A Sketch-based Approach for Multimedia Retrieval

Thesis submitted in partial fulfillment
of the requirements for the degree of

MS (by research)
in
Computer Science

by

Koustav Ghosal
201207530
koustav.ghosal@research.iiit.ac.in

Center for Visual Information Technology.
International Institute of Information Technology
Hyderabad - 500 032, INDIA
September, 2016

International Institute of Information Technology
Hyderabad, India

CERTIFICATE

It is certified that the work contained in this thesis, titled "A Sketch-based Approach for Multimedia Retrieval" by Koustav Ghosal, has been carried out under my supervision and is not submitted elsewhere for a degree.

_____ _____

Date Adviser: Prof. Anoop Namboodiri

To my family and friends for their constant support

Acknowledgments

I would like to thank my colleagues from **Centre for Visual Information Technology, IIIT-Hyderabad**, who spent their valuable time to read and review this work since its inception. This research owes a lot, if not the most of it, to those group discussions, coffee breaks and midnight snacks.

I am also grateful to my supervisor, Prof. Anoop Namboodiri for his deep and innovative insights towards the various problems addressed in this thesis. I would like to thank Ameya for his contribution in this work, especially in the last chapter. I would also like to thank all the reviewers who gave their thorough and insightful reviews for this work. Most importantly, I take this opportunity to thank all faculty and staff members in CVIT without whose consistent efforts and perseverance, this research would have been impossible.

Abstract

A hand-drawn sketch is a convenient way to search for an image or a video from a database where examples are unavailable or textual queries are too difficult to articulate. In this thesis, we have tried to propose solutions for some problems in sketch-based multimedia retrieval. In case of image search, the queries could be approximate binary outlines of the actual objects. In case of videos, we consider the case where the user can specify the motion trajectory using a sketch, which is provided as a query.

However there are multiple problems associated with this paradigm. Firstly, different users sketch the same query differently according to their own perception of reality. Secondly, sketches are sparse and abstract representations of images and the two modalities can not be compared directly. Thirdly, compared to images, datasets of sketches are rare. It is very difficult, if not impossible to train a system with sketches of every possible category. The features should be robust enough to retrieve classes that were not a part of training.

In this thesis, the work can be broadly divided into three parts. First, we develop a motion-trajectory based video retrieval strategy and propose a representation for sketches that aims to reduce the perceptual variability among different users. We also propose a novel retrieval strategy, which combines multiple feature representations for a final result using a cumulative scoring mechanism.

In order to tackle the problem of multiple modalities, we propose a sketch-based image retrieval strategy by mapping the two modalities into a lower dimensional sub-space where they are maximally correlated. We use Cluster Canonical Correlation Analysis (c-CCA), a modified version of standard CCA, for the mapping.

Finally, we investigate the use of semantic features derived from a Convolutional Neural Network, and extend the idea of sketch-based image retrieval to the task of zero-shot learning or unknown class retrieval. We define an objective function for the network such that, while training, a close miss is penalized less than a distant miss. Our training encodes semantic similarity among the different classes. We perform experiments to evaluate our algorithms on well known datasets and our results show that our features perform reasonably well in challenging scenarios.

Contents

List of Figures

List of Tables

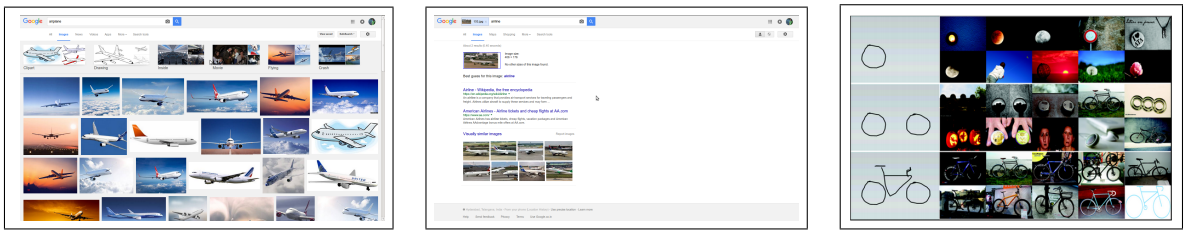
Chapter 1

Introduction

Recent reports [78, 6] indicate that we upload and share over 300 million images every day on social media. When it comes to videos the numbers are 300 hours of video uploaded per minute for YouTube [22]. *Big-Data* is not anymore a trouble of the future but the challenges it poses to data-scientists around the globe are becoming one of the, if not the most active area of research in computer science. The amount of data available is also pushing the state of the art solutions from approaches like supervised learning, graphical models and support vector machines to unsupervised feature learning. In the philosophy of *let data speak for itself*, sophisticated feature learning techniques using Deep Neural Networks and Metric Learning Algorithms have become the de facto standard for last few years. With the surge of social networks, multimedia data (images and video) come with tags and comments, in other words, crowd-sourced annotation. Using this data, machine learning algorithms are evolving fast and making our lives easier with more intelligent search engines, customized recommendation systems, affordable medical diagnosis or automated surveillance systems.

While this enormous quantity of multimedia data is being generated, stored and analysed, it is challenging to organize and retrieve it, as and when required. Some of the most successful and popular platforms used for retrieving multimedia, like YouTube, Google, Bing, Baidu are continuously working to optimize the retrieval time and customize the search with the user. Although Content-based Multimedia Retrieval (CBMR) falls behind text-document retrieval in terms of research and efficient solutions available, it is quite an active area of research for a few years now.

In this thesis, we have worked on a specific area of content-based multimedia retrieval, where the form of query is a simple hand-drawn sketch. We have explored sketch-based query mechanisms for both images and videos. We have focussed primarily on developing a representation for sketches for more accurate and worthwhile results. Issues like search-time complexity and database management have not been investigated. But efforts have been made to retrieve results that are more meaningful and closer to the actual content of the multimedia document being searched for. The proposed representations however, can be fused with any memory-efficient database management technique and fast indexing algorithms to build complete sketch-based multimedia retrieval systems.



(a)Text Based (b) Example Based (c)Sketch Based

Figure 1.1 Query Mode : Different types of content based image retrieval.

1.1 Content Based Multimedia Retrieval (CBMR)

When we search for an image or video in Google or YouTube, we generally provide keywords describing the content. It is the most widespread approach towards CBMR. There are other less known systems that provide alternative forms of queries. For example in *Video Google* [70] the users look for a particular object in a video from an example patch and in *MindFinder* [12] an image is searched from a database of millions of images, using a sketch provided by the user. Most of the research in CBMR has been concentrated around text-based approaches but the other possibilities have not been investigated thoroughly. In this section, we classify multimedia retrieval into different categories based on the type of query and modality.

1.1.1 CBMR : Categories

We divide CBMR in two ways, based on *query type* and *query modality*. One might consider these two genres as two different ways of looking at CBMR. Each categorization can be subdivided into different sub-categories.

Query Type

Depending on the mode of query, there are three primary types of CBMR systems. See Figure 1.1.

1. **Text-based :** In text-based systems, similar keywords from meta data space (tags and annotations) associated with the multimedia data are searched. But the meta data is generally not reliable as it may not represent the actual content in the data or could be misleading. Apart from that, pertaining to perceptual variability, different users may use entirely different queries to search for the same images/videos, requiring NLP techniques for correct interpretation.

2. **Example-based :** Here, an example image, video or audio clip is provided as a query to the system. Results similar to the query are retrieved from the database by directly comparing the query and data in some particular feature space. This paradigm, however, is limited by the fact

that example clips are not always available at hand, in fact their absence being the reason for a search.

3. **Sketch-based :** In sketch-based systems, the query comes in the form of a hand-drawn user-sketch. While such systems address some of the key issues associated with the other two modes, they come with a set of new challenges. We summarize them separately in Section 1.2.

Modality

Modality is the digital format in which data is represented. For example each of text, audio, image and video represent a modality. Based on the modality of the query and results, CBMR can be categorized into *unimodal*, *multimodal* and *hybrid* systems.

1. **Unimodal Systems :** If the query and database belong to the same modality the system is unimodal. For example, systems that present image-image [18], audio-audio [40], video-video [70] combinations as query-result pair are unimodal systems.

2. **Multimodal Systems :** On the other hand text-video (YouTube), text-image (Google Image Search) combinations for the same pertain to multimodal model.

3. **Hybrid Systems :** The hybrid models combine the above two categories. For example, the query itself could be multimodal. For example while retrieving action videos [37], Hu *et al.* used a combination of sketch and text as the query mode.

1.2 Sketch-based Approach : Advantages

Following the discussion in the previous sections we can deduce that Sketch-based Multimedia Retrieval (SBMR) is a multi-modal retrieval strategy. In this section, we discuss the reason why and where the sketch-based queries provide an efficient and intuitive way to search for images or videos.

1.2.1 Video Search

In Computer Vision, motion has been used as the primary *content* of a video in *Content Based Video Retrieval Systems(CBVRS)*. The motion-based analysis of a video is frequent in surveillance, human-machine interaction, automatic target recognition and automotive applications [8]. Most existing approaches primarily have two phases in a CBVRS pipeline. In the first phase, trajectories of different objects are extracted from the videos and are stored. In the next phase, when a query (example videos, keywords or sketch) is presented to the system, it is matched with all the stored trajectories in the database and the corresponding video is retrieved. As mentioned in Section 1.1.1 a serious drawback with example based queries is that an example is not always available in real time scenarios. Text based queries, on the other hand, apart from the meta-data space search problem, are not suitable to describe

(a) (b) (c)

Figure 1.2 (a) A sketch of a car from a front-view (b) Search results that 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.

long and complicated motions. For example, queries like "*the first strike in carrom where three or more carrom men or disks go to pockets*" or "*a particular diving style in swimming tutorials where the swimmer does three somersaults before diving*" are very difficult to frame. In such scenarios, sketch is a very convenient mode to encode many-fold information within the query itself.

1.2.2 Image Search

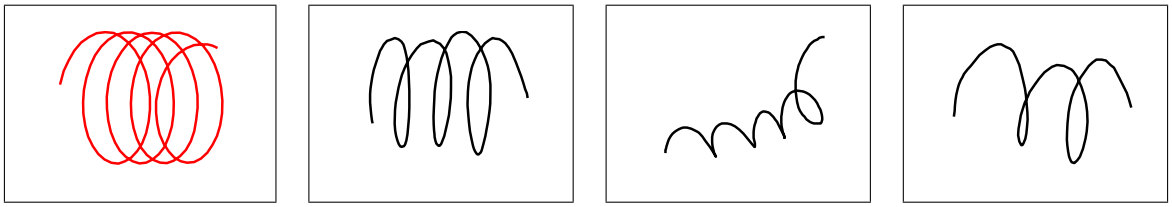
All the advantages involved in video retrieval are there in image retrieval as well, with minor variations. Unlike videos, where we consider only videos with motion, images come in varied forms and categories. It amplifies the scope for errors pertaining to perceptual variability.
For example, as shown in Figure 1.2, if a user needs to search for the image of a car from the front-view, a sketch as in Figure 1.2(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 1.2(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 1.2(c).

1.3 Challenges

However, there are several challenges involved in using sketches for multimedia retrieval. We list some of them in this section, which have been addressed in the subsequent chapters.

1.3.1 Perceptual Variability

The user perception (sketch) of a video is only an abstraction of the same. All the properties of the trajectory like shape, length and position are merely approximations of the trajectory of the object



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

Figure 1.3 (a) Original Motion Trajectory. (b) - (d) Interpretations of the same motion by different users

in the video. A simple Euclidean distance match is not bound to yield any meaningful results. Apart from spatio-temporal variability, the different sketches of the same trajectory by different users also suffer from perceptual variability. In other words, humans perceive motion in a way that is *qualitatively* similar but differ *quantitatively*. This is further elucidated by Figure 1.3.

1.3.2 Multiple Modality

In addition to the perceptual variability, 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.* [46], a simple sketch is a high level sparse representation of the object/scene being searched for. Yong *et al.* [82] found that because of this sparsity, when a sketch is presented as a query to Clarifi [84], cartoon images, which resemble the sketches markedly, are retrieved.

1.3.3 Paucity of Training Data

The paucity of training data for modelling the sketches is a major challenge to sketch-based multimedia retrieval. Although there is an abundance of images of millions of categories over the internet [63], finding the same for sketches is too cumbersome, if not impossible. For example, the TU-Berlin dataset is the biggest known repository for sketches till date.It consists of 250 object categories, contains *chairs* but doesn't contain *sofa*. Thus a model that is trained on this dataset, might retrieve chairs when the sketch of a sofa is provided as a query because of structural similarity.
But the chance of any other object having similar structure to appear is equally likely. The state of the art systems do not have a provision to capture the semantic similarity between different classes. This necessitates a *Zero-Shot* learning [58] model that can retrieve novel classes, which do not appear in training but are semantically similar to the classes that do. In fact, the key incentive of using sketch as a query is it's flexibility to represent *rare* objects that are both semantically and morphologically similar to more familiar objects.

1.4 Major Contributions

In this thesis we investigate the above mentioned challenges and address them methodically. It is organized as follows.

- In Chapter 3, we explore sketch-based video retrieval. A novel representation of a motion trajectory has been proposed that tries to remove the spatio-temporal variability among sketches of different users. Qualitative features have been derived, whose attributes tell us "how" rather than telling us "how much" about the different aspects of a motion.
  We also propose an efficient multilevel cascaded retrieval method with a cumulative scoring mechanism which boosts the retrieval accuracy at each stage of the cascade.

- In Chapter 4, we address the problem of multiple modality in sketch-based image retrieval (SBIR). We model the correspondence between the two modalities, 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. 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.

- In Chapter 5, we address the problem of SBIR in the light of zero-shot learning. We propose a novel feature representation called Deep Features for Semantic Retrieval (DFSR) using Convolution Neural Networks, in which we embed both morphological and semantic properties of object categories. Using an existing knowledge corpus about the semantic similarity among classes [52], the objective function for our network is formulated in such a way that during training, it penalizes the misclassified samples based on their semantic similarity with the original class. In other words a semantically close miss is penalized less than a distant miss, based on the magnitude of the error.

In each chapter, we perform experiments to validate the alleviation of the problems addressed with well-known standard datasets. We evaluate our methods using standard metrics and show that they perform well in challenging scenarios.

Chapter 2

Related Work

2.1 Introduction

In this thesis, we address different aspects of the problem of sketch-based multimedia retrieval. The aspects span across multiple domains and each domain provides us with it's own challenges and limitations , see Figure 2.1. In this chapter, we briefly point out the areas which are releveant to our problem and mention some of the important contributions which have been made in each of these areas.

