Videos dataset

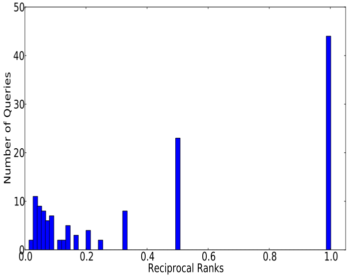


(b) Synthetic Motion dataset

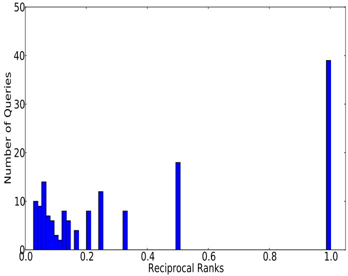
Figure: Accuracy at different top K retrievals.

Results

Quantitative Results : Mean Reciprocal Rank



(a)Pool Videos dataset



(b) Synthetic Motion dataset

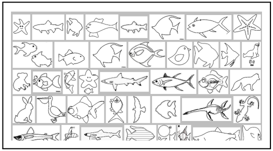
Figure: Mean Reciprocal Ranks

Outline

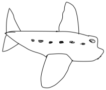
1. Motivation
2. Video Retrieval
3. Image Retrieval
4. Zero-Shot Learning

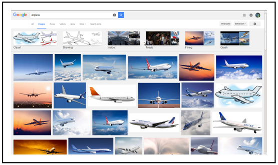
Sparsity of Sketches

Two different modalities



(b)





(a) (c)

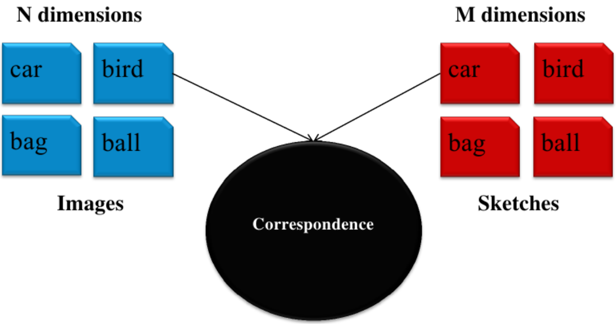
Figure: (a) Query (b) Results (c) Desired Output

Should not be compared directly.

