Ryerson University
Digital Commons @ Ryerson

Theses and dissertations

1-1-2012

Large-scale Content-based Multimedia Analysis And Applications Using Bag-Of-Words Model

Ning Zhang
Ryerson University

Follow this and additional works at: http://digitalcommons.ryerson.ca/dissertations

 Part of the Electrical and Computer Engineering Commons

Recommended Citation
Zhang, Ning, "Large-scale Content-based Multimedia Analysis And Applications Using Bag-Of-Words Model" (2012). *Theses and dissertations.* Paper 1661.

This Thesis is brought to you for free and open access by Digital Commons @ Ryerson. It has been accepted for inclusion in Theses and dissertations by an authorized administrator of Digital Commons @ Ryerson. For more information, please contact bcameron@ryerson.ca.

Large-scale Content-based Multimedia Analysis and Applications Using Bag-of-words Model

by

Ning Zhang

Bachelor of Applied Science, University of Toronto, 2006

A thesis

presented to Ryerson University

in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

in the Program of

Electrical and Computer Engineering

Toronto, Ontario, Canada, 2012

©Ning Zhang 2012

I hereby declare that I am the sole author of this dissertation. This is a true copy of the dissertation, including any required final revisions, as accepted by my examiners.

I authorize Ryerson University to lend this dissertation to other institutions or individuals for the purpose of scholarly research.

I further authorize Ryerson University to reproduce this dissertation by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

I understand that my dissertation may be made electronically available to the public.

Large-scale Content-based Multimedia Analysis and Applications Using Bag-of-words Model

Doctor of Philosophy 2012

Ning Zhang

Electrical and Computer Engineering

Ryerson University

Abstract

This dissertation focuses on the analysis of large-scale image and video data consortia with applications to multimedia indexing and retrieval. Bag-of-words (BoW) model is adopted and improved to suit the efficiency and effectiveness requirements in analyzing large-scale multimedia data. BoW method has been developed from the text retrieval domain and successfully applied in computer vision, such as image scene and object categorization. Specifically, we utilized the BoW model in the domain of image classification and retrieval, tackled challenges of large-scale multimedia applications of video analysis and mobile-based social activity recommendation using visual intents, respectively.

Incorporating the BoW model with unsupervised classification, we propose a scalable and generic approach in video analysis. The method aims at systematically analyzing unlabeled video from its genre identification, frame classification, and event detection. Unlike conventional domain-knowledge dependent approaches, the BoW model is

domain-knowledge independent. Moreover, the system is mainly unsupervised and requires minimum human input. Therefore, our method is capable of processing massive quantity of videos generically. In addition, for the evaluation, sports video has been used as the testing ground.

Combining the BoW model with advanced retrieval algorithms, we propose a mobile-based visual search and social activity recommendation system. The merit of the BoW model in large-scale image retrieval is integrated with the flexible user interface provided by the mobile platform. Instead of text or voice input, the system takes visual images captured from the built-in camera and attempts to understand users' intents through interactions. Subsequently, such intents are recognized through a retrieval mechanism using the BoW model. Finally, visual results are mapped onto contextually relevant information and entities (i.e. local business) for social task suggestions. Hence, the system offers users the ability to search information and make decisions on-the-go.

Acknowledgements

First and foremost, I would like to express my sincere gratitude to my Ph.D. supervisor, Dr. Ling Guan, for his patience, support and guidance throughout my Ph.D. study and thesis writing. His enthusiasm, inspiration, and advice made my Ph.D. journey both delightful and productive. He is truly a role model: not only as a keen and productive researcher, but also as a great human being.

I also would like to thank all the researchers whom I have collaborated with over the past few years. In particular, I want to thank Dr. Tao Mei and Dr. Xian-Sheng Hua from Microsoft Research Asia (MSRA), who shared with me their superb knowledge in the field of multimedia retrieval and analysis. I benefited immensely from their mentorship during my internship at MSRA. My appreciation also goes to Dr. Ling-Yu Duan and Dr. Wen Gao, whom I collaborated with during my visit to Peking University. I am thankful to have had the opportunity to work with them in the field of video retrieval and analysis. Their enormous insights and detailed discussions undoubtedly contributed to the completion of my Ph.D.

It is a blessing for me to be part of the Ryerson Multimedia Research Laboratory (RML). I gained both knowledge and friendship during my stay at RML, working with my fellow researchers and students. My years in RML have been and will always be the most memorable time and enlightenment of my life.

For this dissertation, my gratitude also goes to the defence committee members: Dr. Alagan Anpalagan (Chair of the committee), Dr. Truman Yang, and Dr. Anastasios Venetsanopoulos, for their time and valuable suggestions.

Last but not least, I want to thank my parents for their unconditional love, encouragement, and continual support. Life would be meaningless without your love.

Publications

1. **N. Zhang**, L. Duan, L. Li, Q. Huang, J. Du, W. Gao, and L. Guan, "A Generic Approach for Systematic Analysis of Sports Videos", ACM Trans. on Intelligent Systems and Technology (TIST), Vol. 3, No. 3, Article 46, May 2012.

2. T. Mei, S. Li, Y-Q. Xu, **N. Zhang**, Z. Chen, J-T. Sun, "Gesture-based Visual Search", US Preliminary Patent, Filed in April, 2011. (submitted patent)

3. **N. Zhang**, T. Mei, X-S. Hua, L. Guan, S. Li. "Interactive Mobile Visual Search for Social Activities Completion Using Query Image Contextual Model", Proc. of IEEE International Workshop on Multimedia Signal Processing (MMSP), 2012.

4. **N. Zhang**, T. Mei, X-S. Hua, L. Guan, S. Li, "TapTell: Understanding Visual Intents On-the-go", Proc. of ACM International Conference on Multimedia (ACM MM), pp. 777-778, Scottsdale, AZ, USA, November 2011.

5. **N. Zhang**, T. Mei, X-S. Hua, L. Guan, S. Li, "Tap-to-Search: Interactive and Contextual Visual Search on Mobile Devices", Proc. of IEEE International Workshop on Multimedia Signal Processing (MMSP), pp. 1-5, Hangzhou, China, October 2011.

6. **N. Zhang**, L. Li, L. Duan, Q. Huang, J. Du, W. Gao, and L. Guan, "Automatic Video Genre Categorization and Event Detection Techniques on Large-scale Sports Data", Twentieth IBM CASCON, pp. 283-297, Toronto, Canada, November 2010.

7. **N. Zhang** and L. Guan, "An efficient framework on large-scale video genre classification", Proc. of IEEE International Workshop on Multimedia Signal Processing (MMSP), pp. 505-510, Saint-Malo, France, October 2010.

8. L. Li, **N. Zhang**, L. Duan, Q. Huang, J. Du, and L. Guan, "Automatic sports genre categorization and view-type classification over large-scale dataset", Proc. of ACM International Conference on Multimedia (ACM MM), pp. 653-656, Beijing, China, October 2009.

9. **N. Zhang**, and L. Guan, "Graph Cuts in Content-Based Image Classification and Retrieval with Relevance Feedback", Advances in Multimedia Information Processing: PCM, pp. 30-39, HongKong, China, December 2007.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background

Living in information era, we are surrounded by an enormous amount of digital content. According to Bohn and Short [4], the estimated size of newly created digital data in 2011 is about 1800 exabyte (1 exabyte=1 billion gigabytes), roughly 100 times more than the production in 2002 ($2 \sim 3$ exabyte). This number is equivalent to a ten-fold average annual growth rate. In terms of image and video content, according to the latest released statistics, YouTube hosts more than 120 million copyrighted claimed videos and serves four billion video requests per day [1]. Facebook, on the other hand, hosts about 50 billion photos (2010), 15 billion of which are tagged [2]. Another statistical result shows that Facebook had 845 million monthly active users and 483 million daily active users on average in December 2011 [3]. Undoubtedly, digital content, including images and videos, are deeply rooted in our daily life, from desktops and laptops to mobile phones and tablets. Large-scale content-based multimedia data organization and analysis not only helps to retrieve users' desired information, but also serves the basis and first step to multimedia applications such as image/video classification and retrieval, as well as the recent boom of cross-platform mobile visual search and recommendations.

[1] http://www.youtube.com/t/press_statistics

[2] http://www.usatoday.com/tech/news/2010-07-21-facebook-hits-500-million-users_N.htm

[3] http://newsroom.fb.com/content/default.aspx?NewsAreaId=22

1.2 Challenges and Relevant Technologies

Large-scale image and video search and classification have drawn tremendous interest from research communities. Different from small scale content-based multimedia analysis, unique challenges of large-scale multimedia analysis include, among others:

- First, automatic image classification, with minimum human labeling and intervention. According to a recent study, among web-based image and video consortia, only $5-10\%$ of total data are labeled [5]. The majority of multimedia data cannot be retrieved using current textual-based search engines. Moreover, successful image classification will be useful in frame-based video analysis, such as genre classification and event detection.

- Second, image retrieval, including efficient database index, compact storage, and quick and accurate retrieval performance. Since large-scale databases consist of millions of images, computational efficiency of both off-line and on-line is crucial.

- Third, integration with cross platform-based applications. With the emerging technologies of mobile devices and cloud computing, a lot of desktop-based multimedia applications need to be migrated to and find suitable positions in the mobile domain.

1.2.1 Large-scale Content-based Image Classification and Retrieval

From the multimedia application perspective, large-scale image classification serves as a middle agent to link low-level features and high-level semantic events [6, 7]. Supervised methods are favorable choices in the research community. Although the size and diversity of current databases can be managed with those tasks using labeled training data, such tasks become more and more unmanageable with the growing scale of the dataset, as well as the unlabeled content. Therefore, algorithms using unsupervised learning techniques with generality and efficiency ought to be sought for analyzing large-scale multimedia consortia.

On the other hand, content-based image retrieval (CBIR) has attracted researchers in the field of computer vision, machine learning, database technology, and multimedia

for almost two decades. It still remains a popular research direction, especially when considering how to cope with the vast size of and increasing growth of multimedia data. In the beginning of this millennium, Rui, Huang, and Chang stated that there are two major difficulties with large-scale image datasets [8]. One is the vast amount of labor required in manual image annotation. The other, is how to understand different human perceptions towards the same image content. Moreover, how to efficiently index large-scale image archives for fast retrieval is also raised as a fundamental consideration in designing large-scale image retrieval systems [8, 9].

1.3 Thesis Contributions

In response to the above mentioned challenges, bag-of-words (BoW) model is adopted for multimedia analysis in this thesis, in particular, at large-scale image classification and retrieval [10, 11]. BoW model can effectively combines locally extracted feature vectors of either an image or a video frame. It focuses on the characteristics of the local feature ensemble, and treats individual local descriptors uniformly. The merits of the BoW include a homogenous process in which it compactly represents images or video frames for classification, as well as its availability in large-scale image retrieval due to its success in text retrieval. Contributions of this thesis are presented as follows.

1.3.1 BoW in Unsupervised Classification and Video Analysis

The first contribution of this thesis is to use the BoW model for unsupervised classification in video analysis. A distinguishing yet compact representation of the video clip is proposed using the BoW model. Candidate videos are indexed and represented as a histogram-based interpretation using the learned BoW model. The advantage of using the BoW model is that labeled data is not required. Therefore, video analysis can be realized towards large-scale applications.

Using the above mentioned method, this thesis presents a systematic and generic approach by using proposed BoW based video representation. The system aims at event detection scenario of an input video with an orderly sequential process. Initially, domain-knowledge independent local descriptors are extracted homogeneously from the input video sequence. The video's genre is identified by applying the k-nearest neighbor (k-

NN) classifiers onto the obtained video representation, with various dissimilarity measures assessed and evaluated analytically. Subsequently, an unsupervised probabilistic latent semantic analysis (PLSA) based algorithm is employed at the same histogram-based video representation to characterize each frame of video sequence into one of the representative view groups. Finally, a hidden conditional random field (HCRF) structured prediction model is utilized for interesting event detection. In evaluation, sports videos are used as the testing ground.

1.3.2 BoW in Retrieval and Mobile Image Search

The second contribution of this thesis is to explore the BoW's merit in mobile visual search by effectively incorporating user interaction. Efficient and scalable indexing and non-linear fast retrieval algorithms are adopted in handling large-scale images. Human interaction is included in the loop. Therefore, specific user perception and distinguishing request is delivered to lead the system into achieving a customized search result.

Based on the above idea, an interactive mobile visual search application aimed at social activity suggestion is developed using a coined term "visual intent", which can be naturally expressed through a visual query incorporating human specification. To accomplish the discovery of visual intent on the phone, we developed *TapTell*, an exemplary real application on the Windows Phone 7. This prototype takes advantage of user interaction and rich context to enable interactive visual search and contextual recommendation. Through the *TapTell* system, a mobile user can take a photo and indicate an object-of-interest within the photo via a *circle* gesture. Then, the system performs a search-based recognition by retrieving similar images based on both the object-of-interest and surrounding image context. Finally, the contextually relevant entities (i.e. local businesses) are recommended to complete social tasks.

1.4 Organization of the Thesis

The rest of the thesis is organized as follows. Chapter 2 reviews, in detail, the BoW model and its computer vision applications of image classification and retrieval. In Chapter 3, a generic and systematic approach in large-scale video analysis is proposed. The BoW model is used to represent video clips for unsupervised genre categorization and view

classification. Semantic event detection is achieved by using middle level view classification of sampled video frames after its genre categorization. In Chapter 4, a mobile-based visual search utilizing image context is proposed so that social task recommendation can be completed as the end result. The BoW model is used in indexing and visual recognition under the contextual model. Once the retrieved result of a visual query is connected with the established textual database, a more accurate text-based search is used in the recommendation. Finally, Chapter 5 presents conclusions and future research directions.

Chapter 2

Literature Review on Related Works

2.1 Introduction

In this chapter, we review the BoW model, its original proposal, and early years of applications in computer vision. We also present recent advanced methodologies utilizing the BoW model in large-scale image classification and retrieval.

The phrase "*a picture is worth a thousand words*" has been frequently cited in published works in image classification and retrieval. On one hand, the phrase indicates the convenience of using rich content conveyed by digital images in overcoming language barriers and textual description limitations. Early systems such as QBIC [12, 13], VisualSEEk [14], Photobook [15], and Virage [16], successfully delivered image retrieval systems and achieved acceptable performance using low-level global features.

On the other hand, an image may carry too much information making it beyond the distinguishable representation merely using global features. Moreover, different humans perceive and interpret the same image content differently. Since one global feature is limited in capturing detailed aspects of the image, it is difficult to use the single global feature to represent various interpretations. In recent years, local invariant features have been developed to tackle the aforementioned challenges. Such keypoint based salient patches contain rich local information and they are more resilient in dealing with various conditions and occlusions. An ensemble of local features is treated as a virtual bag containing visualwords to visually represent images or video frames. This paradigm is called the BoW model. Figure 2.1 illustrates such an idea of combining salient patches

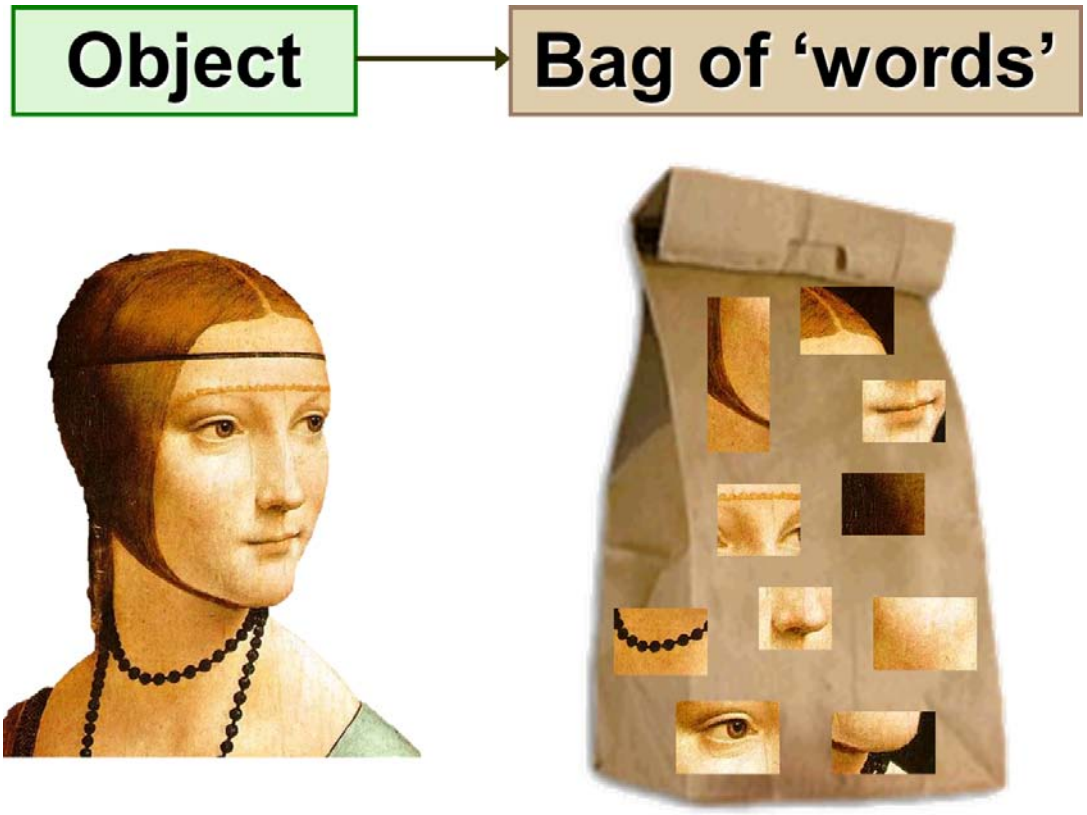


Figure 2.1: Illustration of bag-of-words concept [1].

in a virtual ensemble.

The term BoW was coined in the information retrieval and natural language processing research areas. The idea is to generate an unordered collection of textual words for representation in document retrieval, each of which is weighted equally disregarding the grammar connection and word order [10, 11]. Such a BoW model and notation has been adopted in the field of computer vision for similar visual feature classification and retrieval [17, 18]. Similar to its document retrieval counterpart, a single image or video frame is treated by extracting unordered local visual descriptors as visual "words" to represent the image "document". Then, the image itself becomes a bag of "words" for the later classification and retrieval processes. A detailed literature review of the BoW model utilization in computer vision is given in the following section.

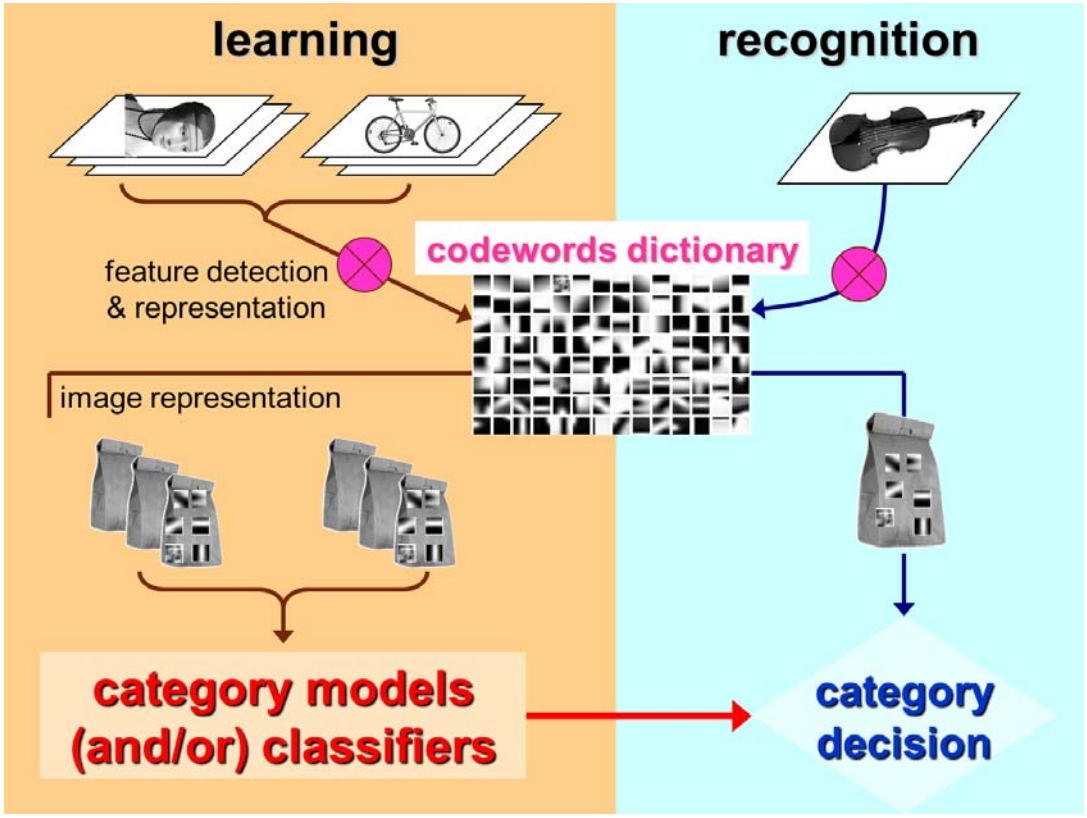


Figure 2.2: Illustration of bag-of-words framework in computer vision [1].