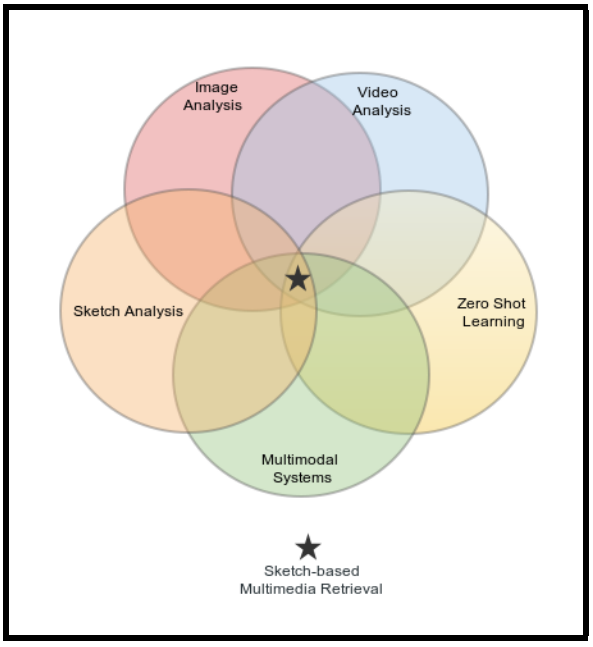


Figure 2.1 The domain of sketch-based multimedia retrieval spans across multiple domains of Image Analysis, Video Analysis, Sketch Analysis, Multimodal Systems and Zero Shot Learning.

2.2 Content Based Image Analysis

This work is closely associated with image classification and retrieval, both of which are two classical problems in Computer Vision and one finds a plethora of sophisticated techniques for these tasks. They have several applications like medical imaging, remote sensing, robotics, fashion, biometrics etc. During it's initial days, the research in this area was limited by the unavailability of enough data. But with the advent of hand-held cameras and major development in imaging technology, this area evolved rapidly, and so did the complexities involved.
Traditional techniques like SIFT [49] and HOG [21] and Fisher vectors [60] perform well for image classification tasks. They have been extensively used on datasets like Caltech [31] and Pascal [26].

For the past few years, research in this domain has been mostly directed towards Deep Learning. Convolutional Neural Networks have been around for a while, being first introduced by Lecun *et al.* [44] in 1989, but have recently become popular with the success achieved by Krizhevsky *et al.* [42]. In 2010, the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) was introduced with a dataset of 1000 classes and over 1 million images. Different groups participate in the challenge and publish their scores which get better every year.
As pointed out in [71], based on the user requirement image search can be divided into three categories

Year	Datasets	Features
Before 2010	Caltech, PASCAL	HOG, SIFT, FISHER [79]
After 2010	ImageNet, MS Coco	Deep Features [42]

Table 2.1 Research in Image Analysis can be roughly divided into two phases. Pre and Post Image Net Challenge. Introduction of Deep Learning and use of CNN's have brought a major change in the paradigm.

— *search by association* where the objective of the search is not fixed, *exact search* where the user seeks for an exact match and *category search* where the user looks for images of similar categories. Sketch is an efficient way to frame the queries in all of these categories. A careful sketch aims at finding the exact match whereas a rough sketch could look for similar categories.
However, images are dense sources of information whereas free-hand sketches are sparse representations of objects or scenes. Thus the feature representations which work well with images fail to work well on sketches. A direct mapping from images to sketches is not possible and a different approach is necessary. We try to propose solutions which address this problem.

2.3 Content Based Video Analysis

Video Retrieval using motion trajectories can broadly be categorized into two different modules of a pipeline — *Trajectory Extraction* and *Query Indexing*.

In trajectory extraction, the objects are initially extracted from a key frame and then tracked across successive frames. Foreground segmentation in videos has been an extensively researched problem and several algorithms have been proposed for the cases of static [86], [74], [30] and dynamic backgrounds [54], [50], [68]. For tracking, standard techniques like Kalman Filters [4], Mean Shift Algorithm [16] and Double Exponential Smoothing [38] have been proposed. Once the trajectories are extracted, they are modelled by motion features like velocity, acceleration, curvature and length.

VideoQ [13], which is one of the first Content Based Video Retrieval Systems using sketch (sCBVRs), takes as an input a sketch, containing colour and shape based features and uses wavelet decomposition to model each trajectory. Alternative approaches to process trajectories like statistical modelling, Principal Component Analysis of sub-trajectories, MPEG based motion flow extraction methods have also been proposed. An exhaustive survey of these techniques can be found in Hu *et al.* [39]. This paradigm has been applied to the problem of event detection and activity classification as well [37]. Bashir *et al.* [2] and Cuntoor *et al.* [19] proposed HMM based approaches for trajectory based activity classification. Basharat *et al.* [1], Saleemi *et al.* [64] , Stauffer *et al.* [75] have modeled traffic behavior using spatio-temporal information from videos. Dyana *et al.* [23] have used a multispectro-temporal curvature scale space (MST-CSS) representation to describe a video object.

Areas like Video Summarization and Action Recognition are also closely related. For example in [17], the authors present an algorithm which provides narratives for the video in the form of stroboscopic images. In another work [15], Chen *et al.* develop a method for video summarization using metadata associated with the videos.

An illustration of the most common problems in video analysis has been provided in Figure 2.2 [1].

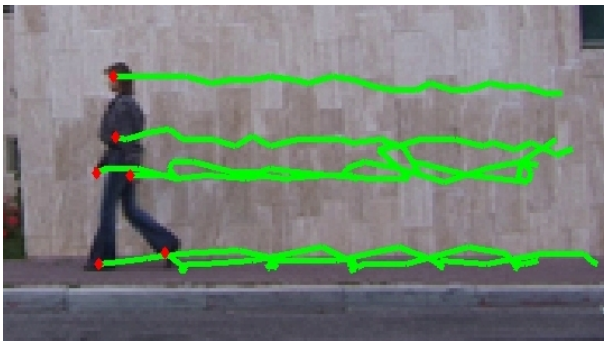
2.4 Sketch Based Approaches

In the last two decades, sketch recognition has been mainly limited to understanding gestures, mathematical symbols, alphabets and digits [67, 33]. A more generic sketch recognition framework was proposed by Eitz *et al.* [24], where they extracted SIFT-like features from sketches. Cao *et al.* [11] used a symmetry aware flip invariant descriptor. Li *et al.* [47, 46] suggested similar solutions using star-graphs and multi-kernel feature learning. Rosalia *et al.* [66] encoded sketches as Fisher vectors which

[1] Image Courtsey : www.google.com

(a)

(b)



(c)

(d)

Figure 2.2 Different tasks associated with video analysis. (a) Summarization [15] : Generating Dynamic narratives from a movie clip is a type of summarization task (b) Traffic Analysis: Sketches can be used to identify trajectories from a traffic video (c) Action Recognition [7] : Sketches can represent actions as well (d) Tracking : It is an important task in video analysis

performed well. Recently, Yang *et al.* [82] designed a Deep Neural Network architecture on sketches.

Inspite of the plethora of motion-trajectory based video retrieval systems, according to our knowledge, there are very few generic sketch-based CBVR systems [13], [36], [81]. In [13], colour, shape and appearance of the objects have also been used for describing *content* of videos. These features work well when the database consists of videos that vary widely in content. On the contrary, if the videos are similar *e.g.* Pool or Billiards videos, ball trajectories have greater saliency than colour, shape *etc*, in terms of representing the video. Unlike surveillance videos, these trajectories are unconstrained with respect to direction and position. Sub-trajectory based matching as done by Chang *et al.* [13] is ideal for event search where the trajectories are short and number of sub-trajectories is limited. In case of longer trajectories, the temporal information is also important alongside spatial information and cannot be ignored.

Our work on videos is inspired and closely related to the work by Bashir *et al.* [3]. They have represented trajectories as a temporal ordering of the sub-trajectories by using Principal Component Analysis, Spectral Clustering and String Matching. Like their work, ours also relies on a stable trajectory extraction algorithm. But there are two fundamental differences between the two. Firstly, they have used *query by example* where the intention was to retrieve a similar set of trajectories from the database. Our query is *sketch based*, where the user intends to find an exact match. Since it is sketch-based, different users interpret the same motion according to their own perception, which differs quantitatively. Secondly, we have introduced a novel scoring mechanism that combines motion features like shape and direction in an efficient manner to refine the results.

Our work on images is motivated from and closely related to the work by Yang *et al.* [83] and Ghosal*et al.* [29]. Like [83], we also train a Convolutional Neural Network with the sketches of TU-Berlin dataset. However, we formulate a different objective function in which we penalize the closer misses lesser than the distant ones and thus encode inter-class semantics. We follow the Cluster-Canonical Correlation Analysis method adopted in text-image modalities by Rasiwasia *et al.* [62] to create the correspondence between the sketch and image modality during cross-modal retrieval.

2.5 Multimodal Systems

As explained in Section 1.3.2, sketches and images have sparse and dense representations respectively. Thus the same feature representation fail to perform in both these modalities simultaneously. This scales the problem at a higher level of difficulty and essentially this problem becomes a multimodal problem.

In other words, this scenario arises when we aim at learning from a source data distribution a well performing model on a different (but related) target data distribution. Examples of such cases, as pointed out by Patel *et al.* [59], include recognizing objects under poor lighting conditions and poses while algorithms are trained on well- illuminated objects at frontal pose, detecting and seg- menting an organ

Method	Summary
Cross-Modal Factor Analysis(CFA) [45]	Joint dimensionality reduction
Bilinear Model [77]	Joint dimensionality reduction
Canonical Correlation Analysis (CCA) [35]	Sample-wise Minimized Correlation
Cluster Canonical Correlation Analysis (CCA) [62]	Category-wise Minimized Correlation
Transfer Learning [61]	In transfer or multi-task learning, different tasks are considered, but the marginal distribution of the source and target data are similar.
Domain Adaptation [69]	Covariate Shift

Table 2.2 A summary of of classical and recent approaches to multimodal problems.

of interest from magnetic resonance imaging (MRI) images when available algorithms are instead optimized for computed tomography and X-ray images, recognizing and detecting human faces on infrared images while algorithms are optimized for color images *etc.*.

Cross-modal retrieval has been an active area for research for quite sometime. Ngiam *et al.* [56] developed a deep network architecture using Restricted Boltzmann Machines on features of two different modalities, audio and video. Klare *et al.* [41] tried to retrieve face images from forensic sketches. By formulating the problem as a joint dimensionality reduction problem, several methods like Cross-Modal Factor Analysis(CFA) [45], Bilinear Model [77] have been tried for the task. CCA was introduced by Hotelling *et al.* [35] to find relation between two sets of variates. Rasiwasia *et al.* [62], 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.
More recent approaches towards this problem involve Visual Domain Adaptation and Transfer Learning. A thorough study of all these approaches was beyond the scope of this thesis and we adapt to the classical approaches.
A brief summary of different approaches to multimodal learning has been illustrated in Table 2.2

2.6 Zero-Shot Learning

The major components of zero-shot learning is a set of *known* classes $\mathcal{Y}$, which is used to train a classifier $\mathcal{H}$ and a set of *unknown* classes $\mathcal{Z}$, which are not a part of the training. While training we

incorporate information about $\mathcal{Z}$ into $\mathcal{H}$ using a semantic knowledge database $\mathcal{K}$. The modality of $\mathcal{K}$ is independent of $\mathcal{Y}$ or $\mathcal{Z}$ and it has a representation which captures the semantic similarity between all classes. It acts as a knowledge database which contains information about both the known and unknown classes and encodes that information within the features during training.
Zero-shot-learning is recently being used in several domains for unknown class retrieval. Palatucci *et al.* [58] built a system which could predict unseen words from MRI signals. Socher *et al.* [72] use a large text corpus to build a vocabulary and used it to predict unseen classes. Norouzi *et al.* [57] propose a method which maps images into the semantic embedding space via convex combination of the class label embedding vectors, and requires no additional training. Elhoseiny *et al.* [25] propose a text-based image search engine which retrieves unknown categories of objects.
To the best of our knowledge, this is the first approach to perform zero-shot learning in a sketch-based image retrieval system.

In this work, the paucity of training data led to the idea of using zero-shot learning. While sketch is a convenient tool to depict rare/unseen classes, finding sketch samples for every category is expensive and tedious. Thus our problem suits well to this domain as well.

Chapter 3

Sketch-based Video Retrieval

3.1 Introduction

Motion has intrigued researchers in science and technology, sports, art, music, literature and films for ages. The trajectory of a missile, *Tiki-Taka* of Spanish football, the revolution of earth around the Sun, suspicious movements in railway stations — all these activities can be represented using a single or a combination of multiple motions. While motion itself conveys a lot of information for describing an event, depicting it (textually or pictorially) becomes a huge challenge for us. Depiction of motion in art has been there in five primary forms [20]- *Dynamic balance, multiple images, affine shear, blur, vectors*.

Dynamic balance or broken symmetry deals with the pose of an object in an image from which the

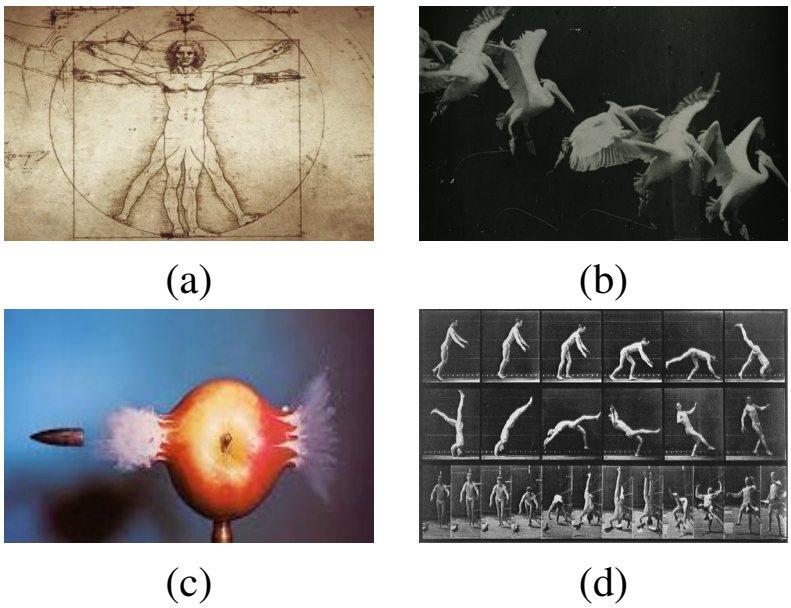


Figure 3.1 Masterpieces of the past [2] (a) *Vitruvian Man* by Leonardo Da Vinci (b)*Flying Pelican* by Étienne Jules Marey (c) *Bullet* by Edgarton (d) *Headspring* by Muyibridge

activity or event is predicted. A video can be summarized by overlaying key frames on one another and creating stroboscopic images [17]. (Figure 3.1 (b)). Affine shear and blur are two well known methods that are used to represent motion in graphic engines and comics. Vectors, on the other hand, are closest to human perception when it comes to representing motion [28].
As we mentioned in Section 1.2, in this work, we use sketches of motion trajectories as queries to search for videos. In this chapter, we propose a new method of modelling sketch based queries which attempts to extract the qualitative features of motion by minimizing the perceptual variability. We also develop a multilevel filter for indexing a query, in which the search results are refined at each stage using a cumulative scoring mechanism. Finally, we show the effectiveness of our algorithm on a dataset of real pool videos and a synthetic dataset containing simulated videos having very complex motion trajectories.