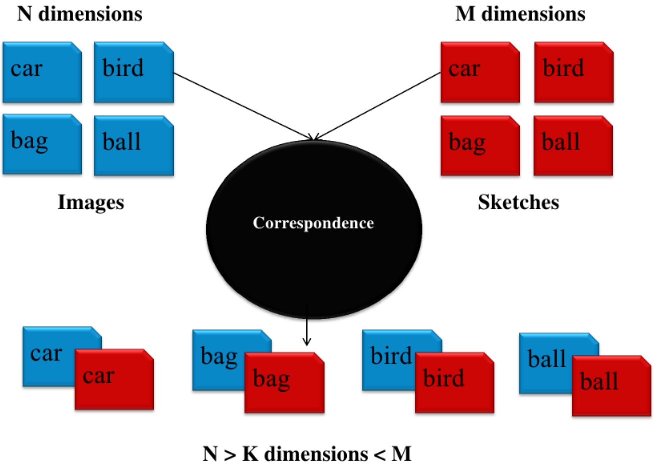
Our Model

Two Modalities

Our Model
Correspondence

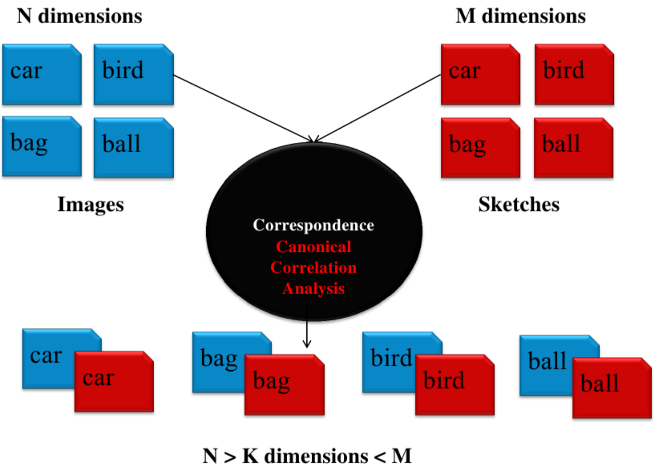


Our Model

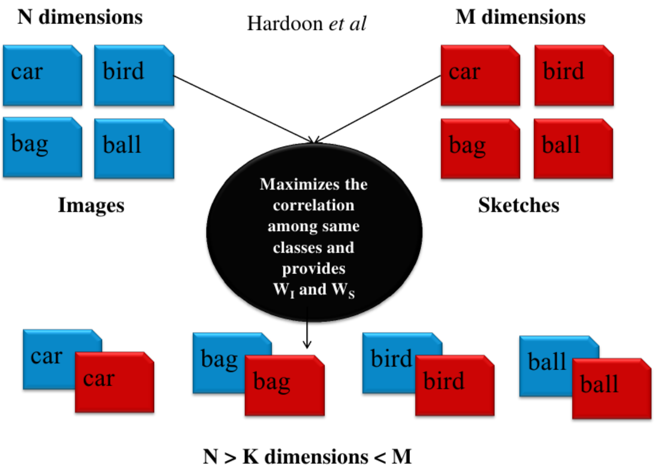


N dimensions
M dimensions
car
bird
bag
ball
Images
car
bird
bag
ball
Sketches
Correspondence
car
car
bag
bag
bird
bird
ball
ball
N > K dimensions < M

Our Model


N dimensions
M dimensions
car
bird
bag
ball
Images
car
bird
bag
ball
Sketches
Correspondence
Canonical
Correlation
Analysis
car
car
bag
bag
bird
bird
ball
ball
N > K dimensions < M

Our Model
Two Modalities



Standard CCA

Given two sets I and S,

$$I = (I_1, I_2, \dots, I_n)$$

$$S = (S_1, S_2, \dots, S_n)$$

we try two find two subspaces,

$$P^I =< W_I, I >, P^S =< W_S, S >$$

such that that their correlation is maximized.

$$\begin{aligned} \rho &= \max_{W_I, W_S} corr\left(P^I, P^S\right) \\ &= \max_{W_I, W_S} \frac{\langle P^S, P^I \rangle}{\|P^S\| \|P^I\|} \end{aligned} \tag{1}$$

Standard CCA

continued...

As derived in [Hardoon et al., 2004], Equation 1 reduces to,

$$\rho = \max_{W_I, W_S} \frac{W_I' \mathrm{Cov}_{IS} W_S}{\sqrt{W_I' \mathrm{Cov}_{II} W_I} \sqrt{W_S' \mathrm{Cov}_{SS} W_S}} \tag{2}$$

and the covariance matrix of (A_I, A_S) given by:

$$Cov = \mathbb{E}\left[\begin{pmatrix} A_I \\ A_S \end{pmatrix} \begin{pmatrix} A_I \\ A_S \end{pmatrix}' \right] = \begin{bmatrix} Cov_{II} & Cov_{IS} \\ Cov_{SI} & Cov_{SS} \end{bmatrix} \tag{3}$$

Equation 2 can be solved as an Eigen value problem for W_I and W_S.

Cluster CCA

As suggested in [Rasiwasia et al., 2014], we compute the covariance matrices as follows,

$$Cov_{IS} = \frac{1}{M}\sum_{c=1}^{C}\sum_{j=1}^{|I^c|}\sum_{k=1}^{|S^c|} I_j^c S_k^{c'} \tag{4}$$

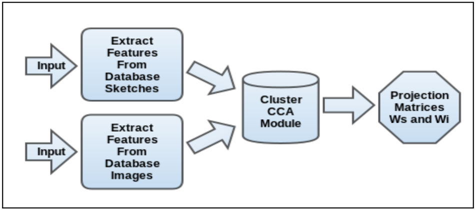
$$Cov_{II} = \frac{1}{M}\sum_{c=1}^{C}\sum_{j=1}^{|I^c|} |S^c| I_j^c I_j^{c'} \tag{5}$$

$$Cov_{SS} = \frac{1}{M}\sum_{c=1}^{C}\sum_{k=1}^{|S^c|} |I^c| S_k^c S_k^{c'} \tag{6}$$

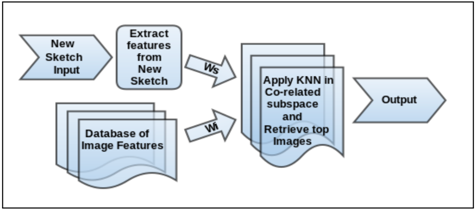
where $M = \sum_{c=1}^{C} |I^c||S^c|$, is the total number of pairwise correspondences across C classes.