2.2 Bag-of-words (BoW) Model

Figure 2.2 shows a framework of the BoW model and its usage in computer vision. In general, there are two parts: learning and recognition. In learning, visual features are extracted from database images or video frames to generate a dictionary of codewords, which is also called a codebook in the literature. Individual images are used to project their features to the codebook to obtain a BoW representation for themselves. They are then categorized by classifiers to get ready for recognition. In recognition, a query or testing image also goes through the BoW model by mapping to the dictionary of codewords. Then, the BoW representation is categorized based on which class the query image belongs to.

2.2.1 Related Works

Initially, the BoW concept was used in the texture image classification. The idea was to use K-means clustering to compute a texton (a frequency-based spatial-frequency selective linear filters) library so that histogram-based representation of texture images could be generated for classification. Notable works include reference [19] [20] [21]. In [19], Cula and Dana proposed a bidirectional texture function, a compact feature for texture image representation. Leung and Malik proposed a three-dimensional textons method for extracting texture image features that adjusts with different lighting and viewing conditions [20]. Varma and Zisserman further extended the work by proposing rotational-invariant filters to achieve viewpoint and illumination independence for texture image classification without using priori knowledge [21]. Lazebnik *et al.* focused on finding scale- and affine-invariant detectors to localize interesting points for computing affine-invariant descriptors [22]. A codebook was built from a clustering method. Previous representative works focused on different feature extractions but all used the BoW model to build a dictionary of codewords for image interpretation. This fact demonstrates the popularity of using the BoW model in the field of texture image classification.

With the help of a more robust and sophisticated local feature extraction, the BoW model has been widely applied in more applications of computer vision, such as view/scence classification, object categorization, image segmentation and stitching, duplicated image/video detection, and concept/object detections in video. In the domain of view/scence classification, pioneering works include Fei-Fei and Perona's proposal of a probabilistic and generative model to categorize natural scene images [23]. In particular, they used a Bayesian hierarchical framework to automatically learn the distribution of codewords. Bosch *et al.* extended the idea and proposed a hybrid approach by incorporating a generative statistical model with discriminative support vector machines (SVMs) and k-nearest neighboring (k-NN) algorithms [24]. In the field of object categorization and matching, Sivic and Zisserman proposed a text retrieval method to match objects in videos [17,18]. Different from previously introduced histogram representation, inverted file systems and document frequency weighings were adopted from text retrieval and were applied in indexing database images efficiently. Sivic *et al.* further proposed a mechanism to learn multiple object categories and content locations using an unsupervised generative statistical model, which is greatly beneficial in processing unlabeled data [25]. Some other

researchers focused on image classification and segmentation concurrently by introducing spatial and regional prior knowledge to the BoW model. Cao *et al.* presented a spatially coherent latent topic model, a generative approach in order to simultaneously recognize categories of objects in the scene images as well as segment those objects [26]. Shotton *et al.* presented a method called the bag of semantic textons to avoid heavy computation of local descriptors and codebook learning [27]. An implicit hierarchical relationship between visual features and region prior information is investigated in learning the regional-based histogram representation. This method achieves both overall image categorization and regional segmentation at the same time in an efficient and automatic fashion.

To tackle the multimedia processing challenges associated with recent boom of large-scale data, the BoW model is among the most popular choices in the research community. It has shown impressive performance in image classification and retrieval. In the following sections, we will first discuss popular local descriptors developed in recent years; and then, focus on two different, but related, computer vision tasks: large-scale image classification and large-scale image retrieval.

2.2.2 Local Descriptors

A key component of the BoW model is to develop an accurate description of visualwords (a locally extracted visual feature). Local feature is based on regional semantic patches, where interest points are detected by their properties such as local extrema of pixel intensity, edge, corners, and etc. Different from global feature, where each image only has one single vector description, there are various numbers of local features for each image. The number depends on how many interest points are detected. Hence, there are two major parts in building local descriptors: feature detection and feature description.

Feature detection is the first process of feature extraction in determining the interest points, which are believed to carry the representative information of an image or a video frame. Subsequently, feature description presents a mathematical operation in obtaining a vector form feature descriptor to represent a local semantic patch or region. Consequently, the ensemble of all local feature descriptors is treated as the representation of the image or the video frame, and used in the BoW model.

Feature Detection

Early work of feature detection includes various edge detectors. Edge detection methods focus on identifying those variations of image intensity level in discontinuities (step edges), local extrema (line edges), and corners which intersect two lines (junction edges) [28]. Edge detection can successfully capture representative image information and has been widely applied in identifying objects' physical, photometrical, and geometrical properties. However, edge detection is vulnerable to various sources of noises such as electronic effects, devices discretizations/quantifications, and lighting conditions. Moreover, for semantic applications such as image classification and retrieval, edge feature is limited in retaining the differentiation of semantic objects from one another. This limitation is because edge detection focuses on line-based outline and silhouette information. It is not able to describe an enclosed regional feature which is critical for semantic similarity and affinity.

Algorithms developed based on semantic patches were recently proposed to overcome edge detector limitations. There are two groups of semantic patches based feature detection: Blob detection and affine-invariant feature detection. Blob detection aims to find the points or regions of the image that are different from the surrounding pixels. Laplacian of the Gaussian (LoG) and Difference of Gaussians (DoG) are the most common blob detectors, where the latter can be viewed as an approximation of the former. Initially, an input image is convolved by a Gaussian kernel in different scale-spaces to obtain Gaussian smoothed images as scale-space representations. Subsequently, a Laplacian operator is applied at the scale-space representation in the case of LoG; or a Difference operator in the case of DoG is applied in the convolved adjacent scale-space Gaussian smoothed images. Finally, a local extrema detector is built based on a 3-D cube, including the 2-D image space and the 1-D adjacent scale-space [2, 29]. The left column of Figure 2.3 shows the Gaussian smoothing operation on the scale-spaced images and their scales group named octave. The Difference operation of the DoG process is shown on the right column of Figure 2.3. In the LoG case, a direct Laplacian operation is applied on the Gaussian smoothed image of each scale-space.

Another blob detection method is called determinant of the Hessian (DoH). The DoH firstly applies the Gaussian kernels to get the scale-space representation. Then, the Hessian matrix, a second-order partial derivative of the scale-space Gaussian smoothed

image is computed at a specific scale. The second stage of the process is to compute the determinant of each Hessian matrix at the specific scale. The last step of the DoH is the same as LoG and DoG methods in finding the local extrema for interest points detection. An approximate version of determinant of Hessian using Haar wavelet is adopted in finding a so called Speeded Up Robust Features (SURF) descriptor, which is much less intense in computation compared to other blob detectors [30].

Some other algorithms vary from the DoH method by treating scale and spatial information separately. Instead of treating them together in one step as a 3-D cube using DoG, LoG, or Hessian matrix applied in DoH method; a Harris-Laplace method proposes to find the spatial location (the 2-D images) by first using a proposed Harris function. Then, it selects interest points using a maximal of a local Laplacian measure over the scales (1-d scale-spaces) [31]. Similarly, another so-called Hessian-Laplace method replaces the Harris measure by the determinant of the Hessian operator, while keeping the remaining steps the same [32].

Blob detection focuses on retaining the information about local regions for differentiation. It has shown superior invariant properties in translation (shifting), rotation, and uniform re-scaling. However, blob detection is vulnerable and subject to perspective distortion. Some other detection methods need to be sought to preserve geometric transformations, and to avoid deterioration by perspective distortion. These detectors should be invariant in geometric transformations, such as skew and stretch. In mathematics, these skew and stretch transformations are defined as affine transformations. Affine transformation is described as a mapping function to preserve straight lines and ratios of distances between points lying on a straight line. However, it does not necessarily preserve angles or lengths [33]. Because of their importance in object recognition, and image/video classification and retrieval, various affine-invariant feature detection algorithms were proposed to focus on describing images by those interest points that are consistent with various affine transformations.

Harris-affine and Hessian-affine detectors focus on finding initial interest points and regions, with a following affine shape adaptation in normalizing the interest regions to achieve the affine-invariant property [34, 35]. A following iterative process refines the initial affine regions to obtain the final stable affine-invariant regions. The difference between Harris-affine and Hessian-affine is in their initial interest points detection stage. The Harris-affine detector relies on Harris corner detection using the second-moment

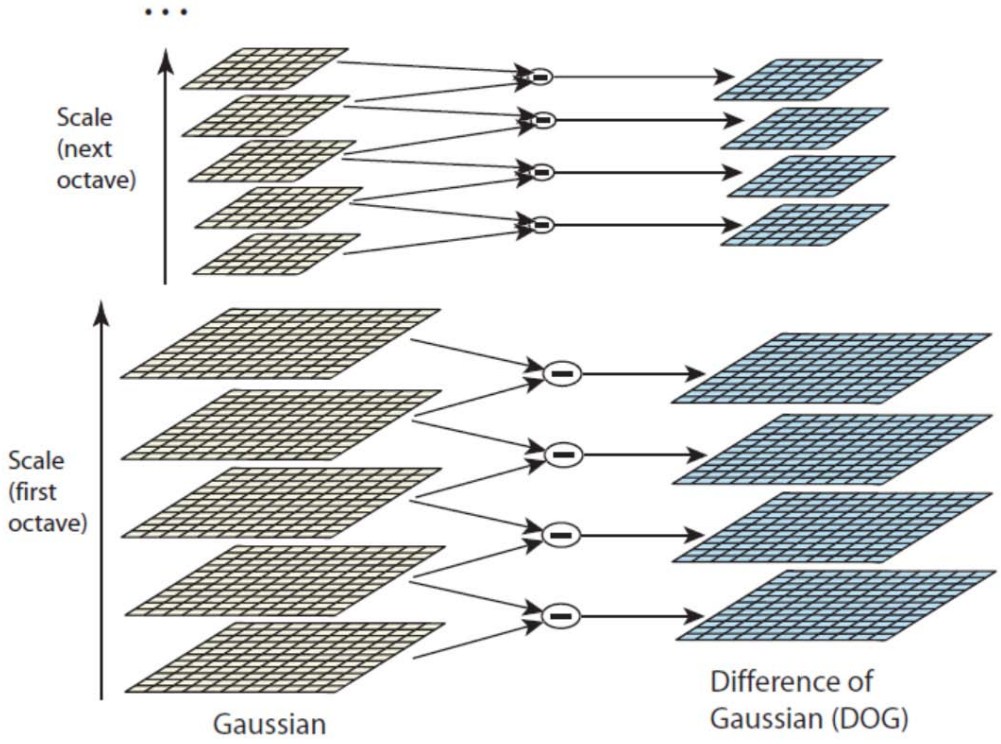


Figure 2.3: Illustration of DoG operation on scale-space representation and its hierarchical scale-group (octave) [2].

matrix, while Hessian-affine uses Hessian matrix for interest points detection.

Some other affine-invariant feature detection includes Edge-based region (EBR) detectors [36], intensity-extrema-based region (IBR) detectors [37], maximally stable extremal regions (MSER) [38], and Kadir-Bradly salient region detectors [39, 40]. EBR focuses on corner points while making use of the nearby edges [36]. The reason is that the edges are stable affine-invariant features subject to various viewpoints and scale changes. IBR starts from detecting intensity extrema over the scale-space to obtain initial interest points. Then, an affine geometric invariant function is applied to explore the surrounding regions to achieve an affine-invariant region. MSER is a method based on a series of processes in applying thresholds at image pixel intensities. This method tries to find those *extremal regions*, such that all pixels inside that MSER have either higher (bright extremal regions) or lower (dark extremal regions) intensity than all the pixels on its outer boundary after applying thresholds. *Maximally stable* describes the property opti-

mized in the threshold selection. As a result, a connected component of an appropriate threshold image is obtained as a MSER. The MSER method is also proven to be affine-invariant [38, 41]. The Kadir-Bradly salient region detector is based on probabilistic density function (pdf) of intensity values over elliptical regions. First, it calculates the entropy value of the pdf over a family of ellipses centered on the interested pixel. The entropy extrema over scales and ellipses are recorded as candidate regions. Then, those candidate salient regions are ranked over the entire image based on the magnitude of the pdf derivative with respect to the scale. Finally, a fixed number of top ranked regions are used as affine-invariant regions.

Feature Description

Feature description is the stage to extract vector representations of identified interest points or regions from previous feature detection stage. Here, we introduce several state-of-the-art feature descriptors which fits the BoW model, including scale-invariant feature transform (SIFT) [2], Gradient location-orientation histogram (GLOH) [42], principle component analysis SIFT (PCA-SIFT), [43], speeded up robust features (SURF) [30], histogram of oriented gradients (HOG) [44], and its compact variation coined compressed histogram of gradients (CHoG) [45].

SIFT feature can robustly identify objects among clutter and partial occlusion. It is claimed to be invariant to uniform scaling, orientation, and partially invariant to affine distortion and illumination [46]. It has also been evaluated as the most resistant to common image deformations, from a comparison study with other local feature descriptors [47]. The SIFT feature descriptor adopts the DoG keypoints detection method on 3D space, consisting of both spatial-space and Gaussian scale-space. A keypoint descriptor is built by first computing gradient magnitude $m(x, y)$ and orientation $\theta(x, y)$ at each sample point, in a region around the detected keypoint, shown as:

$$m(x,y) = \sqrt{(L(x+1,y) - L(x-1,y))^2 + (L(x,y+1) - L(x,y-1))^2} \tag{2.1}$$

$$\theta(x,y) = \arctan((L(x,y+1) - L(x,y-1))/(L(x+1,y) - L(x-1,y))) \tag{2.2}$$

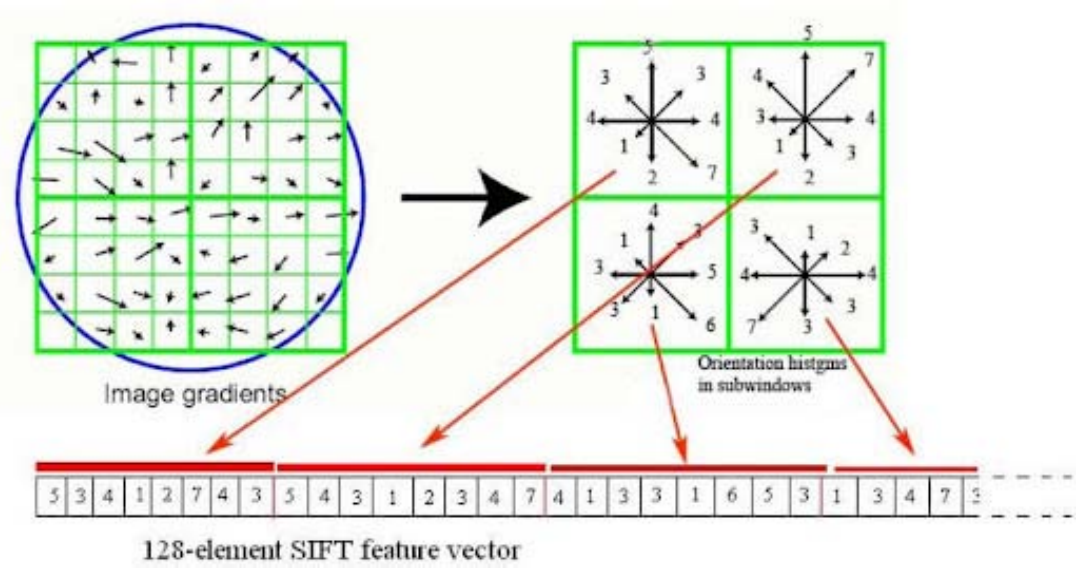


Figure 2.4: Illustration of gradient orientation histogram computation in a down-graded version. A final histogram vector representation concatenated neighborhoods is also shown at the bottom of the illustration [2].

At the scale-space where the keypoint is detected, L and $L(x, y)$ are Gaussian smoothed image and (x, y) pixel smoothed value, respectively. A set of orientation histograms are built on a 4x4 pixel neighborhoods with 8 bins each. Each of these neighborhoods consists of a region of 4x4=16 pixels. This configuration makes a total of 16x16 regions, which are centered at the keypoint. In terms of the vector value, a descriptor of 128 elements is computed as a product of 4x4 neighborhoods and 8 bins each. Figure 2.4 depicts the calculation of the vector descriptor in a down-scaled version of SIFT, proposed by Lowe [2]. In the illustration, a total of 8x8=64 pixels region (instead of 16x16 pixels region) is down-graded to 2x2=4 neighborhoods (instead of 4x4 neighborhoods) in histogram computation, with 8 bins each in histogram.

PCA-SIFT is considered as a variation of the SIFT. It extends SIFT patch-region to a 41x41 pixels patch, which is also centered at the keypoint. A total of 3042 elements are computed as the raw input vector by concatenating the horizontal and vertical gradient maps. Then, a normalization is applied to this vector to minimize the impact of the illumination variation. It is believed that variations of different conditions impacting on the feature vector can be modeled by low-dimensional Gaussian distributions. This is

the reason that Ke and Sukthankar applied the PCA analysis on the raw 3024 elements SIFT vectors [43]. A projection matrix consisting of the top 20 eigenvectors is stored as eigenspace representation and is used for PCA-SIFT mapping. The number 20 is calculated using empirical evaluation. A mapping is conducted using the projection matrix for each new input vector. PCA-SIFT has shown advantages at certain applications and the compact representation is promising in large-scale data mining. However, a pre-built projection matrix is required.

GLOH is designed to increase SIFT descriptor's robustness and distinctiveness. The method is based on a SIFT descriptor computed at a log-polar coordinate system, with a location grid of three bins each in three radius (value at 6, 11, 15), and one bin each at eight angular directions. Thus, a total of 17 (17=3x3+8) location bins are obtained. With each location of 16 bins for gradient orientation, a histogram of 272 (272=17x16) bins is computed. Subsequently, a standard PCA covariance matrix is trained and applied to select the largest 128 eigenvectors for the final GLOH descriptor [42].

SURF focuses on computational efficiency of the local feature. The computation is mainly improved by using an integral image intermediate step for convolutions such that the computation time is reduced [48]. A fast Hessian matrix-based measurement is used in the feature keypoint detection. The feature descriptor is built based on a distribution of Haar wavelets. The integral image is used to speed up the calculation, and a total of 64 dimensions are used as the final feature vector size [30].

HOG uses grid-based dense image descriptors, which is different from the previously introduced salient-based keypoints. This dense grid detection is based on dividing the image window into small spatial regions called "cell". Then, an accumulated 1-D histogram of gradient directions or edge orientations over the pixels (similar as SIFT feature) of the cell is calculated. Finally, histogram entries are combined to represent images, and contrast-normalization is applied to achieve a better illumination and shadowing invariance. Further, Dalal and Triggs introduced two detection methods based on either rectangular or circular log-polar blocks and named them R-HOG and C-HOG, respectively [44]. HOG features can be viewed as a coarse spatial sampling with a fine orientation sampling, followed by a strong local photometric normalization. Therefore, it makes the HOG descriptor particularly suitable for human detection. This is because the individual body movement of humans in image/video shots, which causes lots of noise, is ignored as long as the human maintains a roughly upright position [44].

The recently developed CHoG descriptor provides a compressed version of the HOG feature to satisfy an increasing demand for mobile-based retrieval [49]. This CHoG, as a low-bit-rate compressed feature, fits well with mobile based visual search scenarios and requires low traffic demand through the wireless network. A vector quantization (VQ) process is applied to the gradient distribution to obtain a smaller set of bins than the uncompressed original HOG. This VQ version of the HOG based histogram is encoded by various tree coding techniques such as the Huffman tree and the Gagie tree. It claims to have more than 20 times the bit-rate reduction, while maintaining the baseline image matching performance [45, 50].

2.2.3 Large-scale Image Analysis Using BoW Model

Because of their homogenous procedures in describing images or video frames using representative local features, BoW-based methods enable researchers to conduct large-scale image analysis effectively. Large-scale image classification and retrieval have been carefully studied in recent years to catch up with the ever growing image and video datasets. Image classification and retrieval are highly interrelated research problems. Both of them are based on analyzing distinguished features of the query image, and are in attempts to bring out similar images from the database. Classification focuses on the intra-class commonalities so that the query image can find its suitable class and belonging. Retrieval, on the other hand, focuses on finding the most closely related individual images in the database and returning them as search results. In summary, classification solutions focus on feature ensembles, for instance, the histogram representation of each image. Retrieval solutions focus on both feature ensemble and individual local descriptor matches.