3.2 Motion Features

We first define a feature representation *that captures the constraints among dimensions rather than their quantitative values* [73]. In Sections 3.2.1 and 3.2.2, we explain our strategy to model the sub-trajectories in user sketch and the original videos respectively. At the end of Section 3.2.2, we show how these sub-trajectories can be used to model the entire trajectory. In Section 3.2.3, we derive another set of features that represent directional characteristics of motion.

3.2.1 User Sketch

Query or sketch is obtained as a collection of (x, y, t) points where $x,\ y$ and t represent x coordinate, y coordinate and time respectively. While collecting data, the users were shown some videos, randomly sampled from the dataset and then asked to recollect as many motion trajectories from the videos as they can. But the match was carried out with the longest one among all the trajectories in a given video. In our case, longer trajectories were assumed to be more salient than shorter ones. In a different scenario, there could be other metrics to measure saliency.

The trajectory is first de-noised, freed from outliers and smoothened using the conventional spline interpolation [51]. A video may contain multiple motions. The trajectories are then normalized (empirical results showed that height, normalized to 100 gave best results) in such a way that the relative position and size of the trajectories remain intact. The aspect ratio of each trajectory is preserved. This relative normalization strategy gives the sketch translational invariance and also preserves their relative attributes. The trajectories are subsequently segmented based on curvature. We call each segment a *motion segment* (m-segment), a term inspired from *ballistic segment*, frequently used in modelling handwriting [76] (see Figure 3.2). Here, we assume that a motion can be random and unconstrained but the sub-trajectories follow a strict pattern. Our assumption is based on some fundamental principles of

[2]Image Courtsey : www.google.com

rigid body mechanics and *handwriting*. Unless interrupted by some external force, the m-segments are either linear or circular or parabolic in shape. If we consider each m-segment as an arc of a circle, then

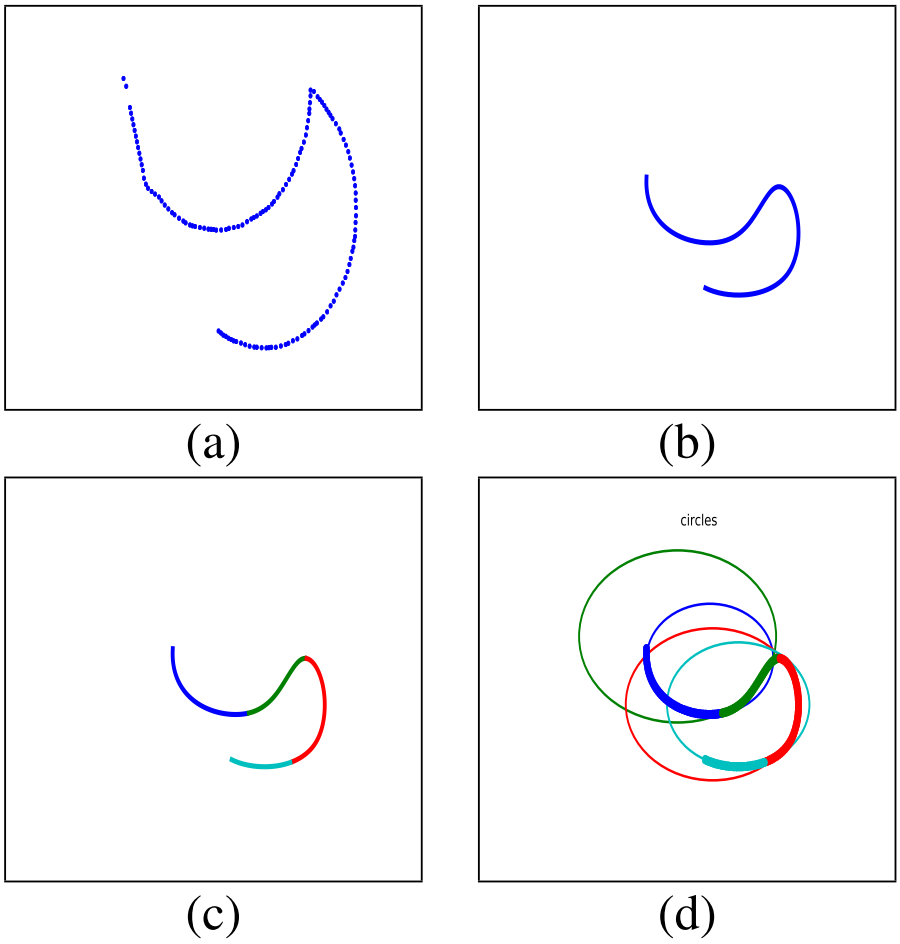


Figure 3.2 A sample motion with the corresponding m-segments: (a) Original (b) Smooth and Normalized (c) m-segments (d) Circle-Based Representation

the corresponding *centre* and *radius* can be used to represent the arcs.

Each m-segment has the form: $S = \{[x_i, y_i] \mid i = 1, 2, \ldots n\}$. A circle is fit by minimizing the squared radial deviations, expressed as

$$J = \min_{x_0, y_0, r} \sum_{i}^{n} x_i^2 + y_i^2 - 2x_0 x_i - 2y_0 y_i + x_0^2 + y_0^2 + r^2 \tag{3.1}$$

where $[x_0, y_0]$ and r is the centre and radius of the circle, respectively.

Let, $-2x_0 = a_1$,$-2y_0 = a_2$ and $x_0^2 + y_0^2 + r^2 = a_3$. Then Equation 3.1 can be expressed in matrix form as

$$(X\ Y\ 1)(a_1\ a_2\ a_3)^T = -(X * X\ +\ Y * Y) \tag{3.2}$$

where $*$ is the Hadamard product of two matrices and $X^T = [x_1\ x_2\ \ldots\ x_n]$ and $Y^T = [y_1\ y_2\ \ldots\ y_n]$ and $[x_i, y_i]\ \in\ S$. Solving Equation 3.2, we get

$$(a_1\ a_2\ a_3) = -(X\ Y\ 1)^+ \times (X * X\ +\ Y * Y) \tag{3.3}$$

where P^+ denotes the *Moore Penrose Pseudo-Inverse* of matrix P. Thus from the solution of equation 3.2, we can find our desired circle parameters as

$$\begin{aligned} x_0 &= -\frac{a_1}{2} \\ y_0 &= -\frac{a_2}{2} \\ r &= \sqrt{\frac{a_1^2 + a_2^2}{4} - a_3} \end{aligned} \tag{3.4}$$

It is interesting to note that small, medium and large values for radius indicate circular, parabolic and linear motion respectively and giving us a qualitative understanding of the m-segment. The radius is mapped to $[0, 1]$ using a hyperbolic tan function. The notion of approximate position of segment S can be represented using the mean $[x_\mu, y_\mu]$, which was experimentally found to be less variant than the centre of the circle. Subsequently, S is represented as,

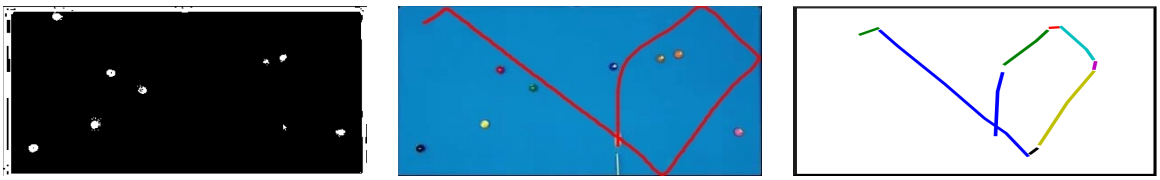
$$S = (\ x_\mu,\ y_\mu,\ r,\ e,\ s\ )$$

where e is the slope of the best line fitting to the points in each segment. The parameters are estimated using a least squares approximation. The slope has been experimentally quantized to 8 directions (N,S,E,W,NW,NE,SW,SE) to minimize the perceptual variability. s represents the normalized length of the arc.

3.2.2 Original Trajectory

In this work, our focus has been more on modelling user perception rather than trajectory extraction from videos. So we have used a set of 100 artificially simulated simple videos (Section 5.3.1) where the background is static and there are only a few objects. But the motion paths have been made very complex. We have also collected a set of 100 Pool Shot videos from many international matches, uploaded on YouTube. Motion trajectories were extracted from the real dataset in the following manner (Figure 3.3).

Firstly, we have denoised each frame using a median filter. Then we have extracted only the board region from each frame using a mask selected from the average frame. Next, we have done background extraction using a thresholding based method. The moving components were tracked in the video using a Gaussian Mixture Model [30] over the binarized frames. The trajectories were extracted from the video using a Kalman Filter [4]. Multiple object tracking was implemented using a variant of Hungarian Algorithm [43]. The raw trajectories were pre-processed and the qualitative features for a segment *i.e* $S =< x_\mu,\ y_\mu,\ r,\ m,\ s >$ are obtained in a similar manner, as discussed in the previous sections.



(a) (b) (c)

Figure 3.3 (a) A Background Extracted Frame (b) Trajectory Extracted from a Video (c) Trajectory after smoothing and segmentation

The m-segments extracted from all the trajectories in the database are clustered using the $k - means$ algorithm to obtain a codebook containing k cluster centers.

The complete trajectory is modelled as a histogram of m-segments, with each bin of the histogram corresponding to each cluster center in the codebook. So, for each trajectory in the database we create a *bag-of-motions* representation, similar to the *bag-of-visual words* representation used to represent images with SIFT features [48]. This same codebook is used to generate the bag-of-motions representation for the query as well.

3.2.3 Order, Direction and Scale

Histogram based features proposed until now do not capture the important motion properties: *Temporal order, Direction and Scale*. As mentioned in Chapter 2, the order in which the sub-trajectories appear, plays a very important role in representing the motion. But it is difficult to compare two motion trajectories having unequal length. *Dynamic Time Warping* (DTW) [55] is an efficient tool which is used to compare time-series data having unequal length. We use DTW to compare the motion trajectories.

We find the change of direction across time. First we re-sample the trajectory to remove the variability in the density of points due to varying speed of the hand movements (Figure 3.2.3). Next, we divide a trajectory into equipoint segments (segments having equal number of points), and the distribution of direction across time has been approximated by fitting a line to each of the equipoint segments. The angle made by each equipoint segment with the horizontal X-axis has been mapped to $[-1, 1]$ using a $sine$ function. (Figure 3.4(c)).

Similarly, scale of motion is an important factor for identifying motion trajectories. The scale is defined as the change of the current position with respect to the starting point. For example, a counter-clockwise spiral motion can be converging or diverging. Shape histogram or change of direction across time cannot differentiate between such motions. But the change of scale across time distinguishes the two.

Summarizing the previous discussions, it can be said that in our query we have conveyed four fold information. Firstly, we have conveyed shape information and the approximate position of each segment in the trajectory in the *bag-of-motions* representation. Secondly, we have features that represent the change of direction and scale of the trajectory across time. Notably, none of the features we derive here use any

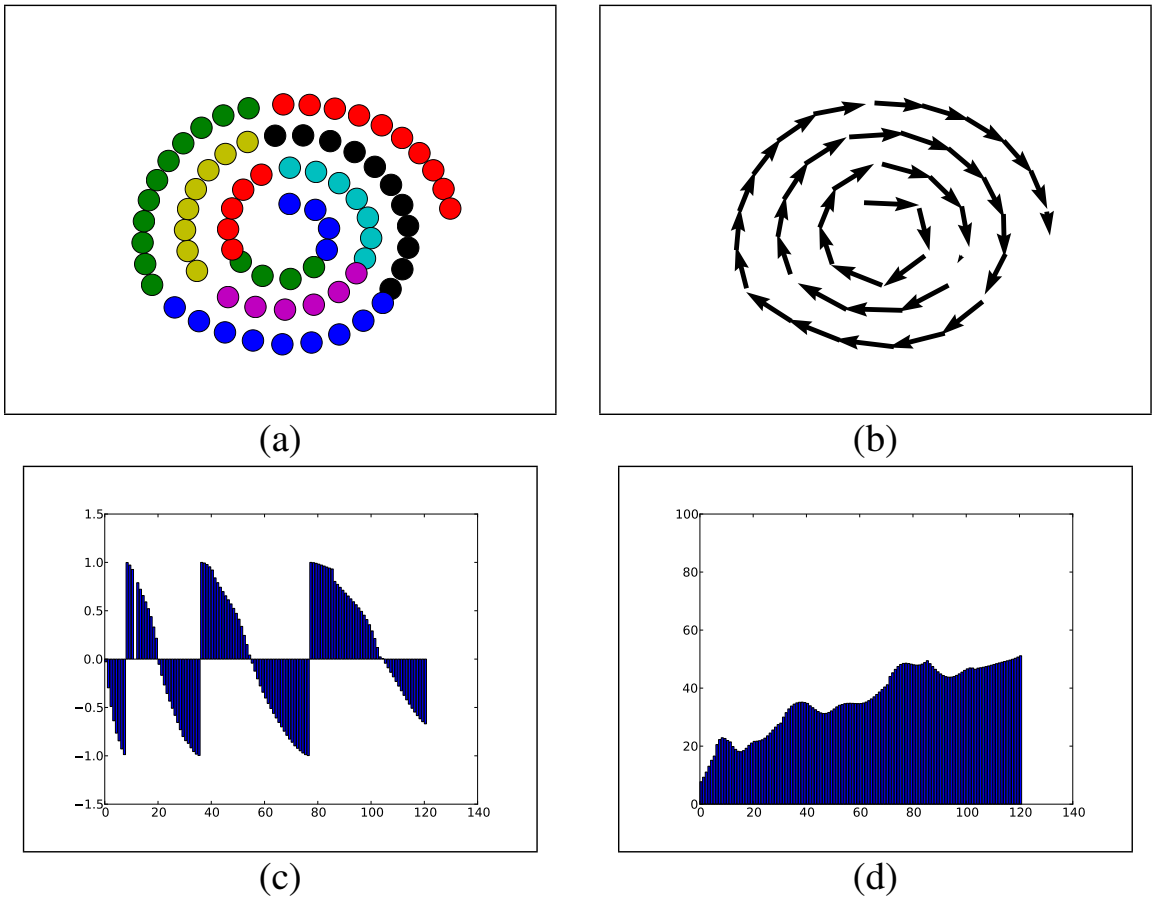


Figure 3.4 A Spiral Motion from our synthetic dataset (a) Points sampled equidistantly in each segment (b) Directions tracked for each equipoint segment (c) Temporal Change of Direction (d) Temporal Change of scale

absolute information. All the features approximate the overall motion, which is intended, to remove the perceptual variability among different users. These features give us a *qualitative* understanding of the motion trajectory.