Pipeline

Training and Testing



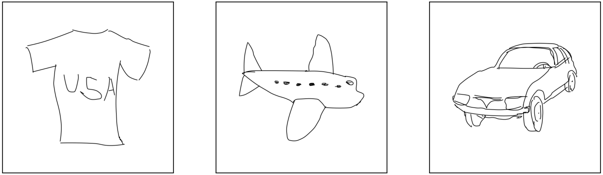
(a)



(b)

Figure: Proposed pipeline (a) Training (b) Retrieval

Datasets



TU Berlin Dataset, 250 categories, 80 sample each category

Caltech 256, Pascal VOC 2007

Features

Table 1 : Summary of Features

Feature	Dimension	Source
CALTECH - SIFT	1000	Vl-Feat. [Vedaldi and Fulkerson, 2008]
CALTECH - HOG	20000	Vl-Feat [Vedaldi and Fulkerson, 2008]
CALTECH - CNN	4096	Krizhevsky *et al.* [Krizhevsky et al., 2012]
PASCAL - SIFT	1000	Guillaumin *et al.*in [Guillaumin et al., 2010]
PASCAL - HOG	20000	Vl-Feat [Vedaldi and Fulkerson, 2008]
PASCAL - CNN	4096	Krizhevsky *et al.* [Krizhevsky et al., 2012]
TU-BERLIN - SIFT-Like	501	Eitz *et al.*in [Eitz et al., 2012]
TU-BERLIN - HOG	20000	Vl-Feat [Vedaldi and Fulkerson, 2008]
TU-BERLIN - Fisher	250000	Rosalia *et al.* [Schneider and Tuytelaars, 2014]
TU-BERLIN - CNN	4096	Yang *et al.* [Yang and Hospedales, 2015]

Results

Qualitative

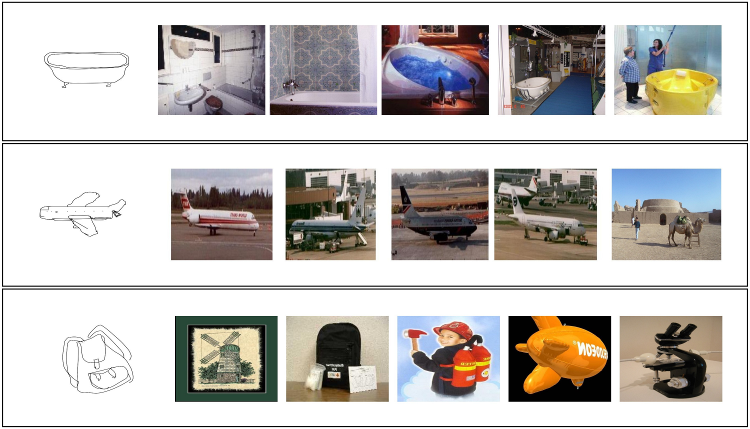


Figure: Example Retrievals From Our System

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

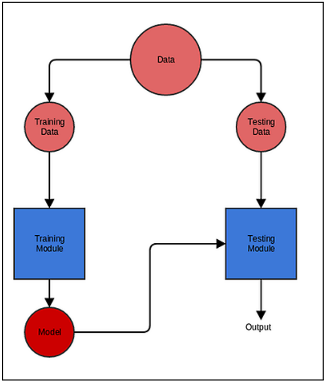
Dataset	SIFT-SIFT	SIFT-HOG	SIFT-Fisher	HOG-SIFT	HOG-HOG	HOG-Fisher	CNN-CNN
Caltech	0.06	0.03	**0.20**	0.14	0.02	0.01	**0.20**
Pascal	0.13	0.12	0.05	**0.18**	0.09	0.06	0.06

Table : Performance improvement in mAP values

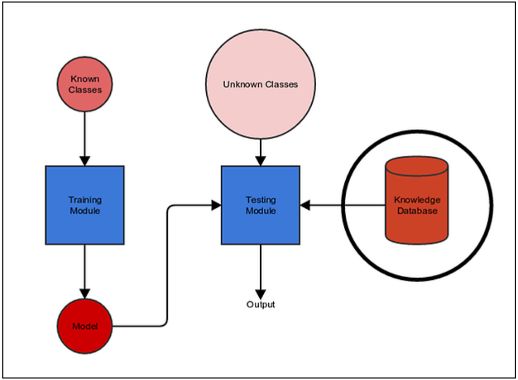
Dataset	Features	Before CCA	After CCA
Caltech	SIFT-Fisher	0.01	**0.20**
Caltech	CNN-CNN	0.01	**0.20**
Pascal	HOG-SIFT	0.01	**0.18**
Pascal	SIFT-SIFT	0.06	**0.13**