Image Classification

Csurka *et al.* proposed a BoW model-based algorithm for visual image classification from seven different classes, including faces, buildings, trees, cars, phones, bikes and books [51]. SIFT feature is used as the local descriptor, and Naïve Bayes, with non-linear supervised support vector machines (SVM), are used as classifiers. Deng *et al.* proposed a database called "ImageNet", which associates images with large-scale ontology supported by the WordNet structure [52, 53]. Currently, about nine million images are indexed and this number is still growing. Among benchmark measurements and comparisons, a spatial

pyramid-based histogram of SIFT local codewords with SVMs classifiers provides the best performance. Zhou *et al.* proposed a method by incorporating vector coding to achieve scalable image classification [54]. They adopted vector quantization coding on local SIFT descriptors to map the features to form a high-dimensional sparse vector. Spatial information of local regions in each image is taken into account and called spatial pooling. Finally, linear SVMs are used to classify the image representations obtained from the spatial pooling.

Although non-linear SVMs classifiers perform well, they suffer from data scalability due to computational complexity. Perronnin *et al.* proposed several methods to improve non-linear SVMs, including square-rooting BoW vectors, kernel-PCA based embedding for additive kernels, and non-additive kernels for embedding [55, 56]. In particular, an algorithm using Fisher Kernels was proposed to build gradient vectors from features, so that linear SVMs could replace those non-linear ones as less computational classifiers [57]. Hence, the scalability issue was alleviated.

Image Retrieval and Visual Search

Sivic and Zisserman proposed a video scene retrieval system called Video Google [17]. The goal is to retrieve similar objects and scenes and localize their occurrences in a video. MSER feature detection and SIFT feature description are used to extract local descriptors. Visual vocabulary is built by K-means clustering. A *term frequency-inverse document frequency* (tf-idf) text retrieval algorithm is used to match each visualword.

Nistér and Stewénius proposed an efficient and scalable visual vocabulary tree, so that building a large-scale retrieval system using the BoW model is possible [3]. The method adopted hierarchical K-means clustering to boost the codebook generation and retrieval process. The idea is that a query visualword does not necessarily need to go through the full comparison with the codebook. Rather, a subset of the codebook (a branch of the hierarchical K-means clustering) is sufficient. This method allows the codebook to scale up from a few thousands, to hundreds of thousands, to millions in size without much computational penalty. Although there is no automatic mechanism to determine the proper codebook size, in general, a larger vocabulary pool size described by the codebook leads to a better description of the query image with less quantization error [58].

Philbin *et al.* proposed a soft weighting scheme for object retrieval in large scale image databases [59]. This soft-assignment maps high-dimensional SIFT descriptors to a weighted combination of visualwords, rather than to a single visualword as hard assignment. The soft-weighting assignment is designed as an exponential function of the distance to the cluster center. This method allows the inclusion of features which are lost in the quantization stage. Jégou *et al.* also suggested to improve the BoW model by aggregating local descriptors into a compact short binary coded image representation called Hamming embedding (HM) [60, 61]. At the retrieval stage, a tf-idf based index is built with an integration of weak geometric consistency verification mechanism to penalize those descriptors which are not consistent in angle and scale.

2.3 Summary

This chapter introduces the BoW model, from its early evolution to recent developments in which it is combined with local feature descriptors. We also present computer vision challenges in large-scale image classification and retrieval. This thesis focuses on multimedia analysis and applications by incorporating the BoW model and algorithms developed in the field of image classification and retrieval. Chapter 3 proposes a systematic video analysis for representing unlabeled video clips using the BoW model. The system is able to categorize video genres and classify sampled framed scenes using unsupervised learning, and eventually detect semantic events. Chapter 4 shifts the BoW model and image retrieval application to a mobile platform, incorporating it with user interaction. It proposes a context-embedded vocabulary tree for an efficient mobile visual search and retrieval. Consequently, contextual entity recommendation, based on associated image content results and their text-based metadata, is suggested to the users.

Chapter 3

Video Analysis Using the Bag-of-words Model

3.1 Introduction

The bag-of-words (BoW) model and its application in image classification have been used in various aspects of video analysis. Because of its robustness in matching semantic objects using local descriptors, the BoW concept has been used in video object reoccurrence detection [62, 63], semantic shot detection [64, 65] and grouping [66], and object-based video retrieval [18, 67]. Some other representative works in video analysis adopted BoW models with feature tracking along the temporal course, including matching semantically similar videos built by local features using spatiotemporal volumes [68]; content-based video copy detection using high-level descriptions derived from the BoW representation [69]; and, person spotting and retrieval based on their faces features in videos [70]. In the field of video event analysis, Zhou *et al.* applied the BoW model to Gaussian mixture models to represent news videos and utilized kernel-based supervised learning in classifying news event [71]. The BoW model was also used in video clip representation in Xu and Chang's work of video event recognition, where a multilevel temporal pyramid was adopted to integrate information from different sub-clips for pyramid match using temporal alignment [72].

Aforementioned video analysis methods using BoW models have their individual merits. However, there is a lack of systematic investigation, which is important in connecting

individual aspects of the video analysis, from raw input video clip genre categorization, to middle level semantic view or shot understanding, to eventually high-level semantic event analysis. Furthermore, large-scale video data often contains many hours with a lot of insignificant information. The nature of large-scale video data is that it requires an automatic and orderly analysis to obtain efficient information extraction. In this chapter, we propose a BoW model to represent video frames and clips. We also propose an unsupervised learning approach to utilize the BoW-based video representation. We manage to tackle a series of video analysis challenges for unlabeled large-scale video consortia. As a result, a systematic analysis of video data is achieved.

In order to evaluate the effectiveness of the BoW model in the systematic video analysis, we need a valid and meaningful test ground. We believe that large-scale sports videos are ideal. First, sports video is truly a large-scale consortia. It also contributes significantly to the total collection of digital content. Second, sources of sports video collection are also various: from daily-basis public recreations to professional sports games broadcasting; from amateur digital camcorder to professional TV broadcasting, and plenteous but low-quality online streamed videos. Third, sports video analysis is closely connected with real applications, due to its huge popularity and vast commercial value.

Although analysis of sports video has drawn much attention in the research community, most of the literature focus on particular sports and tasks, utilizing domain knowledge and production rules [6, 73–76]. Supervised learning is an important characteristic adopted by these works to fill the semantic gap. These stand-alone methods have little inter-connection and also suffer from a lack of generality and scalability to the large-scale data for two reasons. First, with various video content of different themes and cinematographic techniques, domain knowledge associated methods have difficulties in extensibility. Second, labeled data is required for supervised learning, while the majority of multimedia data available is currently unlabeled. In order to tackle these two issues, our proposed algorithm focuses on using a local domain knowledge-independent SIFT feature to represent video clips using the BoW model and utilizes an unsupervised learning paradigm to deal with unlabeled large volume data.

In this chapter, a generic and systematic framework is proposed with experimentations on a large-scale sports video dataset. Three tasks are introduced such that the output from the previous tasks are utilized as the input to the next task. Event detection is the third and final quest with two preceding tasks, video genre categorization and semantic

view type classification. By accomplishing these three tasks, event detection can be achieved with minimum domain knowledge and partially labeled data. Although we perform our methods on sports video, the generic nature makes the proposed framework valid in evaluating other video consortia.

The novelty of this framework lies in the following three aspects:

(1) Domain knowledge-free local descriptors are extracted using a homogeneous process. The BoW model is used to build a histogram-based distribution to represent video clips. The BoW based video representation using local features is the natural selection for generically processing videos due to its domain knowledge-free properties.

(2) An unsupervised classifier with homogeneous process is proposed. This choice of method is because that unlabeled data takes the major portion of all digital content. Thus, an automatic and systematic process can be deployed towards a large-scale dataset. Since sports videos have well defined semantic view types from their production characteristics, local features combined with the BoW model is a perfect candidate in view classification. Such a combination has also been proven successful in computer vision and object recognition (details in Chapter 2). Therefore, a probabilistic latent semantic analysis (PLSA)-based method for semantic view classification is preferred due to its unsupervised nature and applicability to the BoW model.

(3) A structured prediction model is adopted for taking labeled middle-level agents as input to achieve high-level semantics. This choice is because that sports videos have distinguishable temporal patterns often consisting of sequences of middle-level agents. In our work, since semantic view types have been classified in part (2), an appropriate method is to take the view results as input and achieve semantic event detection. Therefore, hidden conditional random field (HCRF) is introduced as a rational choice. The significance of the HCRF is its generalized modeling, which resides in both the relaxation of the Markov property and incorporation with hidden states of the conditional random field (CRF) modeling.

In the following, an overview of the proposed system is first presented with a flowchart, followed by video representation using the BoW model and low-level genre categorization. Then, the proposed techniques are introduced, including unsupervised learning for middle-level view classification and HCRF for high-level event detection. Experimental results are then provided to demonstrate the effectiveness of the proposed method.

3.2 Overview

This section provides an overview from a holistic perspective as illustrated in Figure 3.1. The input video is analyzed systematically using a generic and sequential framework. This video is interpreted in a way such that the result from a preceding process is the input to the next process in a consistent and coherent fashion. There are four modules in total: module 0 is the infrastructure for low-level feature extraction and video representation using the BoW model. Module $1-3$ are tasks introduced in this thesis. The highlights of this framework include the following.

(1) A generic foundation using domain knowledge-free local feature was developed to represent input sports videos. This method fits the general framework in sports video analysis and provides an alternative solution to alleviate generality, scalability, and extensibility issues.

(2) A thorough and systematic structure starting from genre identification is presented, which was ignored in some related work that assumed the genre type as prior knowledge.

(3) A general platform is introduced to associate our method with the abundant and valuable existing literature, as well as various and innovative features input.

At module 0, the low-level local feature utilization incorporated with codebook generation and the BoW model provides an expandable groundwork for the semantic tasks of genre categorization, view classification, and high-level event detection. As our survey shows, the local feature is rarely explored in the domain of the sports videos, though it has been broadly adopted and proved effective in the field of computer vision. Most of the literature discusses domain knowledge and production rules at the feature extraction level. In our structure, a homogenous process is first introduced for extracting domain knowledge-independent local descriptors. The BoW model is used to represent an input video by mapping its local descriptors to a codebook, which is generated from an innovative bottom-up parallel structure. The histogram-based video representation is treated as the sole input (no other feature models) to both the genre categorization and the view classification modules. Such a concise representation built from the BoW model benefits

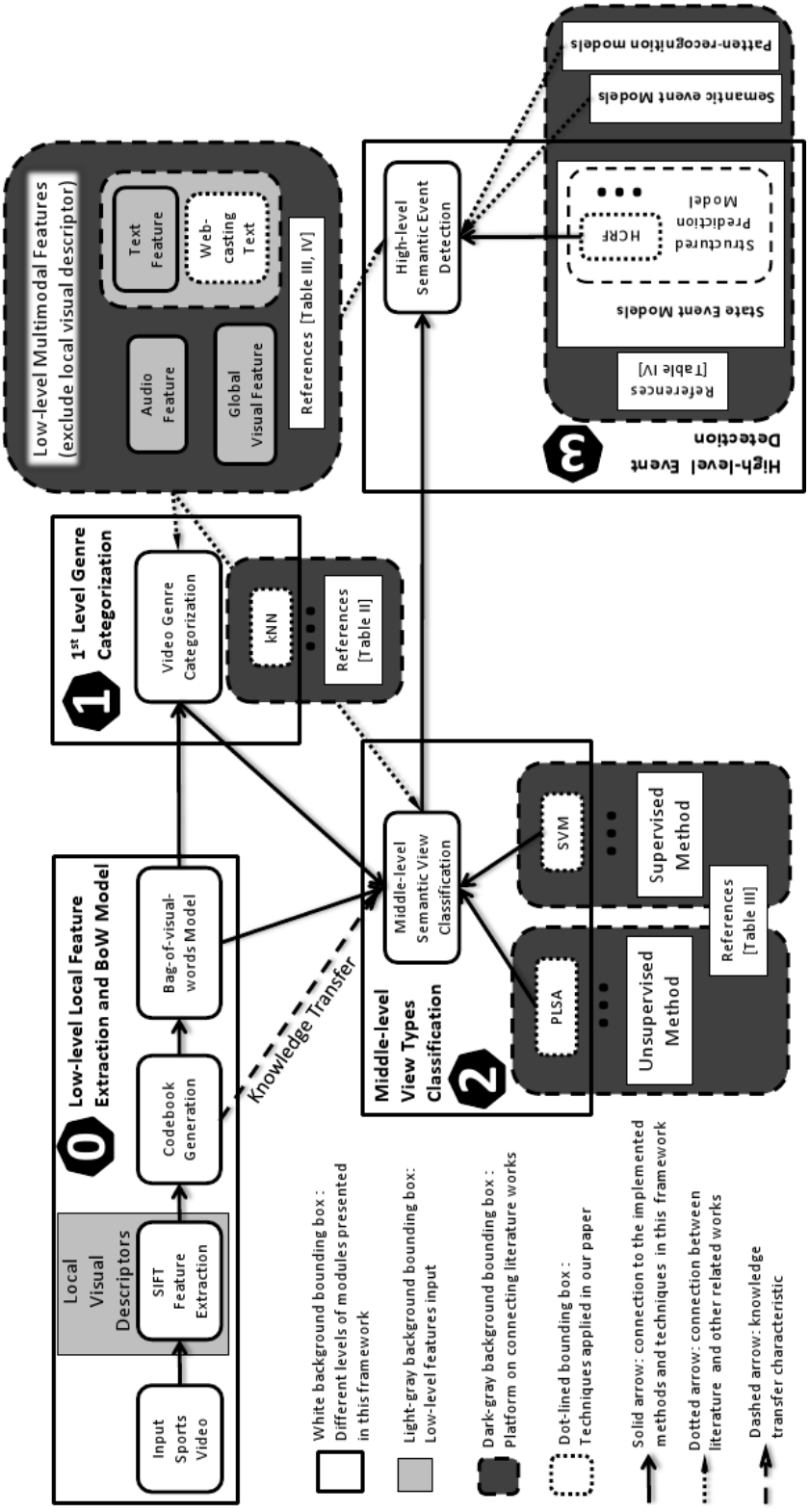


Figure 3.1: A flowchart of the proposed generic framework with one module of generic video representation and three task modules in sequence.

users in homogenously extracting visual features and representing videos in a compact and collective form.

In the 1st module, videos are categorized by genre. Video genre nomenclature is used to describe the video type, which is defined as the highest level of granularity in video content representation. Since the video genre categorization task directly relies on low-level features, the proposed feature extraction of the target video sequence is used in categorization. In large-scale videos, a successful identification of the genre serves as the first step before attempting higher level tasks. For instance, in sports event detection, an unknown "shooting" event is the target quest, which could be from a ball game or a shooting sport. By indiscriminately treating the entire dataset, this event will be searched through all types of sports. However, since sports like figure-skating and swimming have no "shooting" at all, the effort to search this event within those non-relevant sports becomes infeasible. Instead of treating all data indifferently, a more efficient method is to identify the genre of the query video first; and then, deploy middle/high-level tasks. As the survey shows in sports video analysis, most of the related works on view classification and event detection assume the genre by default. This framework, however, provides a system that automatically identifies the genre from various types of sports data before further analysis.

In the middle-level and the 2nd module, semantic view types are classified using an unsupervised PLSA learning method to provide labels for video frames. View describes an individual video frame by abstracting its overall content. It is treated as a bridge between low-level visual features and high-level semantic understanding. In addition, unsupervised learning saves a massive amount of human effort in processing large-scale data. Moreover, the supervised methods can also be implemented upon our proposed platform. Therefore, a SVM model is executed as the baseline for comparison.

Finally in the 3rd module, a structured prediction HCRF model using labeled inputs is a natural fit for the system to detect semantic events. This choice can be justified in that a video event occupies various length along the temporal dimension. Thus, the state event model-based HCRF is suitable to deploy. Less comprehensive baseline methods, such as the hidden Markov model and the conditional random field, can also be applied on this platform.

Besides the three-level modules in the *white background bounding boxes*, this framework, illustrated in Figure 3.1, also highlights the relationship between our system and

existing literature, which are shown in the *dark-gray background bounding box*. Associated Table references are also indicated in each module. Multimodal features excluding local visual features are also introduced at various stages by the literature. The *Dotted arrows* are used to represent these associations. The *solid arrows* denote the proposed and implemented techniques in our work. The *dashed arrow* represents a knowledge transfer characteristic of the generated codebooks. In summary, codebooks generated from certain sports with abundant resources, can be transferred and utilized in classifying other sports materials with scarce resources. The detail analysis is introduced in the Section 3.6.2.

In the following section, module 0 and module 1 are combined and presented, including feature extraction, bag-of-visual-words model, as well as genre categorization.

3.3 BoW-based Video Representation and Genre Categorization

This section covers the first part of our proposed framework, generic feature extraction with the BoW model, and systematic genre categorization. Figure 3.2 illustrates details of each process.

3.3.1 Feature Extraction

Local invariant features are chosen for homogenous feature extraction due to their domain knowledge-free properties. The scale, rotation, and illumination invariant properties make these descriptors good candidates in preserving the similarities for semantic objects and events matching and detection. Global features, on the other hand, rely on domain knowledge and have difficulties in robust concept and event detection, especially in the presence of noise and occlusion [58]. Scale-invariant feature transform (SIFT), developed by Lowe [2], is selected as feature descriptors in this work. The SIFT method extracts key-points of an image and describes these points using local neighborhood regional information. Since no prior and domain knowledge is required, SIFT is an ideal option in the large-scale automatic and homogenous process. By processing image sequences sampled from video clips, each frame is represented by a magnitude of hundreds of SIFT

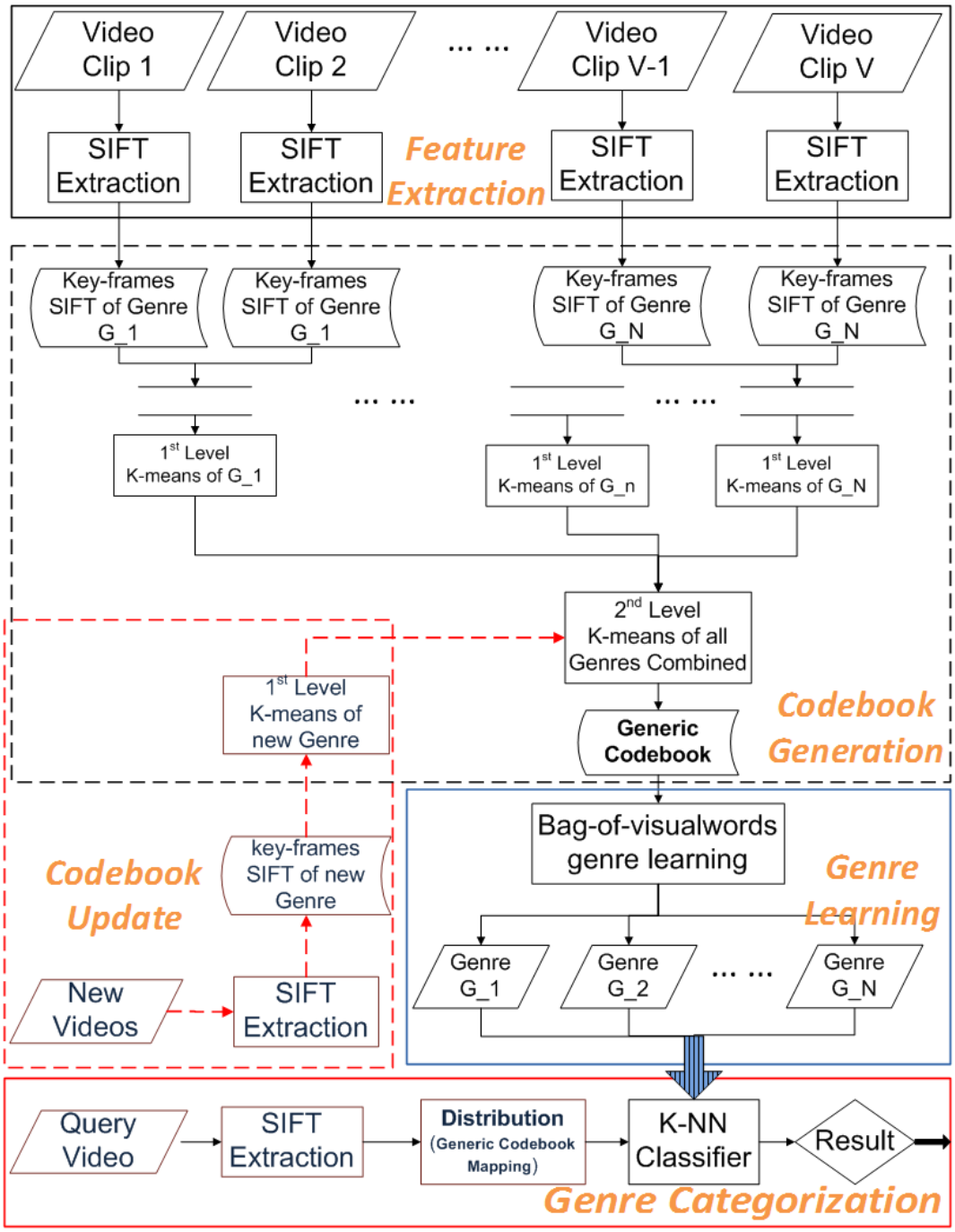


Figure 3.2: Feature extraction and genre categorization framework using data parallelism and bottom-up structure for codebook generation.

descriptors. After homogenous local descriptor extraction, the BoW model is applied, whose effectiveness relies on a robust codebook design. In order to achieve this resiliency, we propose a two-level bottom-up K-means clustering for codebook generation. The advantages of the bottom-up structure are efficiency, scalability, and robustness.