3.3 Retrieval

In this section, we propose our multilevel search strategy. Once a query is given, four sets of features described in the previous sections are extracted and the query is passed through a cascaded filter having four stages.

In the first stage the query histogram is matched with the database of *bag-of-motions*. Each sample X_i in the database is assigned a score $\alpha_i^{(1)}$. This level of filtering finds the trajectories that have similar sub-trajectories. At the next level, a Dynamic Time Warping (DTW) [55] match is performed between the query $Q = [S_1, S_2, \ldots S_M]$ and each sample in database $T_i = [S_1, S_2, \ldots S_N]$, where each S_i is

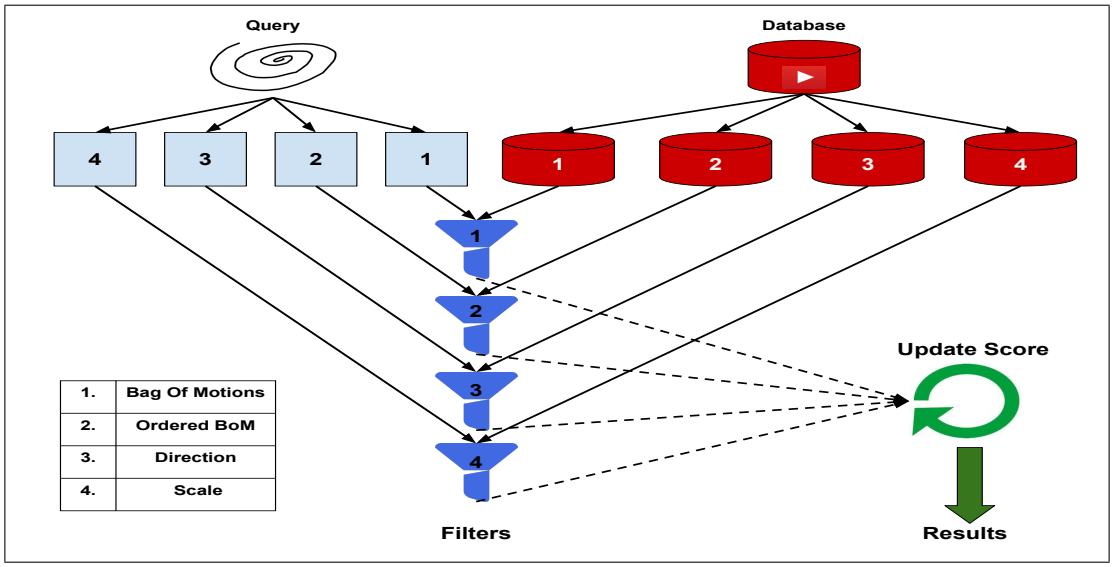


Figure 3.5 Multilevel Retrieval Strategy : The query and original videos in the database (top-left and top-right) are processed and four sets of features are derived in each case. There are four different levels of filtering (four blocks vertically arranged at the center). The functionality of each filter has been shown in the table (bottom-left). After each level of filtering, the score is updated by the score update module (bottom-right). The videos are retrieved based on the final score.

the feature derived in Section 3.2.1 and 3.2.2. A new score $\alpha_i^{(2)}$ is obtained at this level. Apart from preserving the order, the match at this level also facilitates partial trajectory match. At the next two levels DTW match is carried out between the features derived in Section 3.2.3 and scores $\alpha_i^{(3)}$ and $\alpha_i^{(4)}$ are obtained. The final score $\alpha = \sum \alpha^{(i)}$ is calculated. Each of the scores are calculated as a function of the distance of the query from each sample, computed at each level $i.e\ \alpha^{(i)} = f(d^{(i)})$. The value of α is updated after every stage in a cumulative fashion. The final results are retrieved based on the value of α after the fourth stage. The algorithm has been explained using a block diagram in Figure 3.5.

Two important aspects should be considered here. Firstly, although multiple trajectories were extracted from a video, our current system uses only one trajectory sketch to search for the video. Chance of choosing a particular trajectory depends entirely on the user. But in our database we store all the trajectories. However, extending the system to allow a user to specify multiple trajectories in a video can further refine the results. We have not implemented this in our query interface as of now, but we intend to do this in our future work.

Secondly, our current system does not behave like a *regular* cascade and it differs from traditional cascaded systems. Currently, the search-space is not reduced at each level in our algorithm as it happens in cascaded detectors such as [80]. But please note that with successive stages, our features and matching go from weak and efficient to discerning and complex. We can discard samples with the lowest matching scores at each stage, making it a regular cascade.

3.4 Dataset

We have synthesized a dataset, which contains 100 videos of one, two and three body motions. The videos are divided into five sets : (a) a set with linear motions resembling Pool shots (b) a set with mixture of linear and exponential curves as trajectories that resemble moving cars and (c) a third set with respective motions like circular (clockwise and counter clockwise), sinusoidal and spiral. (d) a set of motions that resemble typical motions like sea-saw ride, people jumping side by side, divers diving *etc*. (e) where the motion trajectories are regular geometric shapes like square and triangle. It was found that in animation videos, most of the motion trajectories have regular geometric shapes. The synthetic dataset was created keeping in mind all the different kinds of videos which later can be explored with this kind of retrieval strategy.

We also tested our method on real pool videos. Full match pool videos were segmented into shots using a histogram based approach [85]. Then a dataset of 100 clean videos having a top view of the pool board was created. Each video was shown to different users and they were asked to group the videos which they found perceptually similar.

It was difficult to divide the pool shots into a specific number of classes. To achieve this, we asked multiple users to cluster/group the videos based on their similarity. Pairs of videos within a group were assigned a high similarity score and in different group were given a lower similarity. The similarity scores from multiple users were integrated into a single score matrix and an automatic clustering was performed to arrive at the final class divisions. It was found that the users could identify five different groups from the dataset. The distinction between different classes were done mainly based on shape, direction and position of the shots. The dataset is available for download and can be found online on our website[3].

For collecting query sketches, we sampled 50 videos from each of the datasets and then showed 20 videos (10 from each set) to a user. The user was asked to watch the videos carefully and then sketch two most salient motions that he/she could remember from each video. The (x, y, t) coordinates of the sketches were recorded. The experiment was carried with 25 users. Each video had 5 sample queries.

3.5 Experiments and Results

We have evaluated the effectiveness of our representation using three different standard evaluation metrics as follows.

Precision Recall : The PR curves generated from our experiments with real and synthetic datasets are shown in Figures 3.6 (a) and 3.6 (b) respectively. It can be seen that the area under the curve gradually increases as the query score gets updated with each filter. Simple DTW of the points performs worst. Only Bag-of-Motions based nearest neighbour search performs poorly (red curve). But the results improve significantly as soon as the temporal information is also used and the scores are updated in the

[3] http://cvit.iiit.ac.in/projects/sketchbasedretrieval/

next filter (blue curve). The precision is further tuned using the next filters and the best curve is obtained after the final stage of filtering is completed. There is a significant improvement after the second level than in third and fourth levels. This is because the order in which the sub-trajectories appear play a vital role in distinguishing motion trajectories. We believe the improvement reflects the importance of temporal ordering in modelling long trajectories. Also precision-recall curves in case of the synthetic dataset are better than those in case of the real dataset. This is mainly because, the synthetic dataset has more inter-class variance. The motions have fundamental differences with respect to shape and spatio-temporal properties. But in case of Pool Videos, the trajectories are mostly linear (except the trick shots) and have very little inter-class variance.

Mean Reciprocal Rank : We mentioned in Section 2.4 that our retrieval strategy is intended to find the exact match instead of a class of matches. We found Mean Reciprocal Rank as a good measure to test such an algorithm. The multiplicative inverse of the rank of the first correct answer in a set of retrieval results is obtained. The mean reciprocal rank is the average of the reciprocal ranks of results for a sample of queries.

$$MRR = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{\text{rank}_i}$$

We calculated the MRR with the same set of queries. It was found to be 0.5 and 0.4 on the real and synthetic videos respectively. It indicates that with this approach, the chance of finding an exact match within the top few results is high, which is desirable in our case. Figure 3.7 demonstrates the histogram of reciprocal ranks of all the queries.

Accuracy : We defined accuracy in the following manner. A search was considered successful, if the exact match appeared in the top k results. Figure 3.8 demonstrates the accuracy values on the real and synthetic dataset for k values ranging from 1 to 15. Figure 3.9 shows an example query with the top k retrieved videos.

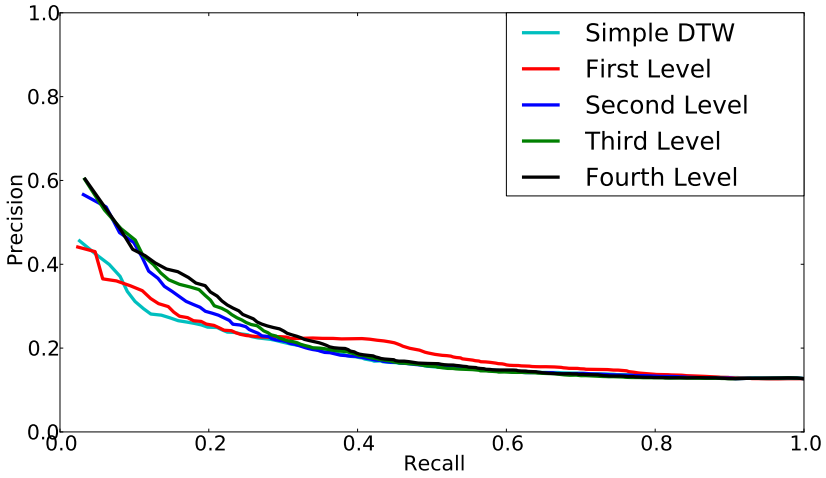
3.6 Summary

In this chapter, we addressed a lesser explored aspect of an extensively explored problem of motion trajectory based video retrieval. One of the limitations of our algorithm is that it relies on strong foreground segmentation and trajectory extraction algorithms which are themselves unsolved problems in complicated videos. Challenges like dynamic background, camera shakes, shadow, camouflage *etc.* [5, 9] are active areas of research in Computer Vision. Moreover, this method cannot be used for retrieving videos where motion is not the most salient feature. The problem can also become very ill posed and difficult when the query is so complicated that it cannot be described in any unimodal format.

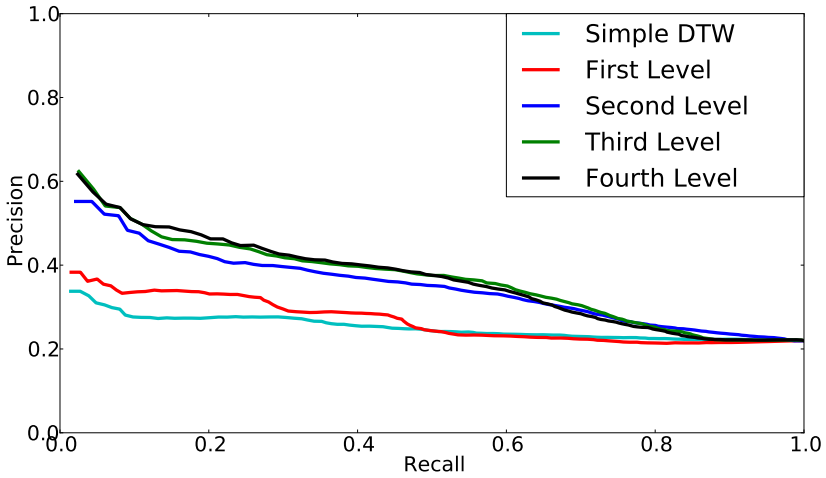
However, it is different from most of the existing sketch based systems because the query is unconstrained. No initial frame is supplied to the user. We have proposed a new representation for the trajectories of objects in videos and the sketch-based query. The features, which depend on perceptual similarity, are qualitative in nature and robust to user-level variations. Moreover, instead of using

only spatial or temporal features the technique uses a combination of those by implementing a novel cumulative scoring mechanism.

A better understanding of the motion perception in humans will enable us to develop more robust features. But, our method can be applied on top of any trajectory estimation method. Also, it can be refined further with object level features like shape, colour and size. Also the fact, that we have shown that our method can be used on real Pool videos with satisfactory results, gives us hope that this approach can be extended to more complicated videos and used to develop accurate and robust multi-modal systems in future.