Outline

Zero-Shot Learning



(a)



(b)

Figure: (a) Standard Classifier (b) Zero-Shot Classifier

Knowledge Database acts as an Oracle, who knows everything

Word 2 Vec

- Word 2 Vec by Mikolov *et al.* [Mikolov et al., 2013] is a vector space, where semantically similar words are mapped together.
- *Apples* and *Oranges* are closer than *Apples* and *Mumbai*.
- The distance between two classes in Word 2 Vec space represents their semantic similarity.

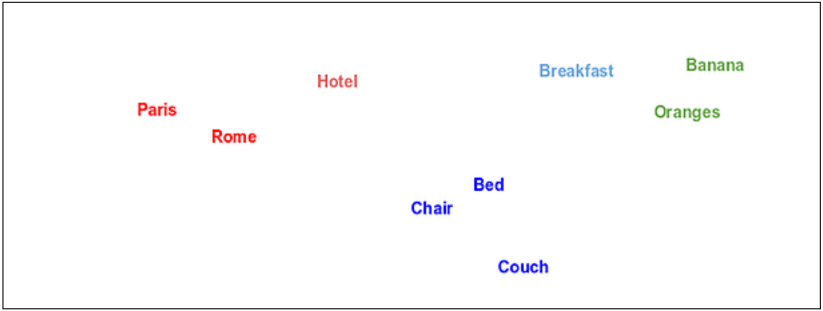


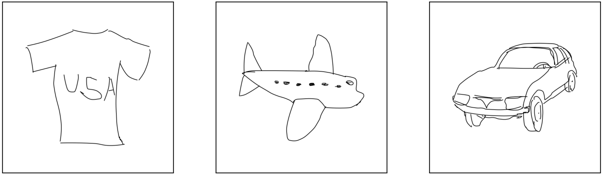
Figure: An artistic expression of Word2Vec vector space.

We train a CNN as in [Yu et al., 2015] for TU Berlin and as in [Chatfield et al., 2014] for Caltech256 dataset and replace the soft-max layer as follows.

$$J(\theta) = \sum_{y \in \mathcal{Y}} \sum_{x^i \in \mathcal{X}_{\mathcal{Y}}} ||w_y - g(x^i)||^2$$

A close miss is penalized less than a distant miss.

Datasets



TU Berlin Dataset, 250 categories, 80 sample each category

Caltech 256, Pascal VOC 2007

Results

Qualitative

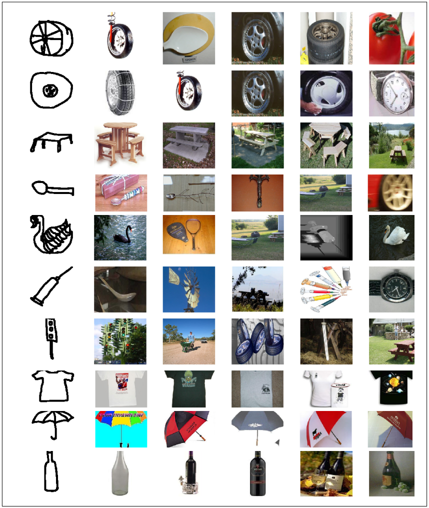


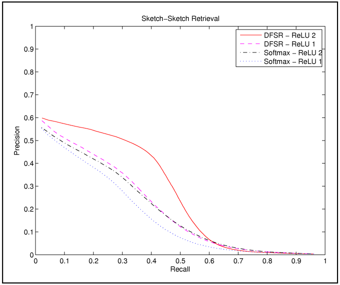
Figure: Example Queries

Experiments

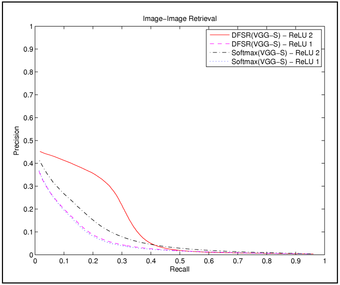
- Sketch Classification.
- Uni-Modal Sketch Retrieval.
- Cross-Modal Retrieval.
- Zero Shot Retrieval.