3.3.2 BoW Model with Two-level Bottom-up Codebook Generation

The BoW model is adopted by first synthesizing a representing codebook using codewords which are exemplars of combining sampled SIFT local descriptors. Consequently, a video clip is characterized by mapping its SIFT feature points to a generated codebook; and then, a histogram distribution is obtained. Compared to the original footage, this compact representation preserves enough information for differentiation, only using a small size in storage. In addition, random noise can be suppressed by using this proposed frequency-based histogram representation.

With the large-scale dataset, efficiency and robustness of the codebook formation have been important concerns for the BoW model. Heuristically, the larger the codebook size, the better the classification results (with certain saturation limitations) [77, 78]. Different codebook sizes have been explored, ranging from several hundred [79, 80] to thousands [17] to hundreds of thousands [77]. Since they all use different datasets, no conclusions have been drawn to make a standard rule. In this article, choices of codebook sizes are based on the empirical studies.

K-means clustering is utilized to generate a codebook by finding and appointing cluster centers as codeword values. In a large-scale domain, satisfactory performance has been reported using a top-down structure for categorization [81]. In that work, a two-layer top-down structure is used for sports genre categorization. At the first-layer, a general codebook (size 800) is generated using single K-means, in which a query video is only categorized to one of the predefined bigger groups consisting of several genres. Such a group is determined by those sports sharing similar semantics. At the second-layer after the membership of the bigger group is identified, an individual codebook (size 200) for this bigger group is used to decide the video genre. For instance, *judo* and *boxing* are combined into a bigger group named *martial arts*, where *martial arts* is used as the first-layer candidate. Subsequently, Judo and Boxing are differentiated in the

second-layer categorization. Although good classification accuracy has been reported, efficiency and robustness are problems for such a method in terms of creating a general codebook using single K-means clustering. This is because most computation of K-means lies in calculating the distances between individual points to their cluster centers in each iteration. A single K-means clustering using large-scale data is heavy in computation and sometimes inaccurate due to K-means own limitations. Since more than 3 million high-dimensional SIFT points are used for building the codebook in our application, one single K-means clustering becomes inefficient.

Therefore, a two-level bottom-up structure is proposed in this work for efficient codebook generation. At the bottom layer, individual genre codebooks are generated in 1st-level K-means clustering. At the upper layer, the 1st-level codebooks are used as the input for the 2nd-level K-means to build the generic codebook. By using this bottom-up structure, we reduce the heavy computation in measuring individual point-to-cluster-center distance in the K-means algorithm. Moreover, since the 1st-level K-means are independent from each other, distributed computing methods can be applied to further reduce the computation time. The numerical analysis is referred to in Section 3.6.1.

Another advantage of bottom-up K-means clustering resides in the system update and scalability. In the case of new genre videos added to the dataset, a codebook update module is applied to find the new genre's individual codebook. The result, together with existing codebooks, is used to generate the new generic codebook by only re-running the 2nd-level K-means. In the case that new videos are imported for an existing genre, the corresponding 1st level K-means is applied to achieve the updated individual codebook; and then, 2nd-level K-means is re-run to update the generic codebook.

3.3.3 Low-level Genre Categorization

Related work

Video genre and its categorization was one of the earliest video analysis which drew researchers' interests. The main task of this genre categorization starts from a diverse group of videos, such as sports, music, news, movies etc., and gradually moves to a more discriminating categorization such as identifying the sports genres. Various works have been highlighted as follows. However, a major and common disadvantage of these works is their heavy dependency on domain knowledge.

Fischer *et al.* [82] first proposed a classification method based on five different video genres. Brezeale and Cook [83] provided an extensive survey in this field. Incorporating the survey and most recent works, a concise summary is provided in Table 3.1. Color features with C4.5 decision trees were used in [84]. Camera motion features with statistical classifiers were chosen to classify six sports genre in [85]. A principal component analysis (PCA) modified audio-visual feature was used to train a Gaussian mixture model (GMM) classifier in [74]. Semantic shots (views) were used to help in genre categorization in [86]. Motion and color, as well as audio features, were applied in [87]. Color features with a hierarchical support vector machine (SVM) were used in [88]. High-level MPEG-7 features were extracted and applied in multi-modality classifiers in [89]. The best classification result at the moment has an accuracy of 95% using a dataset of eight different genres [90]. These methods used various domain knowledge with supervised classifiers to achieve the automatic genre categorizations.

As defined in [91], domain knowledge-based features can be divided into two categories, cinematic-based features and object-based features. The cinematic feature involves middle to high level semantics from common video composition or production rules such as shots/views or events, while object-based features are described by their spacial properties, such as color, shape, and texture, as well as spatial-temporal-based object motions. As Table 3.1 shows, all reviewed works are domain knowledge-dependent, either object-based or cinematic-based. A lack of diversity, that is, the number of different genres in the database, restricts these methods from generality.

Unsupervised genre categorization

In our proposed method, at the genre categorization stage, a query video is expressed as a histogram Q that also uses the generic codebook and the BoW model. Then, a k-Nearest Neighbor (k-NN) classifier is applied with a defined dissimilarity measurement between the query Q and a trained individual genre P. Consequently, the query video is identified as the genre whose distribution is closest to that of the query within measure. Technical details are presented in Section 3.6.1.

By identifying the genre of this query video, subsequent processes are confined to a focused group, and the scale of computation is decreased. Therefore, advanced and sophisticated techniques can be used in middle/high-level video analysis. In the next

Table 3.1: Summary of previous video genre categorization methods.

Authors and Year Published	Number of Genres	Size of Database (hrs)	Domain knowledge		Genre Categorization Method	Accuracy rate
			Object Based	Cinematic Based		
[84]	4	8	Yes	Yes	C4.5 decision tree	83%
[85]	6	33.75	Yes	Yes	statistics based	n/a
[74]	5	5	Yes	No	PCA & GMM	86.5%
[86]	4	n/a	Yes	Yes	decision tree and HMM	91.6%
[87]	3	16	No	Yes	pseudo-2D-HMM	n/a
[88]	6	33.33	No	Yes	hierarchical SVM	94%
[89]	5	5	Yes	Yes	Multimodel	88.5%
[90]	8	100	Yes	Yes	Parallel Neural Networks	95%

step, training data is characterized by frequency-based histogram representation. The individual genre is modularized as a distribution denoted by P using training data of its own kind.

3.4 BoW-based Unsupervised View Classification

Once a video genre is identified, the next step is to achieve view classification of each of the video frames in the query sequence. We present a literature review first, followed by the proposed unsupervised method.

3.4.1 Related Work

We summarize related works so that readers can compare popular supervised means with proposed unsupervised PLSA in this thesis. Additionally, there are only two works using unsupervised techniques based on our study. We present them for completeness of the review [92, 93].

Although there may be different nomenclatures, the fundamental purpose of the middle-level views(shots) is to involve certain production rules to aid in high-level tasks. This frame-based label concept was first introduced by Xu *et al.*, who defined three groups of views: global, zoom-in, and close-up [73]. Ekin and Tekalp [6] used a slightly different notation which includes long-shot, middle-shot, and close-up/out-of-field. Duan *et al.* [7] used a finer view/shot group classification, supported by innovative semantic features. These pioneering methods, along with other works such as [94–96] focus on using decision tree classifiers to link the low-level features to view/shot types. Xu *et al.* [73] and Ekin *et al.* [6] applied color-based grass detector and field/object size to determine view types. Incorporating previously mentioned features, Tong *et al.* [94] added head-area detection, as well as a grey-level co-occurrence matrix(GLCM) to improve the decision tree on classification. Wang *et al.* [95] used field region extraction, object segmentation and edge detection for view type decision making. Duan *et al.* [7] first extended the research from single genre (soccer) to multiple genres (four sports) using individual genre-based decision trees. Different from previous visual feature extraction methods, Kolekar and Palaniappan [96] took a top-down approach. They first used audio features to find exciting video clip. The motion features of the whole image volume along with the

background color information are then utilized for view-type classification. Benmokhtar *et al.* [97] took an approach on feature level fusion using dynamic PCA with information coding neural-network (NN). At the classification level, another NN is used to fuse multi-modality inputs. However, these supervised methods are limited by the labeled data; and thus, constrained from being expanded to larger scales.

Some other researchers pursued unsupervised methods for view classification. Wang *et al.* [92] proposed an information-theoretic co-clustering method, in which mutual information was maximized by treating shot classes and features as two random variables. As a consequence, color histogram and perceived motion energy features are used with a test set of four sports video genres. Zhong *et al.*'s method was inspired from spectral theory conventionally used to solve segmentation problem in graph theory [93]. They proposed a spectral-division algorithm to find the proper video shot clustering, which were tested in three sports videos using the HSV space color feature. Although good performances have been obtained in these methods, the extensibility and flexibility towards diverse genres and large-scale datasets are very limited. This limitation is again due to the domain knowledge dependency of the extracted features.

Table 3.2 compares the aforementioned methodologies from angles of feature utilization and classification techniques. Color and texture are two major global features used by most works. Duan *et al.*'s work is the only one that proposed middle level features developed from low-level global features. The rest of the work either adopted additional popular global feature schemes, such as audio feature or Gabor feature, as well as some production rule-based features, or did not utilize any. While various global features are used, none of the local features have been applied. Moreover, most of the supervised methods (except Duan *t al.*'s work) focus on a single (soccer) sport, while unsupervised techniques use various types of sports.

3.4.2 Unsupervised View Classification

This section introduces the middle-level view classification, where the previously built BoW model is also used as feature representation. Since this work targets large-scale videos, an unsupervised solution is more viable and applicable. Therefore, we chose to use unsupervised probabilistic latent semantic analysis (PLSA)-based models. PLSA has demonstrated promising results in analyzing co-occurrence data of words and documents

Table 3.2: Comparison of view classification techniques in literature, emphasizing on features utilization and classification methods.

Authors and Year Published	Nature of data	Global Features			Local Feature Based	View Classification Method
		Color Based	Texture Based	Others (yes/innov)		
[73]	Soccer	Yes	No	No	No	thresholding (S)
[6]	Soccer	Yes	No	Yes	No	morphological operations (S)
[7]	4 Sports	Yes	Yes	innov	No	Decision Tree (S)
[94]	Soccer	Yes	Yes	Yes	No	Decision Tree (S)
[92]	4 Sports	Yes	Yes	Yes	No	spectral clustering (UnS)
[97]	Soccer	Yes	Yes	Yes	No	Neural-network (S)
[93]	3 Sports	Yes	No	No	No	Spectral-division algorithm (UnS)
[96]	Soccer	Yes	No	Yes	No	Decision Tree (S)

Note: In the "Global Features" column with "Others (yes/innov)" category: "yes" means other than color and texture global features are used while not innovative, while "innov" means newly designed features are used. For the "View Classification Method" column, S indicates an supervised method, while UnS indicates the unsupervised method.

in text retrieval [98]. From a matrix factorization point of view, PLSA belongs to a subgroup called non-negative matrix factorization, where the factorized matrices are non-negative [99]. Because the codebook paradigm with codewords is adopted in mapping visual features to a probability-based histogram which has to be non-negative, PLSA becomes a more suitable selection compared to other factorization techniques, such as singular value decomposition or principle component analysis.

PLSA relies on the likelihood function of multinomial sampling and aims to reach an explicit maximization of the predictive power of the model. Incorporating the PLSA plate notation in Figure 3.3 with the view classification application, the observed state w is defined as codewords with a predefined codebook of size M. An individual video frame is denoted by d with a total number of training frames N. Latent state z is the view type and parameter K is the total number of view classes, and in this work, K equals four. The likelihood function is given in Equation (3.1). The probabilistic distribution is defined as $p(w_i|d_j)$, where w_i is an individual codeword, and d_j is a training frame. Such distribution can be represented by a sum-of-product of two distributions, $p(w_i|z_k)$ and $p(z_k|d_j)$. The former is interpreted as an impact on codewords by a view type, while the latter is the probability of a particular view type given a training frame. The number of codeword w_i appearing in a frame d_j is denoted as $n(w_i, d_j)$. The argument of maximum posterior (MAP) estimate z^* is optimized by using an expectation maximization (EM) as shown in Equation (3.2).

$$\begin{aligned} L &= \prod_{i=1}^{M}\prod_{j=1}^{N} p(w_i|d_j)^{n(w_i,d_j)} \\ &= \prod_{i=1}^{M}\prod_{j=1}^{N} \Big(\sum_{k=1}^{K} p(w_i|z_k)p(z_k|d_j)\Big)^{n(w_i,d_j)} \end{aligned} \tag{3.1}$$

$$z^* = \arg\max_z p(z|d) \tag{3.2}$$

Since SVMs have demonstrated great performance in the field of classification, it is adopted in our view classification task for comparison purposes. In general, supervised models tend to yield better results but require predefined knowledge. A typical radial basis function (RBF) is used as the non-linear kernel in SVM [100] and shown in Equation

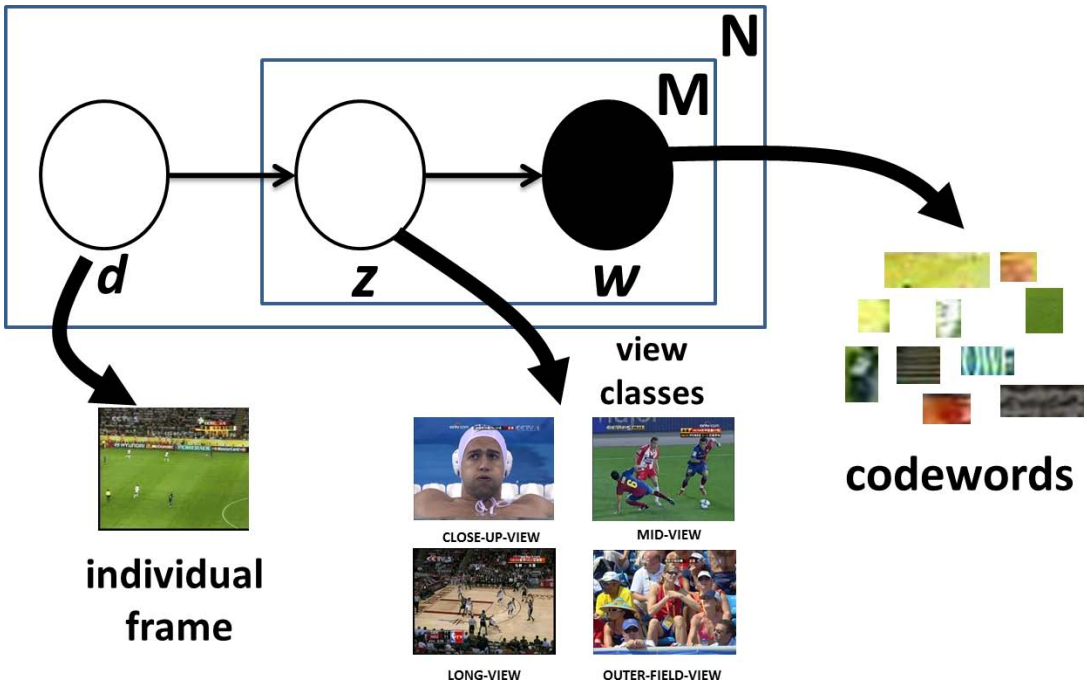


Figure 3.3: Illustration of the PLSA model in plate notation and its connection with view type classification.

(3.3). In this equation, x_i and x_j represent the codewords, and γ is the kernel parameter of the RBF.

$$K(x_i, x_j) = \exp\left(-\gamma \left\|x_i - x_j\right\|^2\right), \quad \gamma > 0. \tag{3.3}$$

Four view types are defined, namely close-up-view, mid-view, long-view and outer-field-view. This definition is also popular among other work in this field [6, 73, 101]. For the PLSA-based model, the number of view types is required, while labeling effort is not needed for individual frames. On the contrary, SVM-based models demand both semantic predefined view types as well as all frames labeled with groundtruth, which could be unaffordable when the video is large in size.

As a result of the view classification task, the query video sequence is labeled with view types. In the next section, models which take labeled video sequence as input for detecting interesting events are introduced.

3.5 High-Level Event Detection

Content-based video event detection is among the most popular quest for high-level semantic analysis. Different from video abstraction and summarization, which targets any interesting events happening in a video rush, event detection is only constrained to a predefined request type (such as the third goal or the second penalty kick in a particular soccer match). In sports videos, a consumer's interest in events resides in the actual video contents, more than just the information delivered. For instance, a user wants to watch particular goals in basketball games, or replays in soccer matches. S/he is not only interested in the information like who/how/what, but more importantly, the visual contents rendered from the sports clips. On the other hand, sports videos also have very strongly correlated temporal structures. In a way, the structure can be interpreted as a sequence of video frames which have patterns and internal connections. This pattern is ubiquitous due to the nature of sports, a competition where players learn from the standard in order to excel. Therefore, an intuitive approach is to find such patterns using certain representation; and in turn, to learn the temporal structure. Luckily, the PLSA algorithm provides such a labeled frame sequence. What we need is a clever technique to analyze portions of the video and determine what structured prediction model to use. In the following, we will first review the literature. Then, we will introduce a coarse-to-fine scheme and hidden conditional random field (HCRF) for event detection.

3.5.1 Related Work

As one of the most popular semantic tasks in video analysis, event detection has been a popular topic from the beginning of multimedia research. Despite different definitions of event detection by different researchers, commonly acknowledged properties of an "event" can be summarized as follows. An event occupies a period of time and is described using salient aspects of the video sequence input, which consists of smaller semantic units or building blocks [102]. Lavee *et al.* also summarized and classified event detection algorithms into three categories: a) pattern-recognition models, b) semantic event models, and c) state event models. Pattern-recognition models focus on direct classification from low-level features, but lacks semantic linkage. Semantic models target high-level semantic rules and constraints with domain-knowledge. These models require a lot of human in-

volvement in creating rules and regulations using prior information. State models utilize abstracted middle-level agents, as well as the intrinsic structure of the event itself.

By comparing these three categories of event modeling with examples in the literature, we think that the pattern-recognition model is heavily dependent on classifiers, which at the moment, are not intelligent enough to understand all semantics from low-level features. On the other hand, the semantic model considerably relies on human expertise; and thus, underestimates the accuracy and efficiency provided by classification tools. From our experience, the state model incorporates the strength of pattern recognition at low-level with classifiers at high-level so that it utilizes both feature extraction power and classification intelligence. Moreover, the state model also accommodates an automatic process and unsupervised learning, which reduces human input into the system. Therefore, state event models are suitable for analyzing large-scale datasets, from both generic and systematic point of views. A coarse-to-fine strategy fits well into such state event models, by first roughly localizing the event with context information and then precisely detecting the event using an advanced structure model.

Although we prefer the state event model for its natural fitness to the proposed systematic approach in this work, two other models are still valued for their efficiencies in analyzing sports videos and utilizations in applications. In the following, state-of-the-art algorithms are summarized and compared.

Support vector machine (SVM) is a popular pattern-recognition model algorithm [102]. Some groups use rich audiovisual features, such as face detection, scoreboard information, and, geometry of the field, to find certain semantic events. Saldier and O'Connor [103] used SVM to classify "scoring" events for four different field sports. Xu *et al.* [104] analyzed tennis videos by using hierarchical-SVM applied on fused audio-visual modalities. Similarly, Ye *et al.* [105] utilized middle-level view labels as well as shot length and camera motion descriptors. An SVM-based incremental learning scheme using updated data is proposed in detecting soccer events, along with a predefined temporal structure. A similar method combining SVM and predefined temporal structure was proposed by Li *et al.* [106], targeting basketball events using optical flow patterns.