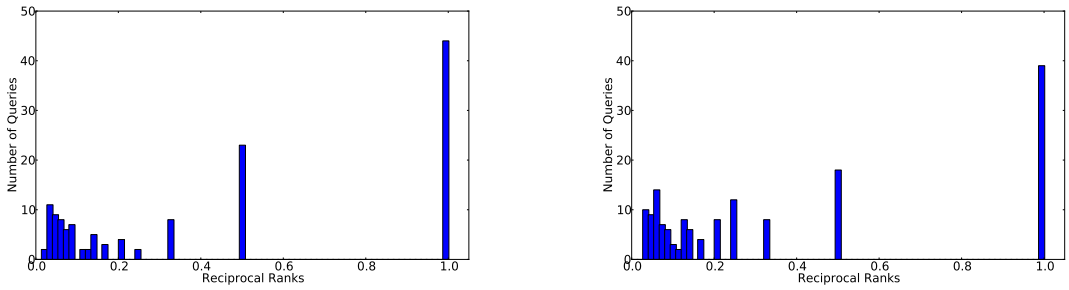


(a) Pool Videos dataset



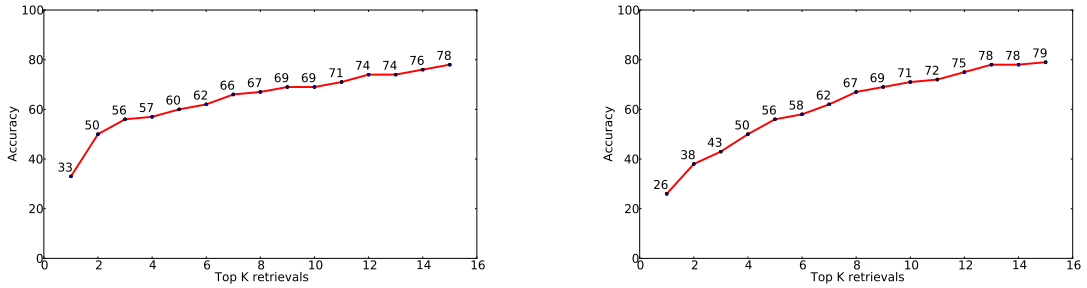
(b) Synthetic Motion dataset

Figure 3.6 Precision Recall Curves (view in colour)



(a)Pool Videos dataset (b) Synthetic Motion dataset

Figure 3.7 Reciprocal Ranks : A high value near one indicates that most of the queries retrieved the exact match as the first result. A value of 0.5 indicates that the second result was the correct match and so on.



(a) Pool Videos dataset (b) Synthetic Motion dataset

Figure 3.8 Accuracy at different top K retrievals. Exact accuracy values at different k are annotated on the curve. It can be observed that the accuracy reaches 70% within top-10 results for both the real and synthetic dataset.

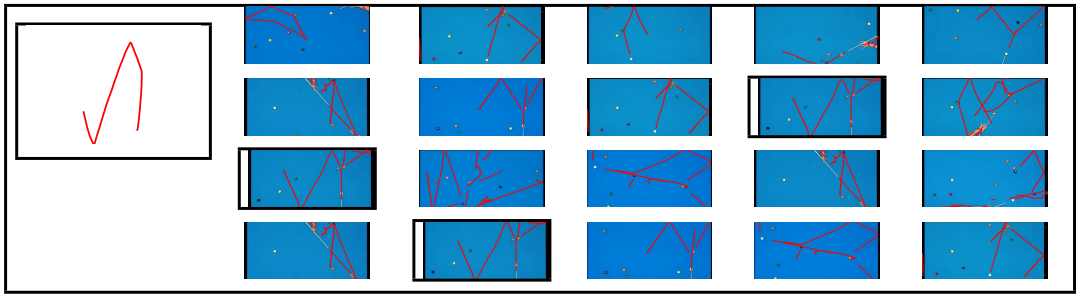


Figure 3.9 The figure on the left is the query. On the right, the four rows correspond to the four stages of our filter. Elements in each row correspond to the top 5 results at each iteration, after the score is updated. The exact match is highlighted using a box. At the first level, the exact match is not found in top 5. But it appears after stage 2 and maintains its position within the top 5 results till stage 4

Chapter 4

Sketch-based Image Retrieval

4.1 Introduction

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.

4.2 Multimodality

In this section, we formulate our problem as a cross-modal retrieval task . In Section 4.2.1, we state the problem formally. In Section 4.2.2, we briefly explain CCA and it's modified version Cluster-CCA.

4.2.1 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 [34], 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,

$$\begin{aligned} \rho &= \max_{W_x, W_y} corr(P_x, P_y) \\ &= \max_{W_x, W_y} \frac{\langle P_x, P_y \rangle}{\|P_x\| \|P_y\|} \end{aligned} \tag{4.1}$$

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 [62], which establishes a one to one correspondence between all pairs of data points in a given class. We explain it in detail in Section 4.2.2.

4.2.2 Cluster-CCA

Rasiwasia *et al.* [62] introduced and used Cluster-CCA for cross-modal retrieval tasks with image and text as two modalities. As derived in [34], Equation 4.1, reduces to the following form

$$\rho = \max_{W_x, W_y} \frac{W_x' \text{Cov}_{xy} W_y}{\sqrt{W_x' \text{Cov}_{xx} W_x} \sqrt{W_y' \text{Cov}_{yy} W_y}} \tag{4.2}$$

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

$$Cov = \mathbb{E}\left[\begin{pmatrix} A_x \\ A_y \end{pmatrix} \begin{pmatrix} A_x \\ A_y \end{pmatrix}' \right] = \begin{bmatrix} Cov_{xx} & Cov_{xy} \\ Cov_{yx} & Cov_{yy} \end{bmatrix} \tag{4.3}$$

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 4.2.1, 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 4.2 can be modified to the following form,

$$\rho = \max_{W_I, W_S} \frac{W_I' \Sigma_{IS} W_S}{\sqrt{W_I' \Sigma_{II} W_I} \sqrt{W_S' \Sigma_{SS} W_S}} \tag{4.4}$$

where the new covariance matrices are defined as follows,

$$\Sigma_{IS} = \frac{1}{M} \sum_{c=1}^{C} \sum_{j=1}^{|I^c|} \sum_{k=1}^{|S^c|} I_j^c S_k^{c'} \tag{4.5}$$

$$\Sigma_{II} = \frac{1}{M} \sum_{c=1}^{C} \sum_{j=1}^{|I^c|} |S^c| I_j^c I_j^{c'} \tag{4.6}$$

$$\Sigma_{SS} = \frac{1}{M} \sum_{c=1}^{C} \sum_{k=1}^{|S^c|} |I^c| S_k^c S_k^{c'} \tag{4.7}$$

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 4.4 can be formulated and solved as an eigen value problem as in [34].

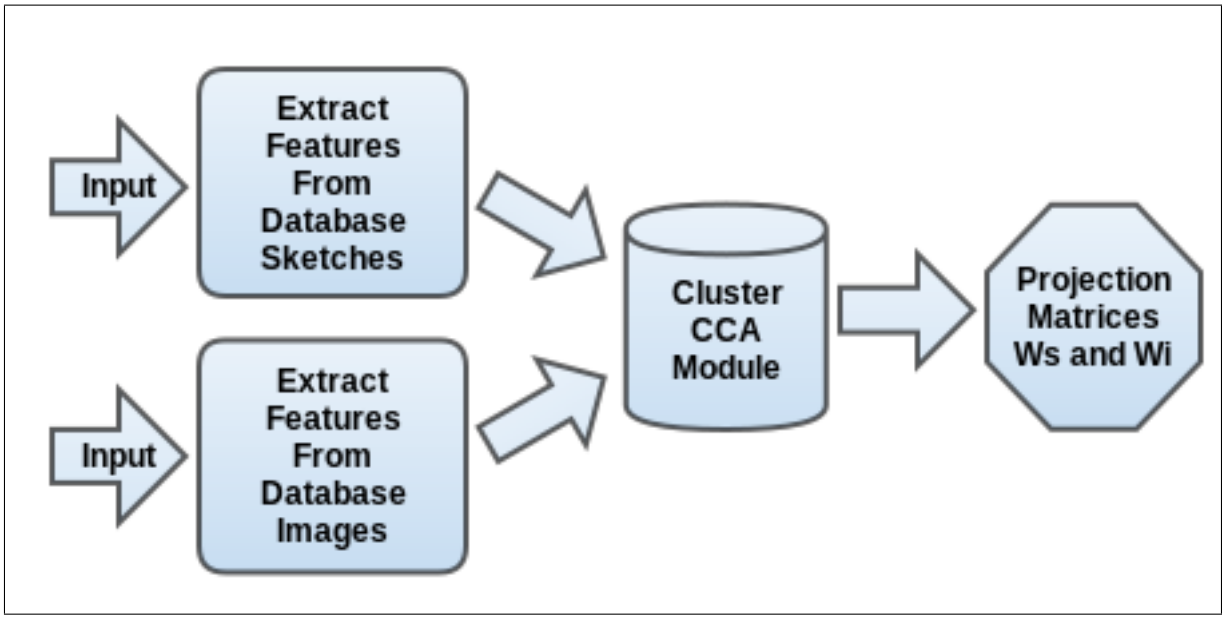
To summarize, in this section we explained, how a modified version of the standard CCA can be used to create a one-to-one correspondence between samples belonging to the same category but to two different modalities, image and sketch. We projected each modality, having different dimensions, into two 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.

4.3 Experiments

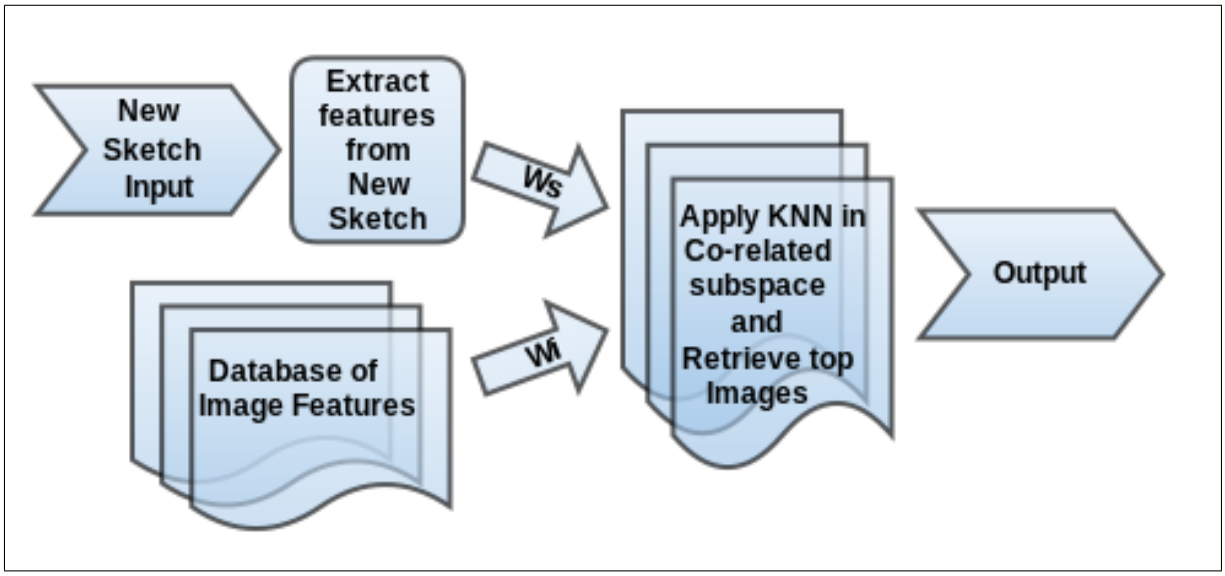
In this section, we quantitatively evaluate the performance of our proposed approach. We use three datasets, PASCAL VOC 2007 [27] and CALTECH-256 [31] for images and TU-BERLIN [24] dataset for sketches. We divide the TU-BERLIN dataset into training and testing sets and use the training set to create 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 4.1. We use PR curves and MAP values as quantitative measures.

4.3.1 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.* [82] and 68% by Rosalia *et al.* [66]. **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.



(a)



(b)

Figure 4.1 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 1 : Summary of Features

Feature	Dimension	Source
CALTECH - SIFT	1000	Vl-Feat. [79]
CALTECH - HOG	20000	Vl-Feat [79]
CALTECH - CNN	4096	Krizhevsky *et al.* [42]
PASCAL - SIFT	1000	Guillaumin *et al.* in [32]
PASCAL - HOG	20000	Vl-Feat [79]
PASCAL - CNN	4096	Krizhevsky *et al.* [42]
TU-BERLIN - SIFT-Like	501	Eitz *et al.* in [24]
TU-BERLIN - HOG	20000	Vl-Feat [79]
TU-BERLIN - Fisher	250000	Rosalia *et al.* [66]
TU-BERLIN - CNN	4096	Yang *et al.* [82]

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

4.3.2 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[1]

4.3.3 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

[1] http://cvit.iiit.ac.in/projects/sketchbasedretrieval/

Table 3: Performance improvement in mAP values

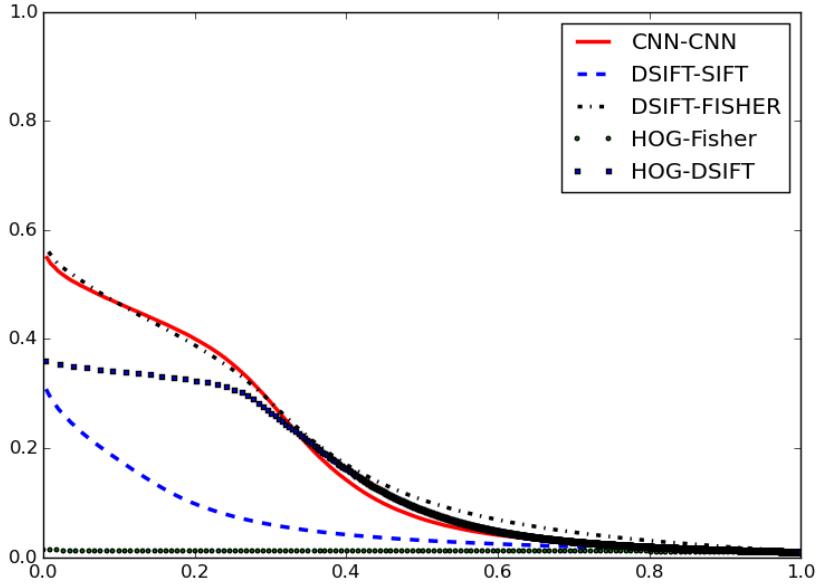
Dataset	Features	Before CCA	After CCA
Caltech	SIFT-Fisher	0.01	**0.20**
Caltech	CNN-CNN	0.01	**0.20**
Pascal	HOG-SIFT	0.01	**0.18**
Pascal	SIFT-SIFT	0.06	**0.13**

we believe more sophisticated feature learning techniques could alleviate this problem.
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.
However, please note that SIFT-Fisher and HOG-SIFT feature combinations could not be compared directly because of unequal dimensions. Hence, we levelled the dimensions by doing a principal component analysis on the larger of the two dimensions.