Algorithm	Features	Classifier	Accuracy
	TU-Berlin Dataset		
Yang *et al.* [Yu et al., 2015]	CNN-Ensemble	SOFT-MAX	74.9%
Yang *et al.* [Yu et al., 2015]	CNN-Single	SOFT-MAX	72.6%
Schneider *et al.* [Schneider and Tuytelaars, 2014]	FISHER	SVM	63.1%
Eitz *et al.* [Eitz et al., 2012]	BOW	SVM	56%
Proposed	**DFSR**	RANDOM FOREST	70.22%

Table: Classification Results show that our features perform reasonably well almost at par with the state of the art methods.



(a)



(b)

Figure: Uni-modal retrieval : (a) Sketch Modality (b) Image Modality

Results
Cross-Modal Retrieval

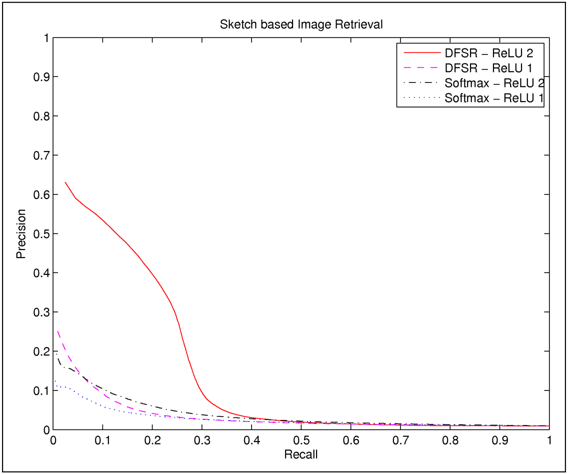


Figure: Cross Modal Retrieval

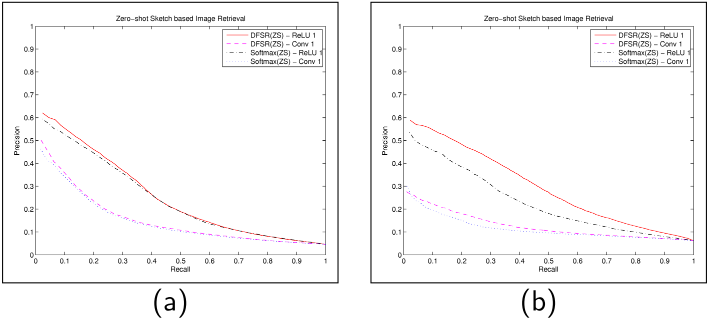


Figure: Zero-Shot Cross-Modal Retrieval : (a) PR curve for the worst performing partition. It can be observed that DFSR outperform the features proposed by Yang *et al.* [Yang and Hospedales, 2015]. (b) PR curve for the best performing partition.

Summary

Figure: Related Domains

- Qualitative features for video retrieval.
- Cluster CCA for Multi Modal Image Retrieval.
- Deep Features for Semantic Rerieval.

References I

Chatfield, K., Simonyan, K., Vedaldi, A., and Zisserman, A. (2014). Return of the devil in the details: Delving deep into convolutional nets. In *British Machine Vision Conference*.

Eitz, M., Hays, J., and Alexa, M. (2012). How do humans sketch objects? *ACM Trans. Graph.*

Guillaumin, M., Verbeek, J., and Schmid, C. (2010). Multimodal semi-supervised learning for image classification. In *CVPR*.

Hardoon, D., Szedmak, S., and Shawe-Taylor, J. (2004). Canonical correlation analysis: An overview with application to learning methods. *Neural computation*.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *NIPS*.

Li, Y., Hospedales, T. M., Song, Y.-Z., and Gong, S. (2015). Free-hand sketch recognition by multi-kernel feature learning. *CVIU*.

References II

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Rasiwasia, N., Mahajan, D., Mahadevan, V., and Aggarwal, G. (2014). Cluster canonical correlation analysis. In *AI Statistics*.

Schneider, R. G. and Tuytelaars, T. (2014). Sketch classification and classification-driven analysis using fisher vectors. *TOG*.

Vedaldi, A. and Fulkerson, B. (2008). VLFeat: An open and portable library of computer vision algorithms. `http://www.vlfeat.org/`.

Yang, Y. and Hospedales, T. M. (2015). Deep neural networks for sketch recognition. *arXiv preprint arXiv:1501.07873*.

Yu, Q., Yang, Y., Song, Y.-Z., Xiang, T., and Hospedales, T. M. (2015). Sketch-a-net that beats humans. In *BMVC*.

Thank you.
koustav.ghosal@research.iiit.ac.in