Some semantic event models using rules and logic and semantic relationships are presented. Babaguchi *et al.* [107] used closed caption text streams with audiovisual features and the intermodal correlation between them to search a "touch down" event from four hours of American football videos. Zhang *et al.* [108] also focused on superimposed cap-

tion frames and used decision trees to decide the event, such as "scoring" or "last pitch" for baseball games. Ekin *et al.* [91] incorporated production rules and soccer sport rules to detect certain events such as "goal", "referee", and "penalty-box".

In terms of state event models, one of the earliest works targeting structures of videos was from Nepal *et al.* [75], who empirically studied the temporal model in basketball videos based on manual observation, using heuristic methods and low-level audio-visual features. Duan *et al.* [101] also generated a temporal structure using multimodality with heuristic experience on tennis events. Another approach of learning temporal structure is from the data mining perspective, where Tien *et al.* [109] focused on a tennis match event detection by creating a max-subpattern tree and learning the frequent patterns from it.

Another important branch of state event models are structured prediction models such as hidden Markov models (HMMs) and their variations, Bayesian networks, as well as discriminative conditional random fields (CRFs). Zhang *et al.* [110] proposed an HMM-based statistical method for classifying middle-level agents generated from web-casting texts. Tong *et al.* [111] used Bayesian networks to classify "shoot" and "card" events in soccer videos, by applying decision tree-based intermediate-layer concept units. Mei and Hua [112] proposed an innovative mosaic-based middle-agent for key-event mining using HMMs. Wang *et al.* [113] proposed a CRF model on detecting semantic soccer events, and the performance turned out to be better than both SVM and HMMs. A similar algorithm was also proposed by Xu *et al.* [114] using CRFs for basketball and soccer event detection where a webcast text feature was obtained to achieve middle-level concepts. An interesting event tactic analysis is proposed by Zhu *et al.* [76], which is beyond the conventional event and adopts the cooperative nature and tactic patterns of team sports. Extensive experiments have been conducted on soccer.

Table 3.3 provides a comparison of the aforementioned literature from a feature utilization point of view. Most of the methods utilize multimodality schemes of features input. By comparing the number of events processed, it appears that the state event model has better scalability in examining various event scenarios. It is also interesting to point out that local visual features have not been utilized in any of the methods. In addition, many of the methods, especially state event models, require middle-level semantic agents to bridge the gap between the low-level features and the high-level events. Such middle-level agents have to be labeled data. However, for the generic method presented

in this work, we tackle event detection problem using the input obtained by unsupervised learning and unlabeled data.

3.5.2 Hidden Conditional Random Field (HCRF) Model

Before learning the temporal patterns, a starting and entry point of an event needs to be seized. A two-stage coarse-to-fine event detection strategy is suitable for this scenario. The first stage is a rough event recognition and localization utilizing rich and accurate text-based information either from web-casting text or optical character recognition (OCR) techniques of the scoreboard update. In the second stage, precise video contents associated with the semantic event have been detected in terms of event boundary detection and accuracy analysis. The coarse-to-fine techniques have been proven effective and accurate [115]. Web-casting text for coarse-stage event detection and video alignment was studied and analyzed such as replaying scenes and various goal and shot scenes detection in soccer video [116, 117].

Since the proposed framework targets the generic learning model that can be extended to large-scale datasets, we rely on visual content, that is, the local features extracted and middle-level views classified from such features. To demonstrate the effectiveness of the proposed model, we focus on a particular basketball score event detection. We adopted the previously developed scoreboard update detection method for a coarse-stage process in order to obtain the time-stamp [115]. The fine-stage process focuses on robust and accurate visual content detection from the score event. The video sequence is analyzed by distinguishing the actual score event from false alarm events, such as timeouts or intermission, which are also concurrent with scoreboard information. We propose a HCRF-based structured prediction model utilizing previously classified views, thereby completing the generic approach. For example, the HCRF model can be used to detect the score event in basketball for exciting events and highlights. Such an HCRF technique belongs to the state event model defined in related works. Therefore, HCRF takes the labeled sequences as input in a natural and seamless fashion. On the other hand, HCRF is a comprehensive model which can be degraded to hidden Markov models (HMM) or conditional random fields (CRF) with certain constraints. The merits of HCRF compared with the other two models are its resilience and robustness with a combination of both the hidden states and the Markov property relaxation. Technical details are examined

Table 3.3: Comparison of event detection models emphasizing feature utilization from both low-level features and middle-level semantic agents.

Event-Detection Algorithm Category	Authors and Year Published	Nature of data	Number of Events	Low-level Multimodal Features	Visual Features		Middle-level Semantic Agents
					Global-Based	Local-Based	
Patten-Recognition Model	[104]	Tennis	5	AVM	Yes	No	Yes
	[103]	4 field sports	2	AVS	Yes	No	No
	[105]	Soccer	1	n/a	n/a	n/a	Yes
	[106]	Basketball	5	VM	Yes	No	No
Semantic Event Model	[107]	Football	3	VST	Yes	No	No
	[108]	Baseball	2	VT	Yes	No	No
	[91]	Soccer	3	VS	Yes	No	Yes
State Event Model	[75]	Basketball	1	AVMT	Yes	No	No
	[101]	Tennis/Soccer	16	AVMT	Yes	No	Yes
	[111]	Soccer	2	VM	Yes	No	Yes
	[113]	Soccer	5	AVM	Yes	No	Yes
	[110]	Basketball	5	VT	Yes	No	Yes
	[109]	Tennis	4	AVS	Yes	No	No
	[112]	Soccer	3	VM	Yes	No	Yes
	[114]	Soccer/Basketball	17	VTS	Yes	No	Yes
	[76]	Soccer	6	VMTS	Yes	No	Yes

Note: In the "Low-level Multimodal Features" column, various features are utilized, including audio (A), visual (V), text (T), motion feature (M), and video shot detection (S), as well as an "n/a" label in the case when no low-level feature mentioned in the related works.

in the following.

There are several advantages of using HCRF in large-scale datasets, rather than HMM, or CRF models. First, HCRF relaxes the Markov property, which assumes that the future state only depends on the current state. In our generic framework, video frames are uniformly decimated and sampled, regardless of the temporal pace of the video itself. In some cases, several consecutive frames have the same labeling, while in other cases, different labels are assigned. Markov property-based models such as HMM are appropriate for the former scenarios, but not suitable for the latter ones, since the future state in HMM only cares about the current state label, but not previous states. On the other hand, HCRF is flexible and takes surrounding states from both before and after the current state. Thus, HCRF is more robust for dealing with large-scale homogeneous processes and uniform sampling with no prior knowledge. For instance, if a key frame immediately preceding the current state is missed due to uniform sampling, such information loss could be compensated by including and summing up distant informational frames (both previous and future) from uniform sampling without misclassifying the event.

Second, HCRF has merit in its hidden states structure, which helps to relax the requirement of explicit observed states. This relaxation property is also an advantage in dealing with large-scale uniformly sampled video frames. It is because of this configuration, CRF model outputs individual result labels (such as event or not event) per state and requires separate CRFs to present each possible event [114]. In HCRF, only one final result is presented in terms of multi-class events occurring probabilities. From the point of view of robustness, a CRF model can be easily ruined by semantically unrelated frames due to automatic uniform sampling. A multi-class HCRF, on the other hand, can correct the error introduced by such unrelated frames using probability-based outputs [118].

Moreover, HCRF is also appealing for allowing the use of not explicitly labeled training data with partial structure [118]. From the literature, HCRF has been successfully used in gesture recognition [118, 119] and phone classification [120].

Figure 3.4(a) illustrates an HCRF structure in which label $y \in Y$ of event type is predicted from an input $\mathbf{X}$. This input consists of a sequence of vectors $\mathbf{X} = \mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_m, ..., \mathbf{x}_M$, with each $\mathbf{x}_m$ representing a local state observation along the HCRF structure. In order to predict y from a given input $\mathbf{X}$, a conditional probabilistic model defined in [118] and in Equation (3.4) is adopted. In the equation, model parameter θ is used to describe the

local potential function ψ, which is expanded in Equation (3.6). A sequence of latent variables $\mathbf{h} = h_1, h_2, ..., h_m, ..., h_M$ are also introduced in Equation (3.4), which are not observable from the structure of Figure 3.4(a). Each h_m member of $\mathbf{h}$ corresponds to a state of s_m. The denominator $Z(\mathbf{X};\theta)$ is the normalization factor, which is expanded in Equation (3.5).

$$P(y|\mathbf{X},\theta) = \sum_{\mathbf{h}} P(y,\mathbf{h}|\mathbf{X},\theta) = \frac{\sum_{\mathbf{h}} e^{\psi(y,\mathbf{h},\mathbf{X};\theta)}}{Z(\mathbf{X};\theta)} \tag{3.4}$$

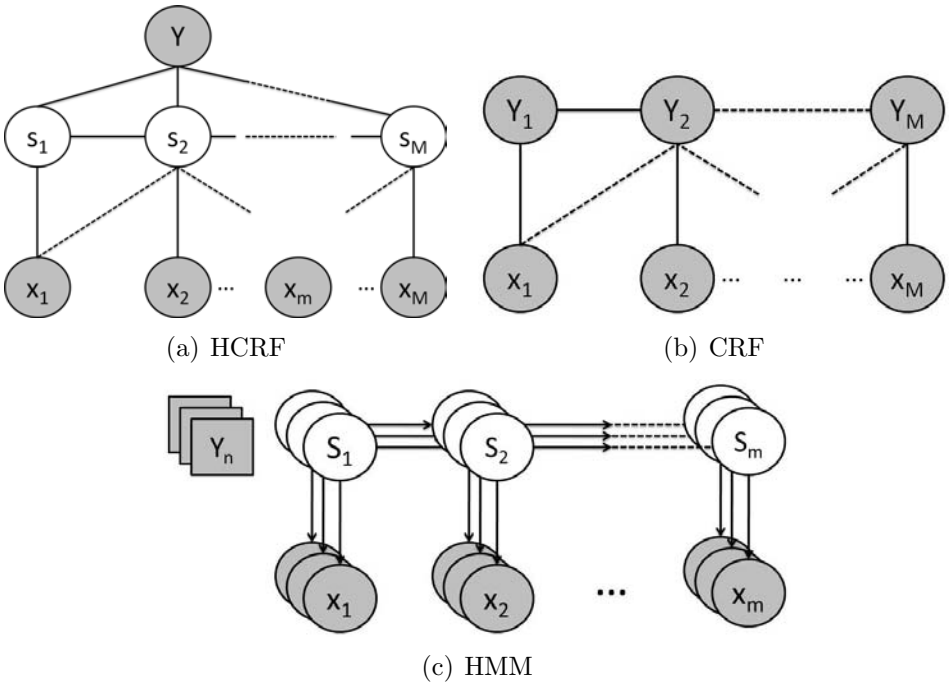
$$Z(\mathbf{X};\theta) = \sum_{y',\mathbf{h}} e^{\psi\left(y',\mathbf{h},\mathbf{X};\theta\right)} \tag{3.5}$$

$$\psi(y,\mathbf{h},\mathbf{X};\theta) = \sum_t \sum_k \theta_k^1 f_k^1(y,h_t,\mathbf{X}) + \sum_t \sum_k \theta_k^2 f_k^2(y,h_{t-1},h_t,\mathbf{X}) \tag{3.6}$$

In the event detection application, each $\mathbf{x}_m$ from $\mathbf{X}$ is a vector descriptor called local observation. In the notation, the $\mathbf{x}_m$ value at a time t is defined as $\mathbf{x}_m(t) = [p_{ws_1}(t), p_{ws_2}(t), p_{ws_3}(t), p_{ws_4}(t), p_{wc}(t)]$, with each entry of $\mathbf{x}_m(t)$ calculated from an average result of a sliding window centering at time t, as Figure 3.5 shows. The first four entries of $\mathbf{x}_m(t)$ are the probabilities of four possible view types, where $p_{ws_{j=1,2,3,4}}(t)$ associates with close-up-view, mid-view, long-view, and outer-field-view by $j = 1, 2, 3, 4$ respectively. The fifth $p_{wc}(t)$ value is an associated directional motion descriptor, introduced by Tan *et al.* [121]. The formula to calculate the average values at time-stamp t are given in Equation (3.7), where individual frame-based probabilities are $p_{s_{j=1,2,3,4}}$ and p_c.

$$\begin{aligned} p_{ws_j}(t) &= \frac{1}{N} \sum_{\tau=t-N/2}^{t+N/2} p_{s_j}(\tau) \quad with \;\; j = 1,2,3,4 \\ p_{wc}(t) &= \frac{1}{N} \sum_{\tau=t-N/2}^{t+N/2} p_c(\tau) \end{aligned} \tag{3.7}$$

A label and training sequence pair is defined as $(y_i, \mathbf{X}_i)$ with the index number $i = 1, 2, ..., n$. For each pair, $y_i \in Y$ and $\mathbf{X}_i = \mathbf{x}_{i,1}, \mathbf{x}_{i,2}, \mathbf{x}_{i,m}, ..., \mathbf{x}_{i,M}$ are the event label and observed states as Figure 3.4(a) depicts. For instance, $\mathbf{x}_{i,m}$ is interpreted as the



(a) HCRF (b) CRF

(c) HMM

Figure 3.4: Structured prediction models: (a) hidden conditional random field (HCRF); (b) conditional random field (CRF); (c) hidden Markov model (HMM).

m^{th} sampled time state of the i^{th} training sequence, where $\mathbf{x}_{i,m}(t) = [p_{i,ws_1}(t), p_{i,ws_2}(t), p_{i,ws3}(t), p_{i,ws4}(t), p_{i,wc}(t)]$.

During HCRF training, parameters θ_k^1 and θ_k^2 need to be learned. As Equation (3.6) shows, θ_k^1 and θ_k^2 are coefficients for the state feature function f_k^1, which contains a single hidden state, and the transition feature function f_k^2, which involves two adjacent hidden states, respectively. In order to find the optimal parameters, a log-likelihood objective function is used, as shown in Equation (3.8), with a shrinkage prior (the second term in the equation) in order to avoid the excessive parameter growth. A limited-memory version of the Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) quasi-Newton gradient ascent method [122] is applied to find the optimal $\theta^* = argmax\pounds(\theta)$. The L-BFGS algorithm is chosen due to this method's efficiency and performance from both theory [123] and application [114].

During the optimization process, the conditional probability in Equation (3.8) is substituted by the explicit form in Equation (3.4) to get Equation (3.9). Then, partial derivatives of a training sample $\pounds_i(\theta)$ with respect to θ_k^1 and θ_k^2 are derived in Equations

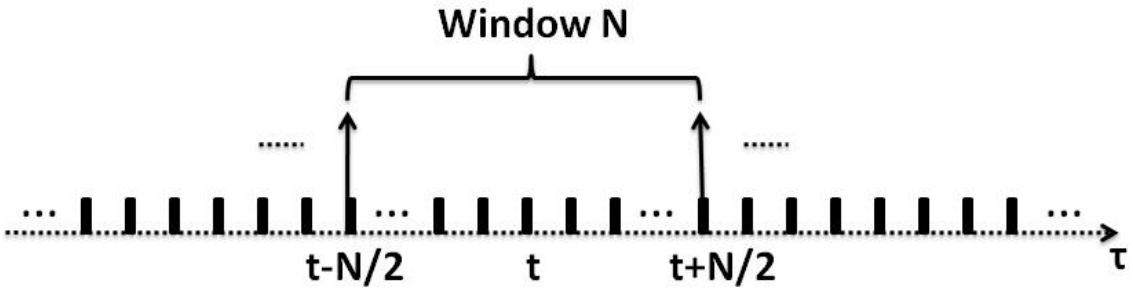


Figure 3.5: HCRF input shown in Equation (3.7), by sliding window average result on view types of decoded image sequence.

(3.10) and (3.11), respectively.

$$\pounds(\theta) = \sum_i \log p(y_i|\mathbf{X}_i, \theta) - \frac{1}{2\delta^2}\|\theta\|^2 \tag{3.8}$$

$$\pounds(\theta) = \sum_i \log\left(\frac{1}{Z(\mathbf{X}_i;\theta)}\sum_{\mathbf{h}} e^{\psi(y_i,\mathbf{h},\mathbf{X}_i;\theta)}\right) - \frac{1}{2\delta^2}\|\theta\|^2 \tag{3.9}$$

$$\begin{aligned}\frac{\partial \pounds_i(\theta)}{\partial \theta_k^1} &= \sum_t P(h_t|y_i,\mathbf{X}_i) f_k^1(y_i,h_t,\mathbf{X}_i) \\ &\quad - \sum_{t,y'} P(h_t,y'|\mathbf{X}_i) f_k^1(y',h_t,\mathbf{X}_i)\end{aligned} \tag{3.10}$$

$$\begin{aligned}\frac{\partial \pounds_i(\theta)}{\partial \theta_k^2} &= \sum_t P(h_{t-1},h_t|y_i,\mathbf{X}_i) f_k^2(y_i,h_{t-1},h_t,\mathbf{X}_i) \\ &\quad - \sum_{t,y'} P(h_{t-1},h_t,y'|\mathbf{X}_i) f_k^2(y',h_{t-1},h_t,\mathbf{X}_i)\end{aligned} \tag{3.11}$$

3.5.3 Comparison with Conditional Random Field (CRF) and Hidden Markov Model (HMM)

For comparison purposes, we also utilized conventional CRF models as depicted in Figure 3.4(b). By following definitions in [124], the conditional probability function is shown in Equation (3.12), with the normalization factor in Equation (3.13). The potential function

is defined in Equation (3.14), where $v_j(Y_{t-1}, Y_t, \mathbf{x})$ is a transition feature function between state positions t and $t-1$ within the observation sequence; while $s_k(Y_t, \mathbf{x})$ is a state feature function at state position t. Parameters λ_j and μ_k are estimated for transition and state feature functions, respectively.

$$P(\mathbf{Y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \cdot exp\left(\sum_{t=1} F(\mathbf{Y}, x, t)\right) \tag{3.12}$$

$$Z(\mathbf{x}) = \sum_{Y'} exp\left(\sum_{t=1} F(Y', \mathbf{x}, t)\right) \tag{3.13}$$

$$F(Y, \mathbf{x}, t) = \sum_j \lambda_j v_j(Y_{t-1}, Y_t, \mathbf{x}) + \sum_k \mu_k s_k(Y_t, \mathbf{x}) \tag{3.14}$$

The HMM algorithm is also provided in Equation (3.15) and depicted in Figure 3.4(c).

$$\begin{aligned} P(Y|X) &= P(X,Y)/P(X) \\ &= \prod_t P(X_t|Y_t) \cdot P(Y_t|Y_{t-1}) \end{aligned} \tag{3.15}$$

The aforementioned three structured prediction models use different decision-making schemes for the final event detection. For the HMM, the query sequence is tested. The highest likelihood of the HMM provides the final decision in event detection. On the other hand, in the CRF model, since each state variable $Y(t)$ requires a label, as Figure 3.4(b) shows, a majority-rule voting scheme in which the most event labels along the Y sequence decide the event result. For the HCRF model depicted in Figure 3.4(a), a multi-class training process recognizing all classes at the same time is adopted. Therefore, a detected event with the highest probability is considered the final result for the query sequence.