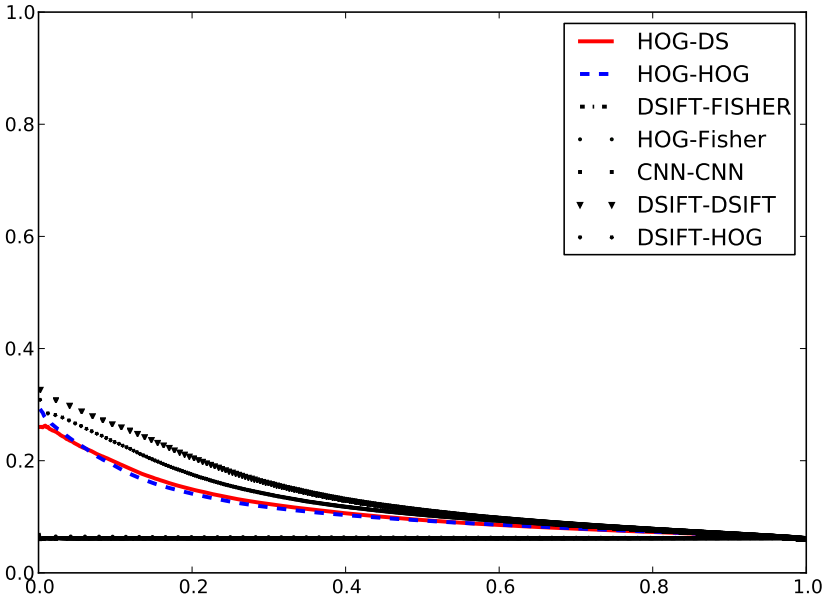
Precision Recall: The PR curves in Figure 3 are suggestive of the greater complexity of the PASCAL dataset in comparison with Caltech. The poor PR curves indicate that the correspondence doesn't work well with complex images. In Figure 4.3, we provide results from two example queries for airplane and backpack.

4.4 Summary

In this chapter, 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.



(a) Caltech



(b)Pascal VOC 2007

Figure 4.2 Precison Recall Curves

(a) Success

(b) Failure

Figure 4.3 (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.

This is an interesting problem and there are a lot of areas which can be explored in future.We realised that an interesting area might be finding representations which are generic enough to retrieve unknown classes. We explore that aspect of the problem in the next chapter.

Chapter 5

Using Deep Features for Zero-Shot Retrieval

5.1 Introduction

Convolutional Neural Networks have recently been successfully used for sketch recognition. However, the state of the art features in sketch-analysis are tuned towards classification and perform poorly during retrieval. They also fail to capture the semantic relationship among the different categories, which is essential for a new class retrieval.

We propose a novel feature representation for sketches called Deep Features for Semantic Retrieval (DFSR), which captures the semantic relationship among different object categories. We introduce a new loss function for the network that penalizes a distant miss more than a close miss, based on the magnitude of the error. Hence, by differentiating between the classes using semantic similarity, we encode the inter-class semantic relationship. We perform multiple experiments to show that the performance of our features is comparable with the state-of-the-art and significantly better in uni-modal retrieval, cross-modal retrieval and zero-shot retrieval.

5.2 Features

The major components of zero-shot learning is a set of *known* classes $\mathcal{Y}$, which is used to train a classifier $\mathcal{H}$ and a set of *unknown* classes $\mathcal{Z}$, which are not a part of the training. While training we incorporate information about $\mathcal{Z}$ into $\mathcal{H}$ using a semantic knowledge database $\mathcal{K}$. The modality of $\mathcal{K}$ is independent of $\mathcal{Y}$ or $\mathcal{Z}$ and it has a representation which captures the semantic similarity between all classes. It acts as a knowledge database which contains information about both the known and unknown classes and encodes that information within the features during training. Henceforth, known and unknown samples in the semantic feature space $\mathcal{F}$ will be referred to as $y_{\mathcal{F}}$ and $z_{\mathcal{F}}$ respectively.

In this section we explain how we formulate our problem as a zero-shot retrieval problem and propose features, which we call Deep Features for Semantic Retrieval (DFSR). We argue that $\mathcal{F}$ is an intermediate representation between $\mathcal{K}$ and $\mathcal{Y}$ or $\mathcal{Z}$. Samples of the same categories, from the two different

modalities, images and sketches are closer in this semantic space than in their respective original space. We probe the validity of our claim by multiple experiments in Section 5.3.

5.2.1 Word2Vec

We use the vector space model proposed by Mikolov *et al.* [53], in which semantically similar words, mined from a very large text corpora, are mapped to nearby points in a continuous vector space hereafter referred to as *Word2Vec* space. Conceptually, it is based on skip-gram model, which tries to predict contextual words from a sentence or document. Mathematically it is expressed as follows.
Given a sequence of words $w_1, w_2, w_3 \dots w_T$, the skip-gram model tries to maximize the average log probability

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c, c\neq 0} \log\ p(w_{t+j}|w_t) \tag{5.1}$$

where c is the size of the training context.

Interestingly, the learned vectors encode certain linguistic patterns and semantic similarities between words. For example, as pointed out in [53] vec(Madrid) $-$ vec(Spain) $+$ vec(France) is closer to vec(Paris) than any other word in the Word2Vec space.
For our experiments, we use Caltech-256 image and TU-Berlin sketch dataset, each containing 256 and 250 categories of objects, respectively. We calculate a distance matrix $\mathcal{D}$ where

$$\mathcal{D}_{i,j} = ||C_i - C_j||^2$$

where $C_k \in C$ and where C is the union of all class-names in Word2Vec space.
We observed from $\mathcal{D}$ that classes which were semantically similar like *apples* and *mangoes*, were close in the Word2Vec space, whereas semantically exclusive classes like *apples* and *accordion* were distant. Since the Word2Vec space contains representations for millions of words, for even a very large dataset, it is a good choice as the semantic knowledge database $\mathcal{K}$ for zero-shot learning.

5.2.2 Intermediate Representation

Now that we have $\mathcal{K}$, in this section, we explain how we derive $\mathcal{Y}$(the set of known classes), $\mathcal{Z}$(the set of unknown classes) and $\mathcal{F}$(the semantic feature space).
We randomly split Caltech-256 and TU-Berlin dataset into five parts and iteratively, use one part as $\mathcal{Z}$ (set of unknown classes) and the other four parts as $\mathcal{Y}$ (set of known classes). We elaborate more on our experimental set-up in Section 5.3. We train a Convolutional Neural Network on the training data using the architecture provided by Yang *et al.* [83] for TU-Berlin and the one provided by Chatfield *et al.* [14] for the Caltech-256 dataset. However, the features extracted using these networks are standard CNN features and should be mapped to the semantic feature space $\mathcal{F}$.

In order to embed the semantic relationship among different categories into the features, we adopt the strategy of Socher *et al.* in [72]. We replace the final soft-max layer of the CNN as follows.

$$J(\theta) = \sum_{y \in \mathcal{Y}} \sum_{x^i \in \mathcal{X}_\mathcal{Y}} ||w_y - g(x^i)||^2 \tag{5.2}$$

where $\mathcal{X}_\mathcal{Y}$ is the set of samples from $\mathcal{Y}$ and $g(.)$ stands for the convolution which maps the samples $x \in \mathcal{X}_\mathcal{Y}$ to the d-dimensional feature space. w_y is a d-dimensional word vector for category $y \in \mathcal{Y}$.
We discussed in the previous section that in Word2vec space, classes which were semantically closer were mapped close to each other in $\mathcal{F}$. Thus from Equation 5.2 we can deduce that the function $J(\theta)$ penalizes the classes unequally, which is desirable. It's penalty on a close miss is less whereas that for a distant miss is greater. This incorporates a semantic loss for the misclassified samples as compared to the softmax loss, which penalizes any misclassification equally, irrespective of how bad the misclassification was.
For TU-BERLIN and CALTECH-256 datasets we trained the network with initializations from [82] and [14] respectively, using the objective function from Equation 5.2. We call these features, projected on $\mathcal{F}$ as *Deep Features for Semantic Retrieval* (DFSR).

5.3 Experiments

Although the key incentive for proposing the DFSR features is its ability to perform zero-shot learning, we conduct several experiments to investigate it's efficiency in multiple tasks. We perform four different sets of experiments on Classification (Section 5.3.2), uni-modal retrieval (Section 5.3.3), cross-modal retrieval (Section 5.3.4) and zero-shot learning (Section 5.3.5). We explain the datasets used (Section 5.3.1) and report and analyse the results in the following subsections.

5.3.1 Datasets

We use two datasets for our experiments - TU-BERLIN [24] and CALTECH-256 [31]. TU-BERLIN dataset is a free-hand sketch dataset and contains 250 categories of objects, with 80 samples per category. The best recognition accuracies reported till date is 74.9% by Yang *et al.* [82] and 68% by Schneider *et al.* [66].
CALTECH-256 is an image dataset which contains 256 object categories, containing $30,607$ samples. However, the number of classes which are common to CALTECH and TU-BERLIN is 105, totalling to about $14,231$ images.
In our experiments, during classification, in Section 5.3.2 , we use the entire TU-Berlin dataset and divide it into training and validation sets as done by Yang *et al.* [82]. In Section 5.3.3, where we perform uni-modal retrieval for sketches and images separately, we again use the entire TU-BERLIN and CALTECH-256 datasets. However, to keep the training and validation splits consistent across the two datasets, we sample first 80 images from the CALTECH-256. In Section 5.3.4, however, we select 105

classes which are common to both CALTECH-256 and TU-BERLIN and use them for cross modal retrieval. Please note that we do not train a new network for the cross-modal retrieval but use the same features extracted in, Section 5.3.2 and 5.3.4.
However, in Section 5.3.5.1 and 5.3.5.2, for zero-shot retrieval, we split the data randomly into five partitions and then iteratively use each partition as the set of unknown classes. We train different networks with the other four parts at each iteration and test the network on the zero-shot partition.
The novelty in our implementation is that we achieve 32 times compression in the size of training data, by binarizing the sketches, yet retaining the performance with a drop of only 1% accuracy. Also, along the lines of Ravi *et al.* [65], we use morphological operations to retain the finer details of the sketches in deeper layers of the CNN.
We implement a single channel CNN, it's architecture motivated from [82] and use the output from the softmax layer. We compare DFSR features with this feature in our experiments in Section 5.3.2, Section 5.3.3, Section 5.3.4 and Section 5.3.5.

5.3.2 Classification

In recent years, several groups have worked in Sketch Recognition [83, 66] and have compared their performance with humans, who were able to recognize the TU-Berlin sketches with 73% accuracy [24].The best accuracy reported until now is 74.9% of Yang *et al.* [83], which beats humans. Experiments show that our features perform reasonably well using a much simpler CNN architecture.

Algorithm	Features	Classifier	Accuracy
	TU-Berlin Dataset		
Yang *et al.* [83]	CNN-Ensemble	SOFT-MAX	74.9%
Yang *et al.* [83]	CNN-Single	SOFT-MAX	72.6%
Schneider *et al.* [66]	FISHER	SVM	63.1%
Eitz *et al.* [24]	BOW	SVM	56%
Proposed	**DFSR**	RANDOM FOREST	70.22%

Table 5.1 Classification Results show that our features perform reasonably well almost at par with the state of the art methods.

Although in Table 5.1, it shows that features of Yang *et al.* perform better than ours, it is to be noted that using a much simpler architecture we achieve an accuracy close to theirs and better than the others [24, 66]. Moreover, while their soft-max loss tunes their features more towards classification, our Word2Vec features perform well in case of retrieval, considerably better than all others as shown in Section 5.3.3 and 5.3.4. We believe that this little difference in accuracy of classification is acceptable, considering it's superior performance in terms of cross-modal retrieval and zero shot learning. Additionally, as described in Section 5.3.1, we use a 32 times compressed data representation which enables

us to scale up our model to large databases.

5.3.3 Uni-Modal Retrieval

As discussed in Section 5.3.2, DFSR perform better in terms of retrieval than in classification. We confirm our observation, heuristically, on two datasets belonging to two different modalities - sketches and images. Precision-Recall curves are standard evaluation metrics for retrieval tasks and we use them to test our features in Figure 5.1. An improvement in retrieval performance than the features used by Yang *et al.* [82] is indicative of the fact that the features are better mapped in the Word2Vec space than in the original space. While Yang *et al.* uses an additional sophisticated classifier like Soft-Max or Support Vector Machines, to discriminate the classes, we use retrieval to evaluate the system, which is nothing but a K-NN classifier. From this observation, we can conclude that the DFSR features form better clusters in the Word2Vec space than the original CNN features.

5.3.4 Cross-Modal Retrieval

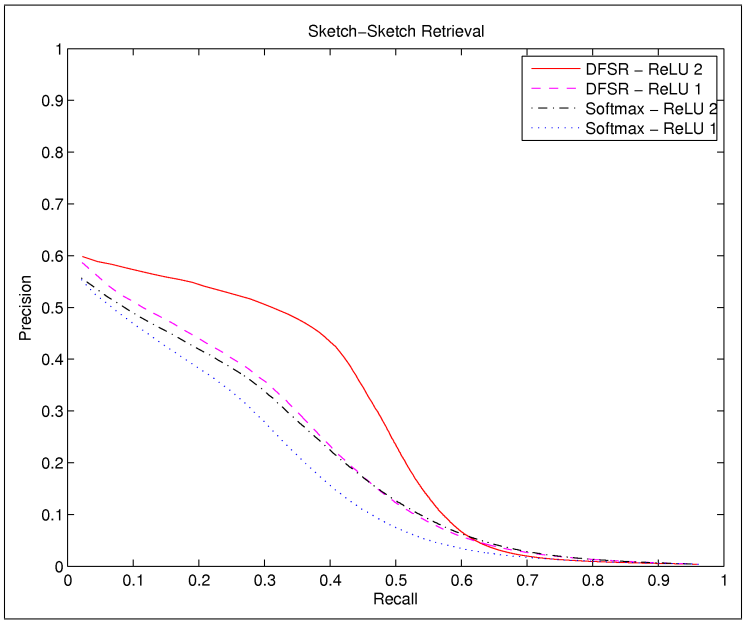
In this section, we show our results on cross-modal retrieval. We train two different models for images and sketches and extract features for each modality. Next, we map these features into a common lower dimensional subspace as explained in Section 4.2.2. Finally, we use the test samples from TU Berlin as a query and retrieve images from Caltech. We show the results in Figure 5.2. In Figure 5.3, we show a few sample queries and their corresponding retrieval output from our system.