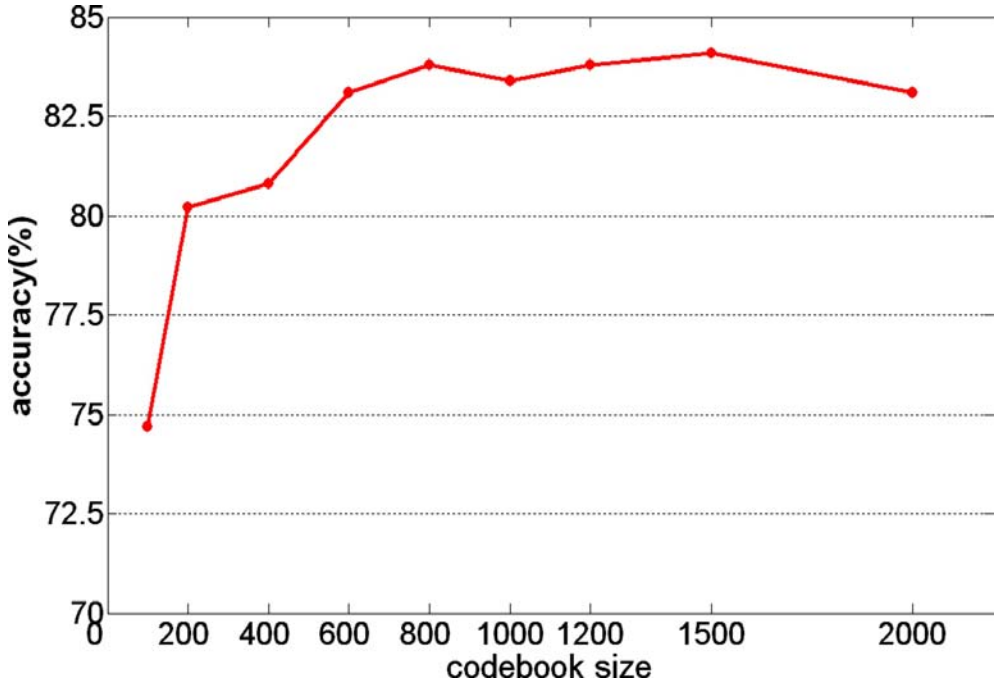
3.6 Experiments and Results

In the following section, experimental results are presented to justify the properties of the proposed generic framework, specifically using a relatively large-scale video collection

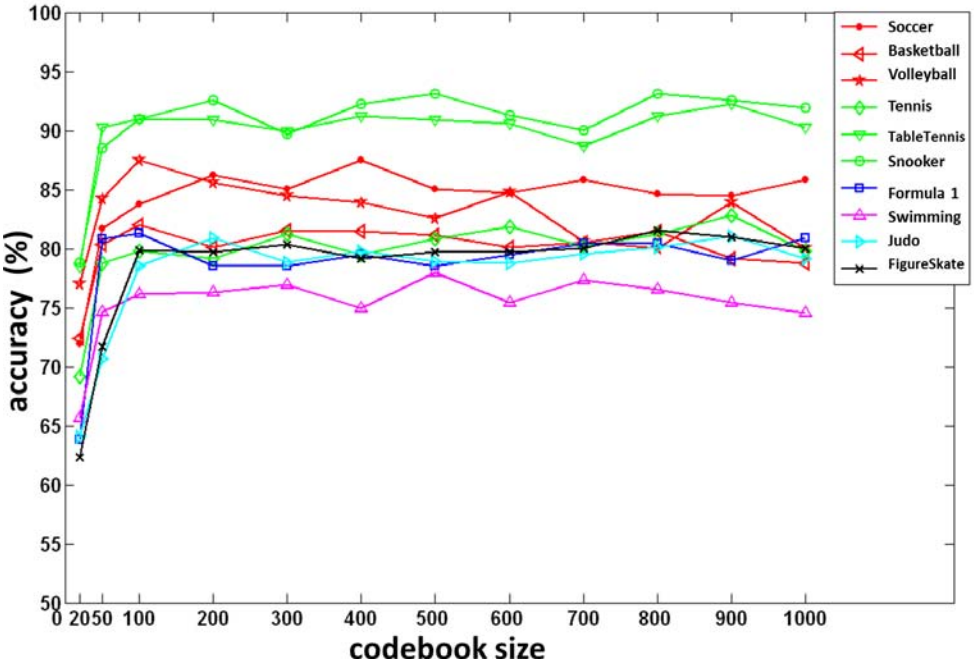
that includes 23 genres with a total of 145 hours gathered by the authors and his co-workers, named the 23-sports dataset. To our best knowledge, this dataset is the most diverse in video genres, collected from both the internet and television. All the video clips have the same length of 167 seconds with a total of 500 uniformly sampled frames at a sampling rate of three frames per second. This dataset is composed with 3,122 clips. In training, 1,198 clips are used, in which a subset of 46 clips (2 clips per sport) are used in codebook generation with a total of 3,112,341 SIFT points. In testing, the other 1,924 clips are selected.

Various codebook sizes were studied at first. Then, the proposed system was evaluated in three experiments, with a particular event detection as its ultimate measurement: (1) genre categorization using the proposed bottom-up codebook generation is analyzed; (2) view classification results are assessed and compared using both supervised and unsupervised classifiers; (3) finally, the coarse-to-fine event detection is examined by investigating the basketball score event. The validity on the score event detection can be extended to other event scenarios with labeled video sequences. The detailed argument can be found in Section 3.6.3.

To investigate the codebook size effectiveness, a subset of the 23-sports dataset of 14 sports was used. The clip numbers of these sports range from 70 to 106, averaging 87, while each individual clip is a uniform 167 seconds in length. Two experiments were conducted on the codebook size selection for genre categorization and view classification, respectively. For genre categorization, the average accuracy performance of all sports as a function of different codebook sizes is shown in Figure 3.6(a). The plot reaches a plateau after codebook size 800, and starts to drop at codebook size 1,500. For view classification, the accuracies of individual sports as a function of different codebook sizes are shown in Figure 3.6(b). Although various accuracy levels are observed for each sport, the individual performance follows a similar plateau trend. Based on these empirical studies, it is concluded that the performances are proportional to codebook sizes, with stable results at codeword ranges of 800–1500 and 800–1000 for genre categorization and view classification, respectively. This study is also consistent with existing research [58,77,78]. In the following experimentation for genre categorization with a total of 23 sports types, it is predicted that the codebook size should be bigger than in the tested 14 sports case. Therefore, a codebook size of 1,600 is chosen, and a codebook size of 800 is also applied as a comparative analysis. For view classifications involving 14 sports, a codebook size



(a)



(b)

Figure 3.6: Empirical studies on codebook size selection. (a) Average sports accuracy performance for genre categorization. (b) Individual sport accuracy performance for view classification.

of 800 is selected.

3.6.1 Genre Categorization Using a K-nearest Neighbor (k-NN) Classifier

In genre categorization, a K-nearest neighbor (k-NN) classifier is applied. Three different dissimilarity measurements are compared, including Euclidian distance (ED), earth mover's distance (EMD), and Kullback-Leibler divergence (KL-div). ED is used for measuring the spatial distance in Euclidian space in between two histograms. EMD is a distance function for achieving the minimal cost in transforming one histogram into the other [125]. The KL-div is a non-symmetric measurement between two probability distributions Q and P defined as $D_{KL}(Q||P) = \sum_i q_i \cdot \ln(q_i/p_i)$ [126]. In this work, q_i and p_i are individual codewords for the query video Q and the trained genre model P, respectively.

Before accuracy performance analysis on genre categorization, codebook generation schemes are examined by comparing both the proposed two-level bottom-up (BU) structure and the baseline single K-means (SK) clustering method [126]. As pointed out by Jain *et al.* [127], K-means clustering is considered a partitional algorithm using the squared error to reach the optimum solution. The sum of squared errors (SSE) is a widely used criterion function for clustering analysis, which quantitatively measures the total difference between all individual points to their clustering centers [126]. An SSE deviation percentage δ_{dev} is defined in Equation (3.16). Let ξ_{BU} and ξ_{SK} represent the SSEs of the bottom-up clustering and the single K-means clustering at the end of each algorithm, respectively. The numerator is the absolute value of the difference between ξ_{BU} and ξ_{SK}, and the denominator is ξ_{SK}. As Table 3.4 shows, the SSE deviation percentages at codebook sizes of 800 and 1,600 are 1.4% and 3.7%, respectively. Thus, we can conclude that in using the bottom-up structure instead of the single K-means clustering for codebook generation, the deviation of SSE is trivial.

$$\delta_{dev} = \frac{\mid \xi_{BU} - \xi_{SK} \mid}{\xi_{SK}} \cdot 100\% \tag{3.16}$$

Codebook computation effort of the bottom-up structure is also compared with single K-means clustering in Table 3.4. Both bottom-up and single K-means clustering are

Table 3.4: SSE deviation percentage δ_{dev} and computation time in codebook generation using bottom-up (BU) and single K-means (SK) structures.

Codebook Size	cb_{BU} =800	cb_{SK} =800	cb_{BU} =1600	cb_{SK} =1600
δ_{dev}	1.4 %		3.7 %	
Computation	4hrs	350hrs	9hrs	648hrs

employed on a single Quad CPU at 2.40GHz with 4.0G RAM machine, in which the bottom-up is only simulated as parallel computing in a serial sequence. To generate a codebook with size 800, the single K-means clustering uses 350 hours, while the bottom-up clustering only takes four hours. When the codebook size is doubled to 1,600, the computation for single K-means and bottom-up clustering are 648 hours and 9 hours, respectively. With a truly distributed processing environment using multiple computers, bottom-up processing time will be further reduced. This comparison of computational complexity demonstrates that our generic framework using robust bottom-up clustering for codebook generation can replace the single K-means in dealing with large-scale and diverse datasets.

For the accuracy performance using k-NN and various dissimilarities, Table 3.5 shows the average genre categorization results for 23 different sports. The proposed bottom-up codebook generation manifests a better and more robust performance than single K-means codebook generation in both EMD and KL-div measurements. By comparing the row-wise's dissimilarities, the bottom-up structure is more consistent with codebook sizes of 800 and 1600. On the contrary, the single K-means codebook generation is unstable for both histogram and mLDA-based distributions. For instance, the performance at a codebook size of 800 using EMD has about a 7% increment from ED dissimilarity (75.33% vs. 68.31%), while the counterpart at a codebook size of 1,600 using EMD has dropped 1.1% from ED dissimilarity (64.28% vs.65.39%). One reason is that the single K-means clustering on over 3 million input SIFT points hardly reaches the optimal value. As a summary, KL-div performs the best among three dissimilarity measures. Using the bottom-up structure, results of the codebook size 1,600 outperform the cases with size 800 in all measurements with consistency. Oppositely, single K-means clustering results are not consistent.

Table 3.5: Average categorization results (%) of 23-sports data with codebook size 800 and 1,600.

Measurement	ED	EMD	KL–div
$\mathbf{cb_{BU}=800}$	61.54	75.80	78.59
$\mathbf{cb_{SK}=800}$	68.31	75.33	73.49
$\mathbf{cb_{BU}=1600}$	65.68	78.94	82.16
$\mathbf{cb_{SK}=1600}$	65.39	64.28	75.75

Note: BU: codebook generated using bottom-up structure. SK: codebook generated using single K-means structure.

Another merit of the bottom-up structure is its preservation of individual genre characteristics from the 1st-level K-means. On the contrary, single K-means codebook generation covers all the data; thus, a weakly distinguishable genre is easily overruled by a strong one. This reasoning explains why with the increase of codebook size from 800 to 1,600, the bottom-up process has about a 4% improvement for KL-div, while the single K-means process has only a 2% increment for KL-div.

The individual sport genre classification result is illustrated in Figure 3.7. On average, a codebook size of 1,600 gives an average of 3.6% higher than the codebook size of 800, which corresponds with the empirical studies from other research groups [58, 78].

To evaluate the generic and extensive properties of our proposed method, experimental results on the 23-sports dataset are compared with results in Li *et al.*'s work [81], where a top-down process was adopted using single K-means as its top layer general codebook. The best performance in two-layer and single-layer structures are 83.83% and 81.2%, respectively [81]. In their work, speeded up robust features (SURF)-based method is adopted. Similar to SIFT, SURF is also a scale and rotation-invariant interesting point feature extraction algorithm, which focuses on the computational efficiency [30]. Although SURF and SIFT adopt different key points detection techniques, these two descriptors are comparable in characterizing local features of sampled frames from a video sequence. Therefore, such a comparison is valid in genre categorization performances, regardless of the feature extraction difference. Considering the increment of data in scale about 27% (145 hrs vs. 114.2 hrs), and in variety about 64% (23 genres vs. 14 genres), using the bottom-up structure with a codebook size of 1,600 and KL-div measurement,

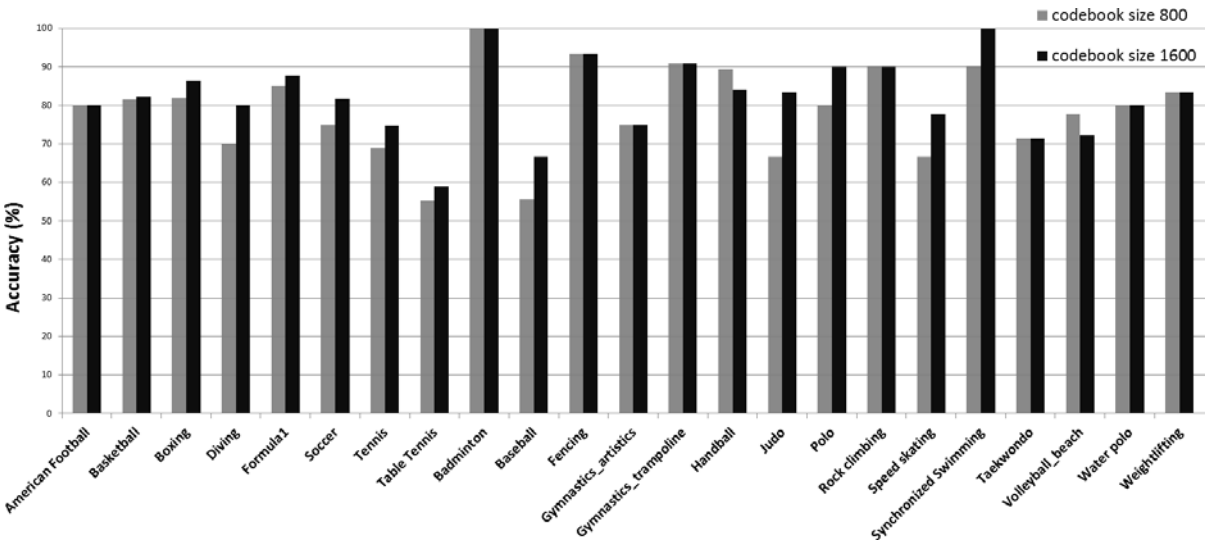


Figure 3.7: Genre categorization for the 23-sports dataset with codebook sizes of 800 and 1600.

our experimentation provides comparable results of 82.16%, with a degradation of 1.67%.

Although the performance is maintained on average, we also observed that the individual performance has been fluctuating. This fluctuation is mainly due to the nature of the adopted k-NN classifier, where distance-based measurement can be overruled by a strong representation in a large and sparse dataset. We acknowledge that k-NN may not be the most robust algorithm towards the very large-scale dataset. However, the k-NN is an efficient method in batch processing. It can be used as a coarse and preliminary execution to quickly prune off the large portion of the irrelevant data.

From a different perspective, generic properties of the proposed method are assessed using various video clip lengths and frame sampling methods. As detailed in Table 3.6, better performance is acquired using longer lengths of video clips, while a generic and automatic uniform sampling method outperforms the key-frame sampling. It is because the proposed method is based on local key-point descriptors. Therefore, a longer video clip with denser sampling frames provides more key-points and consequently builds a better distribution than a shorter clip with less sampled key-frames/shots. Such experimentation demonstrates the merit of our proposed generic method towards a truly large-scale dataset.

Table 3.6: Genre categorization accuracy between various video clips with uniform sampling-based and key-frame/shot-based methods.

3 Minutes Clip		***10 seconds Clip***	
Uniform Sampling	**Key-frame/ Shot**	**Uniform Sampling**	**Key-frame/ Shot**
83.83%	79.41%	71.90%	63.10%

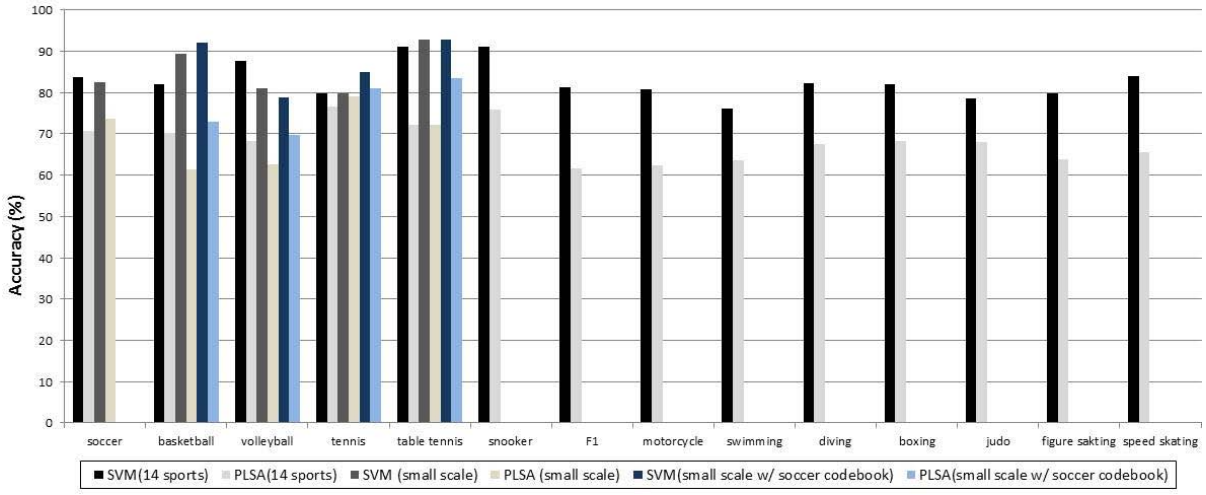


Figure 3.8: View type classification using supervised SVM and unsupervised PLSA. First two columns are with codebook size 800 for 14 sports.

3.6.2 View Classification Analysis Using Supervised SVM and Unsupervised PLSA

Experiments in this section focus on middle-level view classification by utilizing extracted low-level histogram-based representations. A subset of 14 sports of all 23 sports was used as test data. Figure 3.8 compares both supervised SVM and unsupervised PLSA results as the 1st and 2nd columns, respectively. On average, supervised SVM has a classification accuracy of 82.86%, and unsupervised PLSA has an average of 68.13%, in which the SVM technique outperforms the PLSA algorithm by 14.73%.

It needs to be pointed out that this evaluation is based on predetermined semantic view types, which are in favor of the SVM algorithm. It is because such a semantic definition has become considerably involved in SVM training, while barely being used in PLSA training. In the SVM method, labeled training data associated with each

predefined view type are indispensable for building the classifier. On the other hand, the PLSA model training merely requires a specified number of view types, which is similar to the number of clusters needed for training a K-means clustering. Thus, it is anticipated that the supervised SVM method will have better performance than the unsupervised PLSA algorithm.

However, the PLSA model is advanced in its unsupervised characteristics such that the labeled data is not necessary in training. This feature makes the PLSA more suitable than the SVM and significant in supporting the generic framework dealing with large-scale datasets, where automatic processes and minimum human and expertise interventions are essential. For evaluating our proposed framework, a trade-off in the classification accuracy can be afforded, if the ultimate event detection results are comparable using either the PLSA or the SVM view results.

In order to analyze the generic and scalable properties, a subset with small-scale five-sports dataset is applied, including {soccer, basketball, volleyball, table tennis, tennis}. The SVM and PLSA view classification performance of this small-scale dataset is presented in the 3rd/4th columns of Figure 3.8, respectively. The baseline on the small-scale data, the 14-sports, has a 0.27% performance drop in SVM and an improvement of 1.76% in PLSA. With similar results, compared with the five-sport small-scale data, the 14-sport view dataset has a lot more data in both variety and volume.

Based on the preceding analytical results, the extrapolated performance from this current relatively large-scale dataset to a truly large-scale dataset should be maintained, especially for the PLSA method. The reasoning is twofold: first, large-scale data is normally sparse; PLSA, as a generative model, has a characteristics in probabilistically mapping data from a high-dimensional space to a low-dimensional space. Hence, more information brought by the new data can help in finding significant representatives in the lower dimensional space. Second, since the number of view classes are fixed at four types, more variety and volume will not affect the performance much.

Additionally, a knowledge transfer property is investigated by using the same five-sport dataset. It can be seen that an individual sport from insufficient resources {basketball, volleyball, table tennis, tennis} can be assisted by borrowing the codebook from an abundant sport resource {soccer}. As Figure 3.8 depicts, these limited-source four sports in the 5th/6th columns, the codebook transfer mechanism has improved about 2.07 % and 5.05% for the SVM and PLSA on average, respectively. The margin of improvement

using the PLSA is bigger than its counterpart in SVM. This result can be explained by the nature of two different techniques. PLSA is a probabilistic-based dimensional reduction technique. Therefore, more data will provide a more thorough characterization of the low-dimensional model. On the contrary, SVM is a technique mapping from a low dimensional space to a higher dimensional space. More information brought by the codebook may be overwhelmed by the SVM process and may not necessarily provide a better classification in the higher-dimensional space. Therefore, such a knowledge transfer property could help the unsupervised PLSA in further improving its performance for sports with scarce resources.

3.6.3 Event Detection Using Coarse-to-fine Scheme and HCRF-based Structured Prediction Model

In previous experiments, the proposed framework provides an application to identify video genres by directly utilizing domain knowledge-free SIFT descriptors and a BoW model. After the genre is determined, individual frames of the query video sequence are labeled by the middle-level semantic views via either supervised or unsupervised classifiers. In this experiment, the task on basketball score event detection is investigated by employing this labeled video sequence. A two-staged coarse-to-fine scheme is adopted that first detects scoreboard information change, introduced by Miao *et al.* [115]. By adopting this technique, an entry point of an interesting event is located. However, this coarse detection only provides a static frame-based rough estimation as an entry point. Since scoreboard information not only appears in score events, but also in time-out events or intermission events, individual frame-based detection without temporal structured information cannot provide robust and satisfactory results. Therefore, a fine-tuning process in finalizing detection is adopted to ensure that the query video truly conveys the score event as its semantic theme. The proposed HCRF model is deployed as the fine-tuning process after the first-stage coarse detection. Experimental results using this HCRF model are compared with CRF and HMM baselines.

Two video groups consisting of four matches are utilized, which are defined as (a) Dataset A: using two NBA games for training and using another two Olympic Games for testing; (b) Database B: using one NBA game for training and using another NBA game for testing. Frame-based views from the PLSA model and the SVM model are applied to

Table 3.7: Precision and recall results of basketball score events detection at the first (coarse) stage.

Correctly Detected Score (true positive)	Detected Score (correct result)	Correct Total Score (obtained result)	Precision (%)	Recall (%)
231	251	268	92.03	86.19

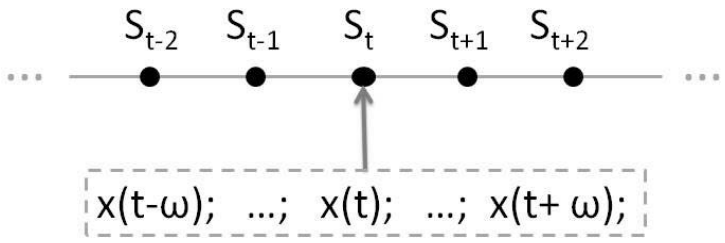


Figure 3.9: Current state influenced by surrounding observed states.

Dataset A and B. Therefore, four combinations of view labels and datasets are defined as $PLSA + A$, $PLSA + B$, $SVM + A$, and $SVM + B$. Each video clip used in both training and testing is automatically decimated and consists of 500 uniformly sampled frames. We use a window size $N = 20$, which is introduced in Figure 3.5 and Equation (3.7) from Section 3.5, with a window N sliding every ten frames. The final number of the states sequence for HCRF is thus calculated as $49 = 500/(20 - 10) - 1$.

The number of approximated events detected after the first stage is given in Table 3.7. The precision and recall of the coarse-stage basketball score detection are 92.03% and 86.19% respectively. In the second stage, the proposed HCRF-based model and state-of-the-art HMM and CRF models are evaluated and compared. The advantage of HCRF over HMM is its relaxation on the Markov property that the current state S_t can be inferred from both current observations, as well as surrounding observations, as illustrated in Figure 3.9. In the experiment, the circumferential range number is selected at $\omega = 0, 1, 2$. As shown in Table 3.8, the HCRF has better performance than the CRF for the same ω values, while both models outperform the HMM baseline. When using different ω values for both CRF and HCRF, $\omega = 1$ provides better results than $\omega = 0$, in which neighboring information assists in better decision-making. However, when $\omega = 2$ is used for HCRF, the performance has been dropped for all cases compared with $\omega = 1$. This performance degradation can be viewed as an overfitting issue, in which

Table 3.8: Performance comparison on score event detection in basketball.

	Accuracy			
	Dataset A (NBA/Olympics)		**Dataset B (NBA/NBA)**	
	SVM+A (%)	PLSA+A (%)	SVM+B (%)	PLSA+B (%)
HMM $\omega = 0$	78.28	75.29	87.50	85.94
CRF $\omega = 0$	78.16	74.57	87.43	86.52
CRF $\omega = 1$	79.52	76.82	88.52	87.89
HCRF $\omega = 0$	80.93	75.53	90.00	90.77
HCRF $\omega = 1$	83.26	80.24	93.08	92.31
HCRF $\omega = 2$	82.09	77.88	91.46	91.77

Note: Dataset A: NBA matches as training, Olympic matches as testing. Dataset B: NBA matches for both training and testing.

adding more surrounding information limits the structured prediction ability. A similar overfitting problem is also observed in gesture recognition research using HCRF [118]. In summary, the proposed HCRF-based model with parameter $\omega = 1$ outperforms both CRF and HMM models. The best results are obtained at 93.08% and 92.31% by taking SVM-and PLSA-based input labels, respectively.

On the other hand, by comparing the proposed PLSA with SVM benchmark, performance discrepancy of the event detection has been shortened, despite the input view classification (as shown in Figure 3.8) has PLSA (70.14%) outperformed by SVM (82.00%) with 11.86%. For Dataset A, the average difference shows that SVM outperforms PLSA by 3.65%, while in Dataset B, such a difference is only 0.47%. This tolerable difference demonstrates the robustness and resilience of structured prediction models in accommodating poorly labeled video sequences from PLSA, yet achieving comparable performance with those labeled sequences from SVM. Therefore, the event detection presented in this work achieves similar results by both unsupervised and supervised learning. However, due to PLSA's reduced human involvement, the unsupervised classifier is preferred in large-scale video analysis.

Experimental result discrepancies using Dataset A and Dataset B are also compared. Although both datasets belong to basketball, Dataset B (with NBA matches for both training and testing) outperformed Dataset A (with NBA matches for training and Olympics matches for testing) by 10.9% on average. It suggests that albeit Datasets

A and B are of the same genre and event detection task, a significant difference exists. Such a difference can be explained by assuming that NBA and international basketball (FIBA) are two different styles of the same genre. In terms of computer vision and structured prediction, NBA and FIBA have related but different temporal patterns even in the same semantic event. Thus, by training/testing in the same style, it is expected to have a better detection rate than training/testing using different styles. This is also an example of the semantic gap–that semantic event recognition with discrepant conditions is still not perfect.

Although there is only one event detection example discussed, it is believed that the method can be extended and generalized to a bigger pool of event scenarios. The reason is fourfold: First, the experiment data of the basketball score event are multi-source and non-simplex. Videos are collected from both internet and TV recordings, and there are different production rules of NBA and Olympics basketball. Second, the video representation module using local features and the BoW model is domain knowledge-free and with no production rules involved. Such a generic approach has been proven to be effective in genre categorization of 23 sports, view classification of 14 sports, and the basketball score event. Third, the event detection algorithm utilizing HCRFs, as well as baseline HMMs and CRFs are structured prediction models and belong to the category of state event model. By comparing the number of events analyzed using different event models from Table 3.3, the state event model, a recently popular approach in literature, is capable in handling more events than the other two model types (i.e. patten-recognition model and semantic event model). In addition, among the state event models, most methods utilize middle-level semantic agents. In our work, the adopted four-category view type definition is one of the most popular classification schemes in literature. Last and most important, the input of our event detection model is a sequence of labeled views which is the result of a domain knowledge-free method (either PLSA or SVM), using generic video representation. With better accuracy achieved by the proposed HCRF-based model than baselines HMM- and CRF-based models, the performance should be maintained with other labeled sequences which could form various event scenarios. Moreover, utilizing sequences labeled by the middle-level agents as input, is also popular among peers' work with state event models [110, 111, 113, 114].

3.7 Summary

This chapter introduces the BoW model, with its incorporation of unsupervised learning algorithms, in analyzing large-scale video dataset generically and systematically. Three video tasks are investigated in a coherent and sequential order. After processing all data indifferently at the feature extraction stage using domain knowledge-free local SIFT descriptors, video sequences are represented by utilizing compact and concise BoW model. Then, a systematic scheme is employed for interesting event detection, by taking the video sequence as query. In this framework, after its genres identified using a k-NN classifier, the query video is evaluated by a semantic view assignment as the second stage using the PLSA model. Both genre identification and view classification tasks utilize the initially processed video representation as input, and unsupervised algorithms as classifiers. Finally in the third task, the interesting event is detected by feeding the view labels into an HCRF-structured prediction model.

Overall, this framework demonstrates the efficiency and generality in processing voluminous data from a large-scale sports collection and achieves various tasks in video analysis. The effectiveness of the framework is justified by extensive experimentation and results are compared with benchmarks and state-of-the-art algorithms. As a conclusion, with little human expertise and effort involvement in both domain knowledge-independent video representation and annotation-free unsupervised view labeling, the proposed generic and systematic method using the BoW model is promising in processing videos, and has the potential for even larger and more diverse datasets.

Chapter 4

Interactive Mobile Visual Search and Recommendation Using the Bag-of-words Model

4.1 Introduction

The bag-of-words (BoW) model and its application in content-based retrieval has shown promising results in desktop-based visual searches. In this chapter, we present a mobile visual search algorithm by combining the BoW model's merit with user interaction through a mobile platform. We proposed an innovative context-aware search-tree based on the BoW paradigm, which includes both user specified region of interest (ROI) and surrounding pictorial context. There is a mutual benefit by combining the visual search using the BoW model with mobile devices.

From a retrieval point of view, although the BoW model has shown promising results in desktop-based visual searches for large-scale consortia, it also suffers a semantic gap. The BoW model is limited by its homogenous process in treating all regions without distinction. Features are extracted homogeneously, and local features are treated without emphasis. Therefore, information provided by a query image without priority can mislead the computer vision algorithm for recognition. Hence, to have a better retrieval result, there is a need to orderly utilize local visual information. Multi-touch screen and its user interaction on mobile-devices offer such a platform for users to select their ROIs as

prioritized information, with surrounding context as secondary information.

From a mobile application perspective, visual search via image query provides a powerful complementary carrier besides conventional textual and vocal queries. Compared to conventional text or voice queries for information retrieval on-the-go, there are many cases where visual queries can be more naturally and conveniently expressed via mobile device camera sensors (such as an unknown object or text, an artwork, a shape or texture, and so on) [128]. In addition, mobile visual search has a promising future due to the vital roles mobile devices play in our life, from their original function of telephony, to prevalent information-sharing terminals, to hubs that accommodate tens of thousands of applications. While on the go, people are using their phones as a personal concierge discovering what is around and deciding what to do. Therefore, the mobile phone is becoming a recommendation terminal customized for individuals—capable of recommending contextually relevant entities (local businesses such as a nearby restaurant or hotel) and simplifying the accomplishment of recommended tasks. As a result, it is important to understand user intent through its multi-modal nature and the rich context available on the phone.

Motivated by the above observations, this chapter presents an interactive search-based visual recognition and contextual recommendation using the BoW model. Smart-phone hardware such as camera and touch screen, are taken advantage of in order to facilitate expressions of user's ROI from the pictures taken. Then, the visual query along with such a ROI specification go through an innovative contextual visual retrieval model to achieve a meaningful connection to database images and their associated rich text information. Once the visual recognition is accomplished, associated textual information of retrieved images are further analyzed to provide meaningful recommendations.

An actual system codename *TapTell* is implemented based on the proposed algorithms and methodologies. A natural user interaction is proposed to achieve the *Tap* action, in which three gestures are investigated (i.e., circle, line, and tap). We conclude that the circle (also called "O" gesture) is the most natural interaction for users, which integrates user preference to select the targeted object. We adopt the BoW model introduced in Chapter 2 and propose a novel context-embedded vocabulary tree. The algorithm incorporates both ROI visual query and the context from surrounding pixels of the "O" region to search similar images from a large-scale image dataset. Through this user interaction (i.e., "O" gesture) and the BoW model with our innovative algorithm, standard visual

recognition can be improved. The *Tell* action is accomplished by recommending relevant entities based on recognition results and associated metadata.

The novelty of the chapter lies in the following aspects:

- We adopt the BoW model and propose a context-aware visual search algorithm in which a novel context-embedded vocabulary tree (CVT) is designed. The algorithm is able to achieve better visual recognition performance by embedding the context information around the "O" region into a standard visual vocabulary tree.

- Based on the proposed context-aware visual recognition, we implemented a real system *TapTell* to understand users' visual intents. The goal is to provide a contextual entity suggestion for activity completion that provides meaningful and contextually relevant recommendations. We utilize the advances of touch screen technology provided at the mobile platform and introduce human experts in loop for a better visual search. We investigate three different kinds of gestures for specifying object (and text) of interest by a user study. We conclude that "O" provides the most natural and effective way to interactively formulate user's visual intent and thus reduce ambiguity. After obtaining the recognition results, we propose a location-aware recommendation which suggests relevant entities for social task completion.

In the following, an interactive mobile visual search using the BoW model and the proposed CVT algorithm is first presented. A viable application, *TapTell*, is introduced in detail to show how to accomplish meaningful contextually relevant recommendations through mobile recognition. Experimental results are provided to demonstrate the effectiveness of the proposed method.

4.2 BoW-Based Mobile Visual Search

This section presents the mobile visual search with proposed context-aware image retrieval using the BoW model. Section 4.2.1 introduces the literature and industrial developments of mobile visual search. Section 4.2.2 presents an overview of the proposed algorithm. Finally, Section 4.2.3 presents the visual recognition by search using the BoW model, with the help of both image context as well as sensory GPS information. Section 4.2.4 summarizes this section.

4.2.1 Mobile Visual Search

Mobile Visual Search in Industry

Due to its potential for practicality, mobile visual search is one of the research areas drawing extensive attention from both industry and academia. Table 4.1 summarizes representative mobile visual search applications from industry. Different from the above mentioned applications, the proposed system is innovative in terms of an interactive gesture-based (using advanced multi-touch function) visual search system to help users to specify their visual intent, with a consequent recommendation based on the visual search results and contextual information. In this perspective, our system leverages visual search results to formulate a second query to accomplish task completion on mobile devices, which is significantly different from existing applications.

Mobile Visual Search in Academia

In academia, the workshop on mobile visual search has been gathering researchers and engineers to exchange various ideas in this field [129]. Quite a few research efforts have been put into developing compact and efficient descriptors, which can be achieved on the mobile end. Chandrasekhar *et al.* developed a low bit-rate compressed histogram of gradients (CHoG) feature which has a great compressibility [45]. Tsai *et al.* investigated in an efficient lossy compression to code location information for mobile-based image retrieval. The performance is also comparable with its counterpart in lossless compression [130].

On the other hand, contextual features such as location information have been adopted and integrated successfully into mobile-based visual searches. Schroth *et al.* utilized GPS information and segmented searching area from a large environment of city to several overlapping subregions to accelerate the search process with a better visual result [131]. Duan and Gao proposed a side discriminative vocabulary coding scheme, extending the location information from conventional GPS to indoor access points as well as surrounding signs such as the shelf tag of a bookstore, scene context, and etc. [132].

Additionally, other researchers targeted practical applications and provided promising solutions. Takacs *et al.* proposed a loxel-based visual feature to describe region-related outdoor object features [133]. Chen and Tsai proposed methods on using image process-

Table 4.1: Summary of mobile visual search applications in the industry.

Application	Features	Techniques	Company
Goggles	product, barcode, cover, landmark, name card, artwork	visual search, OCR	Google
Bing Vision	cover, art, text, barcode	visual search, OCR	Microsoft
Point&Find	palces, 2D barcode	visual search	Nokia
Digimarc	print, article, ads	watermarking	Digimarc
SnapTell	cover (CD/DVD/book/video-games), barcode	visual search	Amazon
Mobots	ads, print, product, cover	visual search	Mobots
Kooaba	cover, print	visual search	Smart Visuals
Layar	direction, points of interest	geo-location, AR	Layar
WordLens	Real-time English/Spanish translation	OCR, AR	QuestVisual

ing techniques to find book spines in order to index book inventories based on bookshelf images [134, 135]. Girod *et al.* investigated mobile visual search from a holistic point of view with practical analysis under mobile device constraints of memory, computation, devices, power and bandwidth [49]. An extensive analysis using various feature extraction, indexing and matching techniques is conducted using real mobile-based Stanford Product Search system. They demonstrated a low-latency interactive visual search with satisfactory performance.

A Summary of the Proposed Work

Aforementioned visual search methods and applications on mobile devices have demonstrated their merits. Alternatively, we believe that combining visual recognition techniques with personal and local information will provide contextually relevant recommendations. Hence, this work proposes a mobile visual search model to suggest potential social activities on-the-go.

We have investigated three types of user interactions (i.e., the tapping, straight line, and circle gestures) to facilitate the expression of the user intent. Then, the visual query goes through an innovative contextual visual retrieval model using the state-of-the-art BoW paradigm, to achieve a meaningful connection to database images and their associated metadata information. Once the user intent expression is predicted by such visual recognition, associated textual information of retrieved images are further analyzed to provide meaningful textual-based social activity and task recommendation.

4.2.2 Overview

Figure 4.1 shows the framework of our visual recognition and activity recommendation model. In general, it can be divided into the client-end and cloud-end. On the client-end, a user's visual search intent is specified by the "O" gesture on a captured image. On the cloud-end, with user selected object and the image context around this object, a recognition-by-search mechanism is applied to identify user's visual intent. We have designed a novel context-embedded vocabulary tree to incorporate the "O" context (the surrounding pixels of the "O" region) in a standard visual search process. Finally, the specified visual search results are mapped to associate metadata by leveraging sensory context (e.g., GPS-location), which are used to recommend related entities to the user.

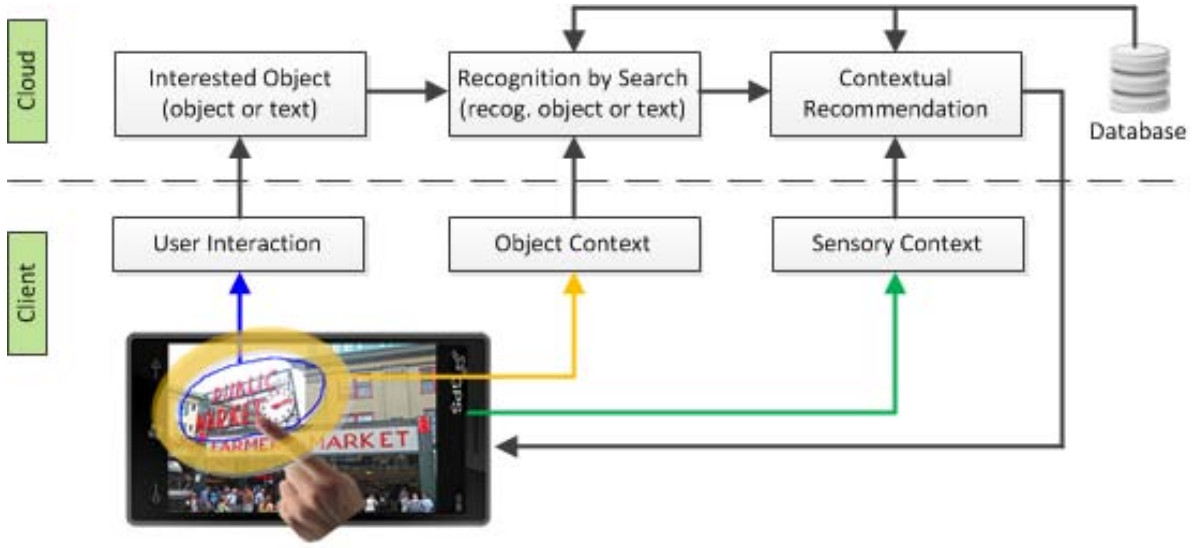


Figure 4.1: Proposed framework of mobile visual search and activity completion model using image contextual model, including 1) "O"-based user interaction, 2) image context model for visual search, and 3) contextual entity recommendation for social activities.

The "O" gesture utilizes multi-touch screen of the smart-phone. Users do not need any training and can naturally engage with the mobile interface immediately. After the trace (the blue thin line in Figure 4.1) has been drawn on the image, sampling points along the trace-line are collected as $\{\mathbf{D}|(x_j, y_j) \in \mathbf{D}\}_{j=1}^{N}$, which contain N pixel-wise positions (x_j, y_j). We applied principal component analysis (PCA) to find two principal components (which form the elliptical ring depicted by thick orange line in Figure 4.1). The purpose of this part is to formulate a boundary of the selected region from an arbitrary "O" gesture trace. We also calculated mean μ and covariances Σ based on $\mathbf{D}$ and non-correlated assumption along the two principal components:

$$\mu = [\mu_x, \mu_y] \qquad \Sigma = \begin{vmatrix} \sigma_x^2 & 0 \\ 0 & \sigma_y^2 \end{vmatrix}. \tag{4.1}$$

Figure 4.2 shows the computation of principal components from the "O" query. Once the principal components are identified, proposed image contextual model for mobile visual search is used to identify the object of interest indicated by the user.