5.3.5 Zero-Shot Learning

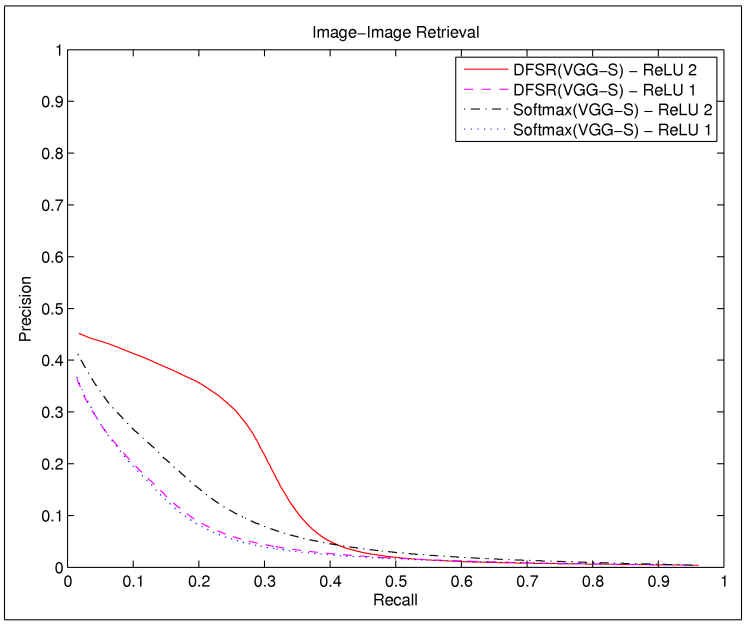
To evaluate the robustness of our system for zero-shot learning, we perform two different sets of experiments. First we ensure that the features perform well for our query modality and then we evaluate it in cross-domain retrieval.

5.3.5.1 Uni-Modal Retrieval

We split the TU-BERLIN dataset into five different parts and iteratively use each part as a set of unknown classes. The other four parts are used to train a network and the retrieval is carried out with queries from the zero-shot classes. Figure 5.4 displays the PR curves for the best and worst performing zero-shot partitions.



(a)



(b)

Figure 5.1 Uni-modal retrieval : (a) PR curve for sketch modality. It can be observed that DFSR outperform the softmax features extracted as mentioned in Section 5.3.1. The TU Berlin dataset contains 250 categories and 80 samples per class. The PR values are mean values across all classes and all the test queries. (b) PR curve for image modality. Caltech 256 dataset was used for the experiment. It consist of 256 object categories, each category containing variable number of samples. We randomly sampled 80 images from each class. The ImageNet model [14], by VGG group, but the soft-max layer was replaced as explained in Section 5.2.2. The PR curve for our features outperform all other state of the art features, in this case, as well

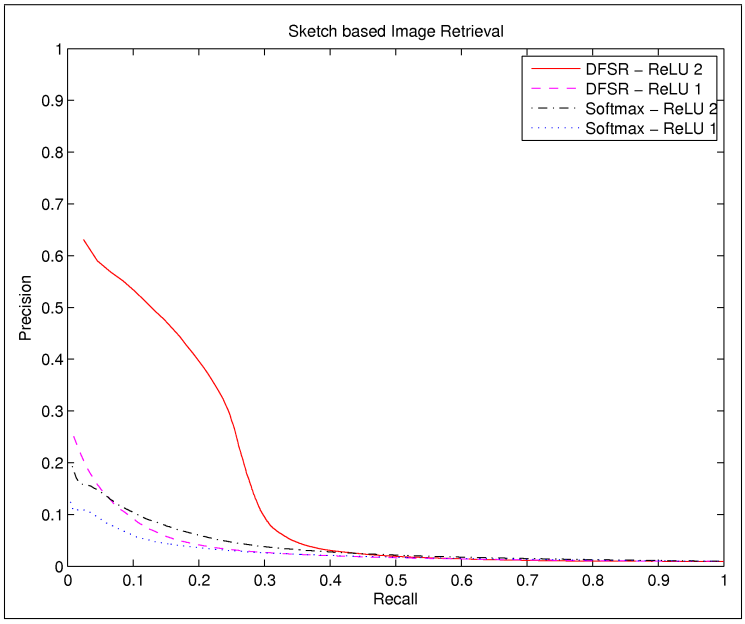


Figure 5.2 Cross-Modal Retrieval : PR curve for cross modal retrieval. The queries were selected from TU-BERLIN and the search was carried out in CALTECH dataset. Features were extracted from TU-BERLIN dataset as mentioned in Section 5.3.1 and from CALTECH dataset using the VGG network [14] with the soft-max layer was replaced as explained in Section 5.2.2. It can be observed from the PR curve that, our features outperform all other state of the art representations.

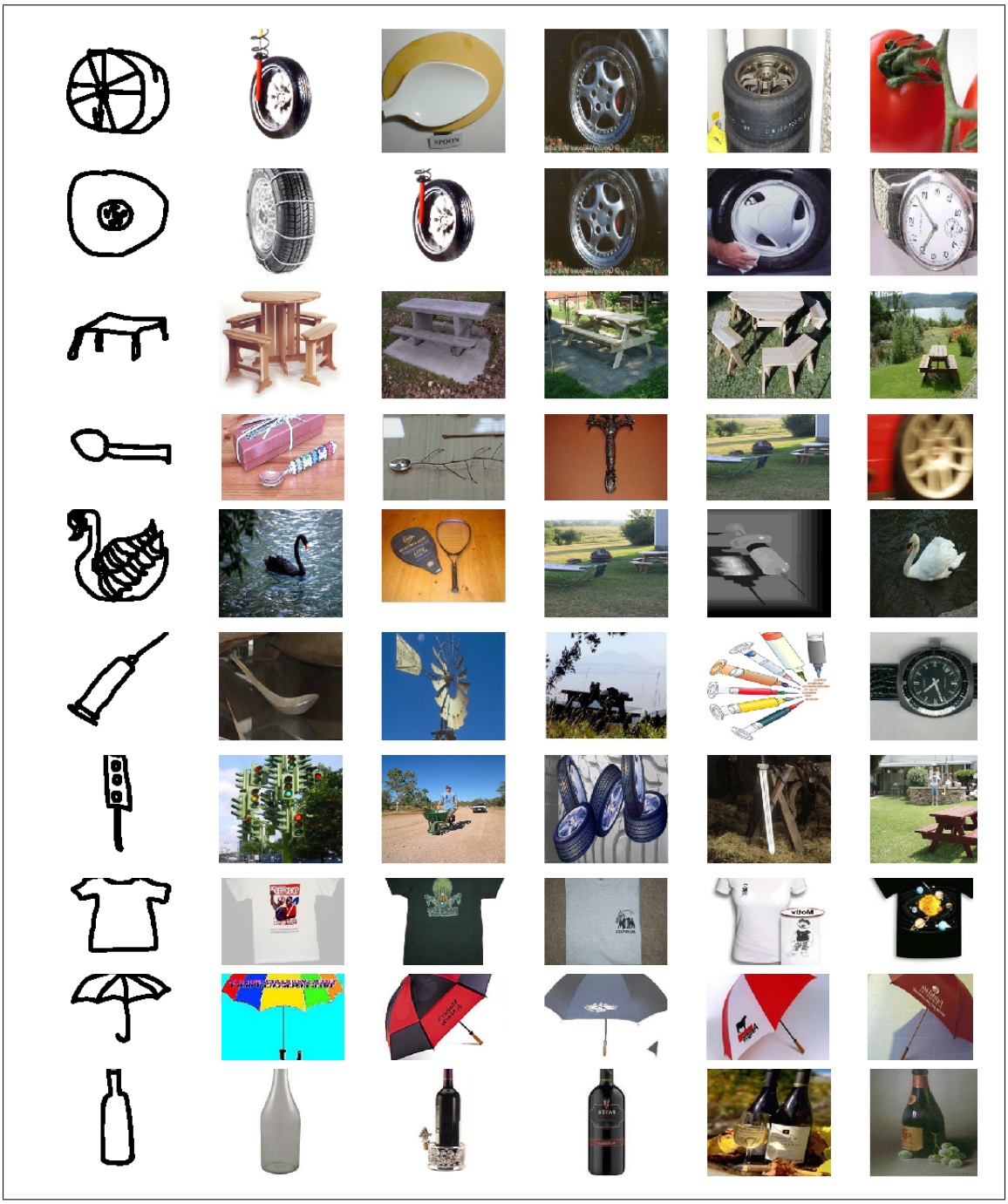
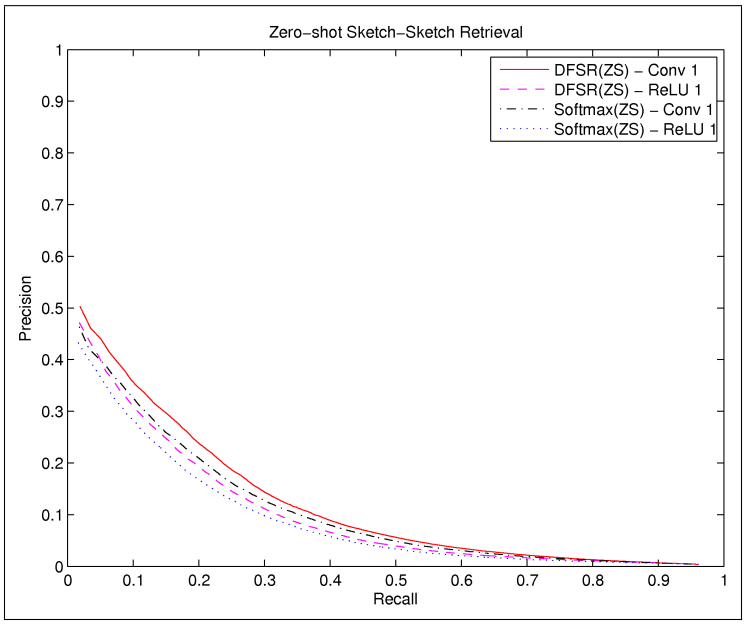
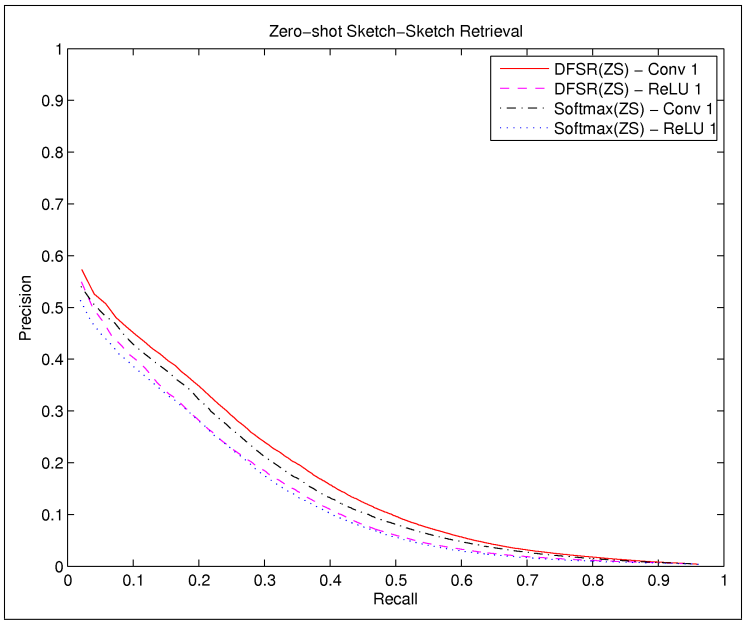


Figure 5.3 Cross-Modal Retrieval : Some sample queries and their corresponding retrievals. Each row corresponds to a retrieval. The first column is the query image and the next five columns are retrieval results from our system. It can be observed from the results that structural information is being encoded in the features. There are a few failure cases as well. For example results are not satisfactory for rows $4, 5, 6, 7$. We can attrribute this failure either to the poor quality of sketch or complexity of the object.



(a)



(b)

Figure 5.4 Zero-Shot Uni-Modal Retrieval : (a) PR curve for the worst performing partition. It can be observed that DFSR outperform the features proposed by Yang *et al.* [82]. (b) PR curve for the best performing partition.

5.3.5.2 Cross-Modal Retrieval

Similarly, in cross-modal retrieval we partition the 105 classes common to CALTECH-256 and TU-BERLIN and perform retrieval experiments. We show the the PR curves for the best and worst partitions in Figure 5.5

5.4 Summary

In this chapter, we proposed a semantic feature space where the semantically related classes are mapped closely. The features are extracted using an existing model of convolutional neural network. In order to encode the semantic relationship we replace the last soft-max layer of the network with a loss function based on Word2Vec distance. Unlike soft-max loss, our loss function penalizes a distant miss more than a closer miss. We perform exhaustive experiments and show the robustness of our features in classification, uni-modal retrieval, cross-modal retrieval and zero-shot retrieval.

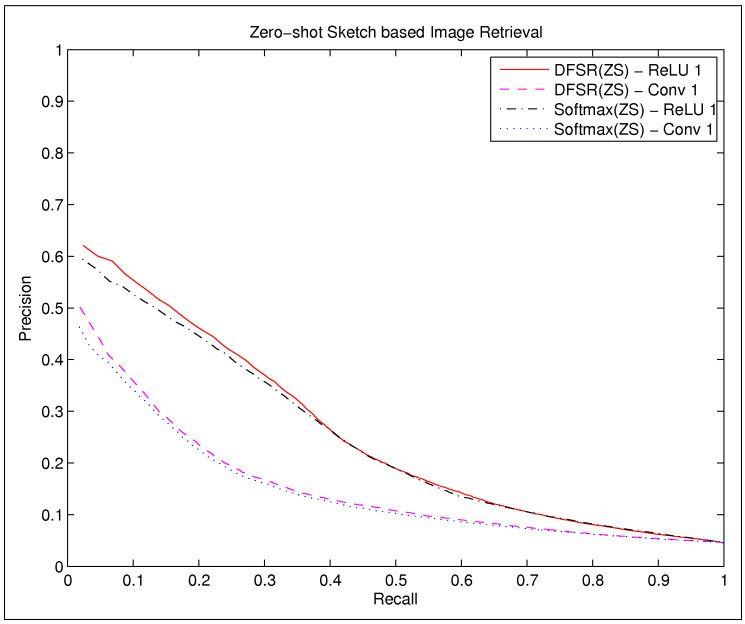
Our approach is limited by the unavailability of enough training samples for the CNN. Existing successful approaches in classification and retrieval are primarily driven by data. In this regard, we believe, addition of more data to our system will improve the performance considerably.

Another problem with our approach is the use of C-CCA. As mentioned before, our loss function in Equation 5.2, tries to map the images to Word2Vec space. We think, using C-CCA on these features distorts the intermediate representation by projecting the non-linear features into a linear sub-space. Ideally, the mapping of the two different modalities to the joint sub-space should be incorporated by the CNN itself.

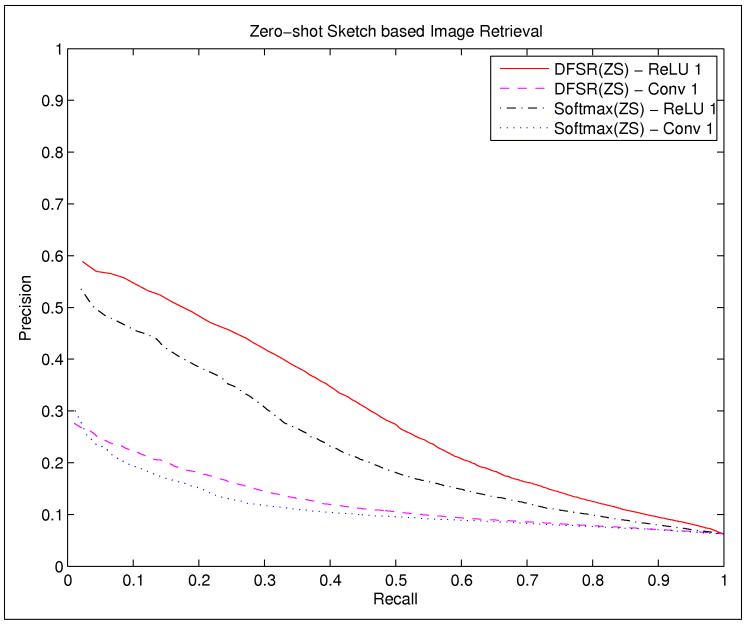
However, to the best of our knowledge, this is the first attempt towards zero-shot learning for sketch-based image retrieval. Moreover, we do not rely on the edge-based similarity between sketches and images but train the two modalities independently. This enables us formulate the problem as a cross-modal problem and connect the abstractions in sketches with real world images, instead of a uni-modal edge-matching.

Also due to the novelty in implementation, mentioned in Section 5.3.1 , our system is scalable to a much larger dataset with the same memory requirements.

Sketch-based multimedia retrieval is an emerging and very active area of research. A thorough study of visual perception of objects in humans and their abstract representation in the form of sketches will enable us to develop better systems. Extending the object level understanding of sketches to the scenes is another track that could be explored from here. We believe that in upcoming years, with rapidly increasing data availability, sketch based multimedia search will open new facets to human computer interaction.



(a)



(b)

Figure 5.5 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.* [82]. (b) PR curve for the best performing partition.

Chapter 6

Conclusion

The problems of sketch-based multimedia retrieval has been attempted in Computer Vision for around two decades. It started to gain importance with the evolution of touch-based devices, but has become popular in the Computer Vision research community only recently. In the last two years, solutions for recognition of sketches (Yu *et al.* [83]) as well as retrieval of images using sketches (Bui *et al.* [10]) were received enthusiastically in both academic conferences (BMVC-2015 and ICCV-2015) as well as in the popular media. In [83], a CNN-based classifier was shown to perform better than humans on TU-Berlin dataset in terms of recognition performance. The recent trend of active research in this area signifies the importance of the problem. However, the problem is still in its initial days of research and there are a several directions to be explored in this field.

For example, a better representation of the queries would require a thorough perceptual study of the art of sketching by ordinary users. Such studies will help us understand the problem more deeply and devising better features. Whether a hybrid mode of query performs better than a single modality or not has not been verified in the real-world. The observations from such a study could improve and extend the existing paradigms of query modelling.
Instead of training the two modalities separately, another direction could be training both the modalities together and propose a joint representation along the lines of Siamese Networks .
Another direction worth exploring is the understanding of sketch-scenes. Undoubtedly, understanding scenes from sketches would be more difficult because of the sparsity of information and the abstraction that sketches present.

The area of sketch understanding is very interesting and solutions to some of the above problems will give insight into the human perception of images and scenes in addition to enabling practical applications such as image and video retrieval and computer assisted art works.

Bibliography

[1] A. Basharat, A. Gritai, and M. Shah. Learning object motion patterns for anomaly detection and improved object detection. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.

[2] F. I. Bashir, A. A. Khokhar, and D. Schonfeld. Object trajectory-based activity classification and recognition using hidden markov models. *Image Processing, IEEE Transactions on*, 16(7):1912–1919, 2007.

[3] F. I. Bashir, A. A. Khokhar, and D. Schonfeld. Real-time motion trajectory-based indexing and retrieval of video sequences. *Multimedia, IEEE Transactions on*, 9(1):58–65, 2007.

[4] G. Bishop and G. Welch. An introduction to the kalman filter. *Proc of SIGGRAPH, Course*, 8:27599–3175, 2001.

[5] T. Bouwmans, F. El Baf, and B. Vachon. Background modeling using mixture of gaussians for foreground detection-a survey. *Recent Patents on Computer Science*, 1(3):219–237, 2008.

[6] S. Brain. Instagram company statistics, 2015.

[7] M. Bregonzio, S. Gong, and T. Xiang. Recognising action as clouds of space-time interest points. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1948–1955. IEEE, 2009.

[8] M. Broilo, N. Piotto, G. Boato, N. Conci, and F. G. De Natale. Object trajectory analysis in video indexing and retrieval applications. In *Video Search and Mining*, pages 3–32. Springer, 2010.

[9] S. Brutzer, B. Hoferlin, and G. Heidemann. Evaluation of background subtraction techniques for video surveillance. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1937–1944. IEEE, 2011.

[10] T. Bui and J. Collomosse. Scalable sketch-based image retrieval using color gradient features. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 1–8, 2015.

[11] 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.

[12] Y. Cao, H. Wang, C. Wang, Z. Li, L. Zhang, and L. Zhang. Mindfinder: interactive sketch-based image search on millions of images. In *Proceedings of the international conference on Multimedia*, pages 1605–1608. ACM, 2010.

[13] S.-F. Chang, W. Chen, H. J. Meng, H. Sundaram, and D. Zhong. Videoq: an automated content based video search system using visual cues. In *Proceedings of the fifth ACM international conference on Multimedia*, pages 313–324. ACM, 1997.

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

[15] B.-W. Chen, J.-C. Wang, and J.-F. Wang. A novel video summarization based on mining the story-structure and semantic relations among concept entities. *IEEE Transactions on Multimedia*, 11(2):295–312, 2009.

[16] D. Comaniciu, V. Ramesh, and P. Meer. Real-time tracking of non-rigid objects using mean shift. In *Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on*, volume 2, pages 142–149. IEEE, 2000.

[17] C. D. Correa and K.-L. Ma. Dynamic video narratives. *ACM Transactions on Graphics (TOG)*, 29(4):88, 2010.

[18] I. J. Cox, M. L. Miller, T. P. Minka, T. V. Papathomas, and P. N. Yianilos. The bayesian image retrieval system, pichunter: theory, implementation, and psychophysical experiments. *IEEE transactions on image processing*, 9(1):20–37, 2000.

[19] N. P. Cuntoor, B. Yegnanarayana, and R. Chellappa. Activity modeling using event probability sequences. *Image Processing, IEEE Transactions on*, 17(4):594–607, 2008.

[20] J. E. Cutting. Representing motion in a static image: constraints and parallels in art, science, and popular culture. *PERCEPTION-LONDON-*, 31(10):1165–1194, 2002.

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

[22] DMR. By the numbers : 120+ amazing youtube statistics, 2014.

[23] A. Dyana and S. Das. Mst-css (multi-spectro-temporal curvature scale space), a novel spatio-temporal representation for content-based video retrieval. *Circuits and Systems for Video Technology, IEEE Transactions on*, 20(8):1080–1094, 2010.

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

[25] M. Elhoseiny, B. Saleh, and A. Elgammal. Write a classifier: Zero-shot learning using purely textual descriptions. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2584–2591, 2013.

[26] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.

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

[28] W. S. Geisler. Motion streaks provide a spatial code for motion direction. *Nature*, 400(6739):65–69, 1999.

[29] K. Ghosal, A. Prabhu, R. Dasgupta, and A. Namboodiri. Learning clustered sub-spaces for sketch-based image retrieval. *Asian Conference on Pattern Recognition*, 2015.

[30] A. B. Godbehere, A. Matsukawa, and K. Goldberg. Visual tracking of human visitors under variable-lighting conditions for a responsive audio art installation. In *American Control Conference (ACC), 2012*, pages 4305–4312. IEEE, 2012.

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

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

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

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

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

[36] J.-W. Hsieh, S.-L. Yu, and Y.-S. Chen. Motion-based video retrieval by trajectory matching. *Circuits and Systems for Video Technology, IEEE Transactions on*, 16(3):396–409, March 2006.

[37] R. Hu, S. James, T. Wang, and J. Collomosse. Markov random fields for sketch based video retrieval. In *Proceedings of the 3rd ACM conference on International conference on multimedia retrieval*, pages 279–286. ACM, 2013.

[38] W. Hu, D. Xie, Z. Fu, W. Zeng, and S. Maybank. Semantic-based surveillance video retrieval. *Image Processing, IEEE Transactions on*, 16(4):1168–1181, 2007.

[39] W. Hu, N. Xie, L. Li, X. Zeng, and S. Maybank. A survey on visual content-based video indexing and retrieval. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 41(6):797–819, 2011.

[40] S. E. Johnson and P. C. Woodland. A method for direct audio search with applications to indexing and retrieval. In *Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on*, volume 3, pages 1427–1430. IEEE, 2000.

[41] B. F. Klare, Z. Li, and A. K. Jain. Matching forensic sketches to mug shot photos. *PAMI*, 2011.

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

[43] H. W. Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

[44] 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.

[45] D. Li, N. Dimitrova, M. Li, and I. K. Sethi. Multimedia content processing through cross-modal association. In *ACM Multimedia*, 2003.

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

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

[48] D. G. Lowe. Object recognition from local scale-invariant features. In *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, volume 2, pages 1150–1157. Ieee, 1999.

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

[50] V. Mahadevan and N. Vasconcelos. Background subtraction in highly dynamic scenes. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–6, June 2008.

[51] G. Medioni and Y. Yasumoto. Corner detection and curve representation using cubic b-splines. In *Robotics and Automation. Proceedings. 1986 IEEE International Conference on*, volume 3, pages 764–769, Apr 1986.

[52] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

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

[54] A. Monnet, A. Mittal, N. Paragios, and V. Ramesh. Background modeling and subtraction of dynamic scenes. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pages 1305–1312. IEEE, 2003.

[55] M. Müller. Dynamic time warping. *Information retrieval for music and motion*, pages 69–84, 2007.

[56] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng. Multimodal deep learning. In *ICML*, 2011.

[57] M. Norouzi, T. Mikolov, S. Bengio, Y. Singer, J. Shlens, A. Frome, G. S. Corrado, and J. Dean. Zero-shot learning by convex combination of semantic embeddings. *arXiv preprint arXiv:1312.5650*, 2013.

[58] M. Palatucci, D. Pomerleau, G. E. Hinton, and T. M. Mitchell. Zero-shot learning with semantic output codes. In *Advances in neural information processing systems*, pages 1410–1418, 2009.

[59] V. M. Patel, R. Gopalan, R. Li, and R. Chellappa. Visual domain adaptation: A survey of recent advances. *IEEE signal processing magazine*, 32(3):53–69, 2015.

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

[61] R. Raina, A. Battle, H. Lee, B. Packer, and A. Y. Ng. Self-taught learning: transfer learning from unlabeled data. In *Proceedings of the 24th international conference on Machine learning*, pages 759–766. ACM, 2007.

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

[63] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *IJCV*, 2015.

[64] I. Saleemi, K. Shafique, and M. Shah. Probabilistic modeling of scene dynamics for applications in visual surveillance. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(8):1472–1485, 2009.

[65] R. K. Sarvadevabhatla and R. V. Babu. Freehand sketch recognition using deep features. *CoRR*, abs/1502.00254, 2015.

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

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

[68] Y. Sheikh and M. Shah. Bayesian modeling of dynamic scenes for object detection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(11):1778–1792, Nov 2005.

[69] H. Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244, 2000.

[70] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pages 1470–1477. IEEE, 2003.

[71] A. W. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *IEEE Transactions on pattern analysis and machine intelligence*, 22(12):1349–1380, 2000.

[72] R. Socher, M. Ganjoo, C. D. Manning, and A. Ng. Zero-shot learning through cross-modal transfer. In *Advances in neural information processing systems*, pages 935–943, 2013.

[73] T. F. Stahovich, R. Davis, and H. Shrobe. Qualitative rigid-body mechanics. *Artificial Intelligence*, 119(12):19 – 60, 2000.

[74] C. Stauffer and W. E. L. Grimson. Adaptive background mixture models for real-time tracking. In *Computer Vision and Pattern Recognition, 1999. IEEE Computer Society Conference on.*, volume 2. IEEE, 1999.

[75] C. Stauffer and W. E. L. Grimson. Learning patterns of activity using real-time tracking. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(8):747–757, 2000.

[76] S. Teja and A. Namboodiri. A ballistic stroke representation of online handwriting for recognition. In *Document Analysis and Recognition (ICDAR), 2013 12th International Conference on*, pages 857–861, Aug 2013.

[77] J. B. Tenenbaum and W. T. Freeman. Separating style and content with bilinear models. *Neural computation*, 2000.

[78] D. Trends. Facebook reveals we upload a whopping 350 million photos to the network daily. 2013.

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

[80] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, volume 1, pages I–511. IEEE, 2001.

[81] C. Yajima, Y. Nakanishi, and K. Tanaka. Querying video data by spatio-temporal relationships of moving object traces. In *Visual and Multimedia Information Management*, pages 357–371. Springer, 2002.

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

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

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

[85] H. Zhang, A. Kankanhalli, and S. W. Smoliar. Automatic partitioning of full-motion video. *Multimedia systems*, 1(1):10–28, 1993.

[86] Z. Zivkovic. Improved adaptive gaussian mixture model for background subtraction. In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, volume 2, pages 28–31 Vol.2, Aug 2004.