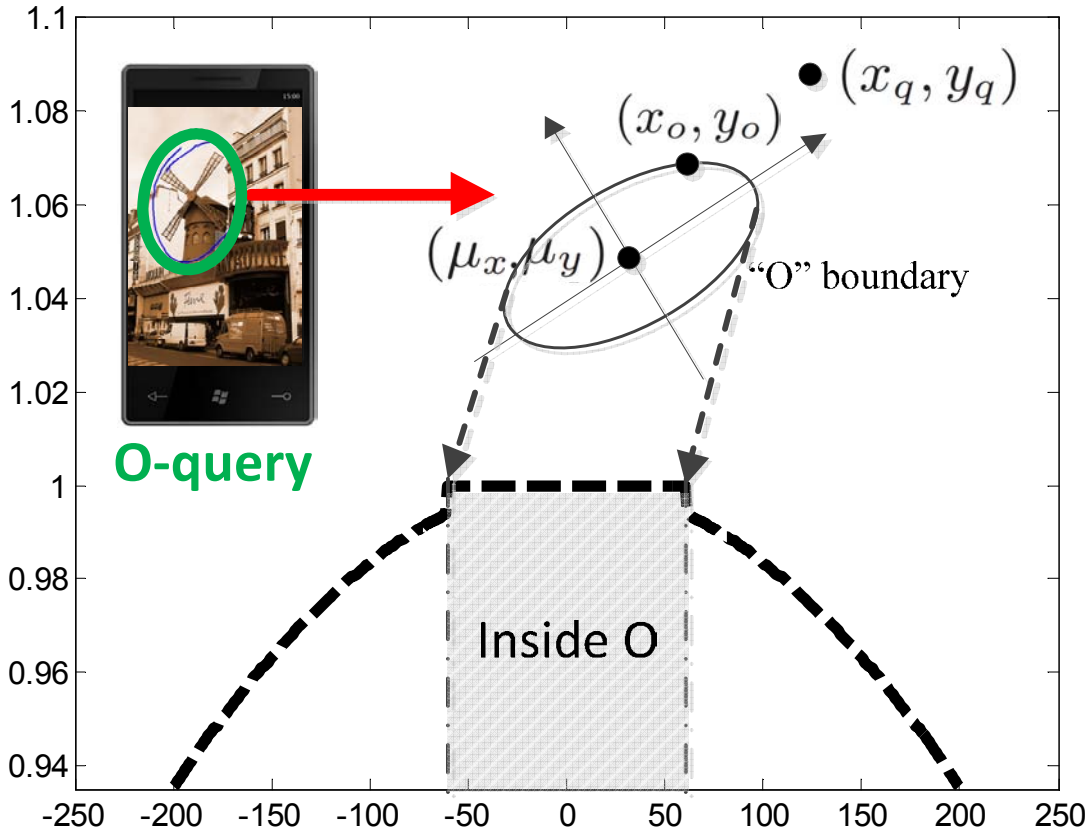


Figure 4.2: Illustration of user indicated "O" query, and the computation of principal components of the query. (μ_x, μ_y) is the center of "O" query, (x_o, y_o) is a pixel on the "O" boundary, and (x_q, y_q) is a query pixel.

4.2.3 Context-aware Visual Search Using the BoW Model

The visual intent recognition method is based on a retrieval scheme using the BoW model with the vocabulary tree proposed by Nister *et al.* [3]. This method provides a fast and scalable search mechanism and is suitable for large-scale and expansible databases because of its hierarchical tree-structured indexing. We adapt this method in the mobile domain, because the "O" gesture fits naturally to provide a focused object selection for better recognition. Different from using the entire image as visual query in [3], we have user-indicated ROI from the "O" gesture (called "O-query"). We design a novel context-aware visual search method in which a CVT is built to take the surrounding pixels around the O-query into consideration. The CVT algorithm focuses on first building a visualwords codebook for the BoW model to map each local feature, and subsequently,

constructing a BoW representation. By establishing a hierarchical K-means clustering for the codebook, this algorithm manages to shorten the codebook generation process. Therefore, it is scalable and efficient for processing large-scale data. Specifically, the CVT algorithm is able to reduce the following ambiguities:

- Sometimes, issuing O-query only in image-based search engines may lead to too many similar results. The surrounding pixels provide a useful context to differentiate those results.
- Sometimes, the O-query may not have (near) duplicates or exist in the image database. Issuing only O-query may not lead to any search results. The surrounding pixels then can help in providing a context to search for the images with similar backgrounds.
- Hierarchically built K-means clustering for codebook generation makes the retrieval process efficient, wherein each queried local feature only goes through one particular branch at the highest level and its sub-branches instead of going through the entire codebook.

The proposed CVT-based visual search method encodes different weights of term frequencies inside and outside the O-query. We will carefully describe the proposed visual search algorithm in Section 4.2.3. We also propose a location-context-based filter process in Section 4.2.3 for re-ranking visual search results based on user's current location (derived from the GPS-enabled images taken by the phone camera). For off-line image indexing, we first extract SIFT local descriptors. Since our target database is large-scale, we utilize the hierarchical K-means to cluster local descriptors and build the CVT. Then, we index the large-scale images using the built CVT and the inverted file mechanism, which is to be introduced in the following sections.

Context-aware visual search

In on-line image searches, given a query image, we can interpret the descriptor vectors of the image in a similar way to the indexing procedure, and accumulate scores for the images in the database with a so-called *term frequency–inverse document frequency* (tf–idf) scheme [3]. This tf–idf method is an effective entropy weighting for indexing a

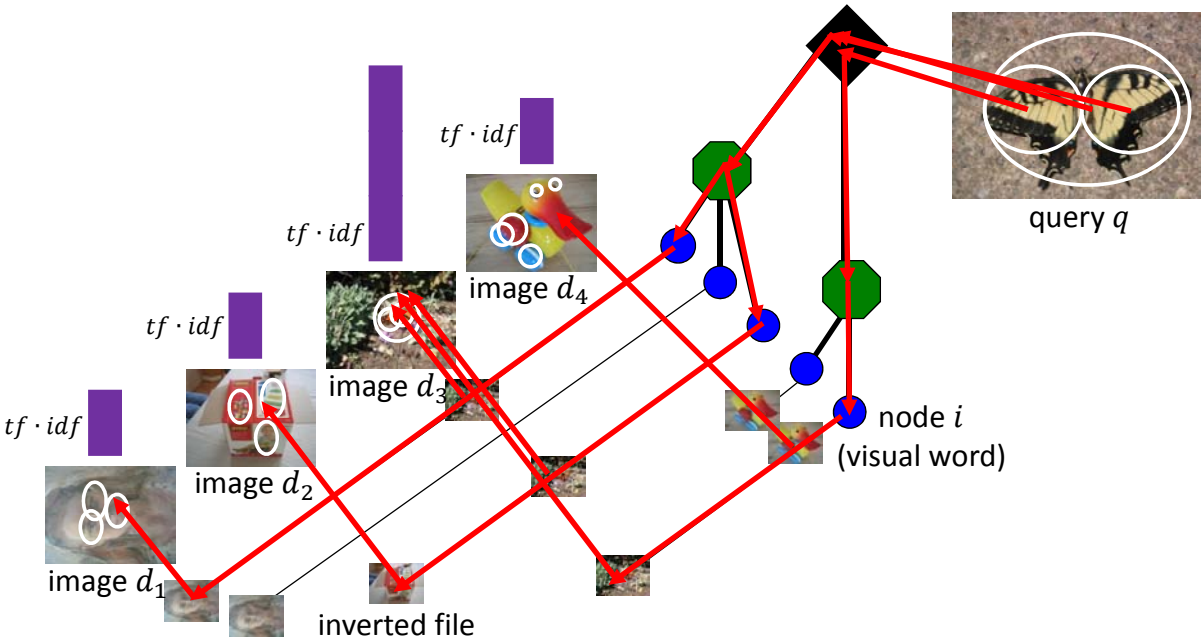


Figure 4.3: Image search scheme with visual vocabulary tree [3]. Note that the white circle in the image corresponds to a local descriptor (not an O-query).

scalable database. Figure 4.3 shows the computation of image similarity based on the tf–idf scheme. In the vocabulary tree, each leaf node corresponds to a visualword i, associated with an inverted file (with the list of images containing this visualword i). Note that we only need to consider images d in the database with the same visual words as the query image q. This significantly reduces the amount of images to be compared with respect to q. The similarity between an image d and the query q is given by

$$\begin{aligned} s(q,d) &= \| \mathbf{q} - \mathbf{d} \|_2^2 \\ &= \Big(\sum_{i|d_i=0} |q_i|^2 + \sum_{i|q_i=0} |d_i|^2 + \sum_{i|q_i\neq 0, d_i \neq 0} |q_i - d_i|^2 \Big) \end{aligned} \tag{4.2}$$

where $\mathbf{q}$ and $\mathbf{d}$ denote the tf–idf feature vectors of the query q and image d in the database, which are consisted of individual elements q_i and d_i (i denotes the i-th visualword in the vocabulary tree), respectively. q_i and d_i are the tf–idf value for the i-th visualword in the query and the image, respectively. Mathematical interpretations are

given by

$$q_i = tf_{i_q} \cdot idf_i, \tag{4.3}$$

$$d_i = tf_{i_d} \cdot idf_i. \tag{4.4}$$

In the above equation, the *inverted document frequency* idf_i is formulated as $ln(N/N_i)$, where N is the total number of images in the database, and N_i is number of images with the visualword i (i.e., the images whose descriptors are classified into the leaf node i).

The *term frequency* representations tf_{i_q} and tf_{i_d} are computed as the accumulated counts of the visualword i in the query q and the database image d, respectively. One simple means for the *term frequency* computation is to use the O-query as the initial query without considering the pixels surrounding the "O". This process is equivalent to using "binary" weights of the *term frequency* tf_{i_q}: the weight is 1 inside "O", and 0 outside "O". A more descriptive and accurate computation is to incorporate the context information (i.e., the surrounding pixels around the O-query) in the vocabulary tree. We design a new representation of the *term frequency* $tf^o_{i_q}$ for the O-query. A "soft" weighting scheme is proposed to modulate the *term frequency* by incorporating the image context outside the O-query, which was neglected in the simple binary scheme. When quantizing descriptors in the proposed CVT, the $tf^o_{i_q}$ of the O-query for a particular query visualword i_q is formulated as:

$$tf^o_{i_q} = \begin{cases} tf_{i_q}, & \text{if } i_q \in \text{O} \\ tf_{i_q} \cdot \min\left\{1, \frac{\Re(x_q,y_q)}{\Re(x_o,y_o)}\right\}, & \text{if } i_q \notin \text{O} \end{cases} \tag{4.5}$$

where $\Re(x_o, y_o)$ and $\Re(x_q, y_q)$ denote the Gaussian distances of the pixel (x_o, y_o) and (x_q, y_q) with respect to the center of O-query (μ_x, μ_y). Figure 4.2 shows the definition of these pixels in the query image q. The Gaussian distance $\Re(x, y)$ for an arbitrary pixel (x, y) is given by

$$\Re(x, y) = A \cdot \exp\Big\{ -\frac{1}{2}\big[\frac{(x-\mu_x)^2}{\alpha\sigma_x^2} + \frac{(y-\mu_y)^2}{\beta\sigma_y^2}\big]\Big\} \tag{4.6}$$

The "soft" weighting scheme shown in Equation (4.5), is a piece-wise, bivariate-based multivariate distribution outside the O-query, and a constant 1 inside the O-query. The

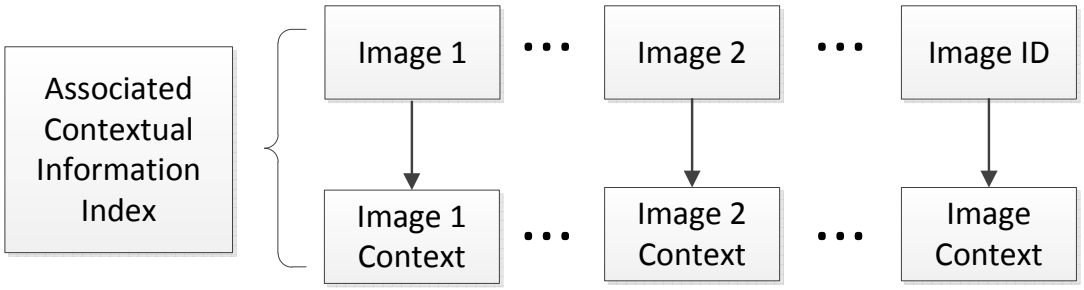


Figure 4.4: Sensory context information index associated with each image.

position (x_o, y_o) is the boundary of the O-query contour where the weight 1 ends. In the case that a visualword i_q is outside the O-query, the modulating term is $\min\left\{1, \frac{\Re(x_q,y_q)}{\Re(x_o,y_o)}\right\}$, such that the soft weighting is guaranteed to be less than 1. The term $\frac{\Re(x_q,y_q)}{\Re(x_o,y_o)}$ is the ratio of which the query point (x_q, y_q) should be weighted with respect to its closest boundary position (x_o, y_o). Mean values μ_x and μ_y are calculated from "O" gesture sample data, while α and β are tunable parameters to control the standard deviation for the bivariate normal distribution. Figure 4.2 also illustrates this "soft" weighting schemes in the CVT when a projection view along one principal axis is sliced and presented. Parameter A is the amplitude value controlling the highest possible weighting scale. Parameters α and β reflect the importance of the horizontal and vertical axis (or directions) when employing the PCA technique. Empirically, we set α with higher value than β to indicate that the horizontal axis is usually more important than the vertical one. This is because most pictures are taken by the phone camera horizontally. As illustrated in Figure 4.3

Location-based filtering

Context information collected by mobile sensors plays an important role to help to identify users' visual intents. As Figure 4.4 illustrates, similar with the inverted file index method, each piece of image context information is indexed with the image itself during the off-line database construction.

In our system, GPS information from sensors is utilized and associated with each image taken by the phone camera. A filter-based process is used to remove the non-correlated images after the initial visual search. This is because GPS as an important context filter can be used to efficiently explore users' true intents by precisely knowing their locations. This process is formulated as:

$$S_L(q,d) = s(q,d) \cdot \phi(\mathbf{q},\ \mathbf{d})$$
$$where \quad \phi(\mathbf{q},\ \mathbf{d}) = \begin{cases} 1, & \text{if } dist_{quadkey}(\mathbf{q},\mathbf{d}) \in \mathrm{Q} \\ 0, & \text{if } dist_{quadkey}(\mathbf{q},\mathbf{d}) \notin \mathrm{Q} \end{cases} \tag{4.7}$$

The visual similarity term $s(q,d)$ is modulated by a location-based filter $\phi(\mathbf{q},\ \mathbf{d})$. This filter is based on the GPS effective region Q, which describes the geographical distance between the query and the database images. We defined $dist_{quadkey}(\mathbf{q},\mathbf{d})$ as the quadkey distance between the query $\mathbf{q}$ and the database image $\mathbf{d}$.

The quadkey method is adopted from the Bing Maps Tile System [1]. It converts the GPS coordinates to a hashing-based representation for fast search and retrieval. We present an example in Figure 4.5 to walk through the steps of conversion from the WGS-84 GPS to a quadruple tiles code. We encode the GPS to a 23 digits number with the ground resolution of possible 0.02m accuracy. The formulation of this distance is computed by the Quadkeys representation. GPS context from mobile sensor is collected first. The standard WGS-84 is encoded to the quadkey representation. In the illustration, pictures of the same landmark (the Brussels town hall) with both the front and the back façades are taken. These two photos have different WGS-84 information, which have 10 out of 15 quadkey digits identical after Bing Maps projection. In other words, the hamming distance between these two codes is 5, which is calculated using tables to approximate a ground distance of about 305m.

4.2.4 Summary

This section proposed a context-aware mobile visual search based on the BoW model and the hierarchical visual vocabulary tree. Contextual GPS information is also used in filtering the visual search result. In the next section, an implementation named *TapTell* is presented based on the CVT algorithm introduced. *TapTell* is able to achieve social activity recommendations through mobile visual searches.

[1] `http://msdn.microsoft.com/en-us/library/bb259689.aspx`

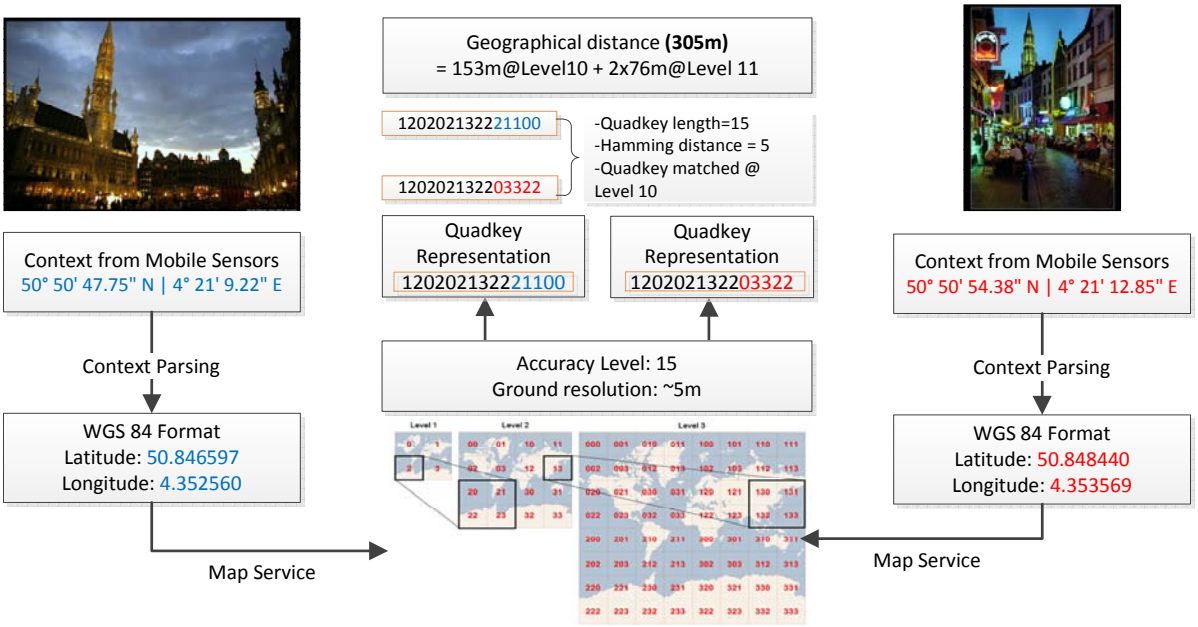


Figure 4.5: Quadkeys quantization and hashing from GPS, and images ground distance estimation using Microsoft Bing Map service.

4.3 *TapTell*: A Mobile Visual Search Implementation

TapTell is a system that utilizes visual query input through an advanced multi-touch mobile platform and rich context to enable interactive visual search and contextual recommendation. Different from other mobile visual searches, *TapTell* explores users individual intent and their motivation in providing a visual query with specified ROI. By understanding such intent, associated social activities can be recommended to users. Existing work has predominantly focused on understanding the intent expressed by text (or the text recognized from a piece of voice). For example, previous research attempts to estimate user's search intent by detecting meaningful entities from a textual query [136,137]. However, typing takes time and can be cumbersome on the phone, and thus in some cases, not convenient in expressing user intent. An alternative is to leverage speech recognition techniques to support voice as an input. For example, popular mobile search engines

enable a voice-to-search mode [2][3]. Siri is one of the most popular applications that further structure a piece of speech to a set of entities [4]. However, text as an expression of user intent has two major limitations. First, it relies on a good recognition engine and works well only in a relatively quiet environment. Second, there are many cases where user intent can be naturally and conveniently expressed through the visual form rather than text or speech (such as an unknown object or text, an artwork, a shape or texture, and so on) [128]. As an alternative, we believe that image is a powerful complementary carrier to express user intents on the phone.

Since *intent* is generally defined as "a concept considered as the product of attention directed to an object or knowledge" [138], we can define *mobile visual intent* as follows:

Definition 1 (Mobile visual intent) *Mobile visual intent is defined as the intent that can be naturally expressed through any visual information captured by a mobile device and any user interaction with it. This intent represents user's curiosity of certain object and willingness to discover either what it is or what associated tasks could be practiced in a visual form.*

The following shows scenarios of mobile visual intent and how expressed intent may be predicted and connected to social tasks for recommendation. The goal is not only to list related visual results, but also to provide rich context to present useful multimedia information for social task recommendation.

- You pass by an unknown landmark that draws your attention. You can take a picture of it. By using visual intent analysis, the related information of this landmark is presented to you.
- You see an interesting restaurant across the street. Before you step into the restaurant, you take a picture of it and indicate your interest using your gesture. By applying visual intent analysis, the information about this restaurant or its neighborhood points-of-interest matching your preference are recommended.
- You are checking a menu inside a restaurant, but you do not speak the language or know the cuisine. You can take a photo of the menu using your phone and

[2] `http://www.discoverbing.com/mobile`
[3] `http://www.google.com/mobile`
[4] `http://siri.com/`

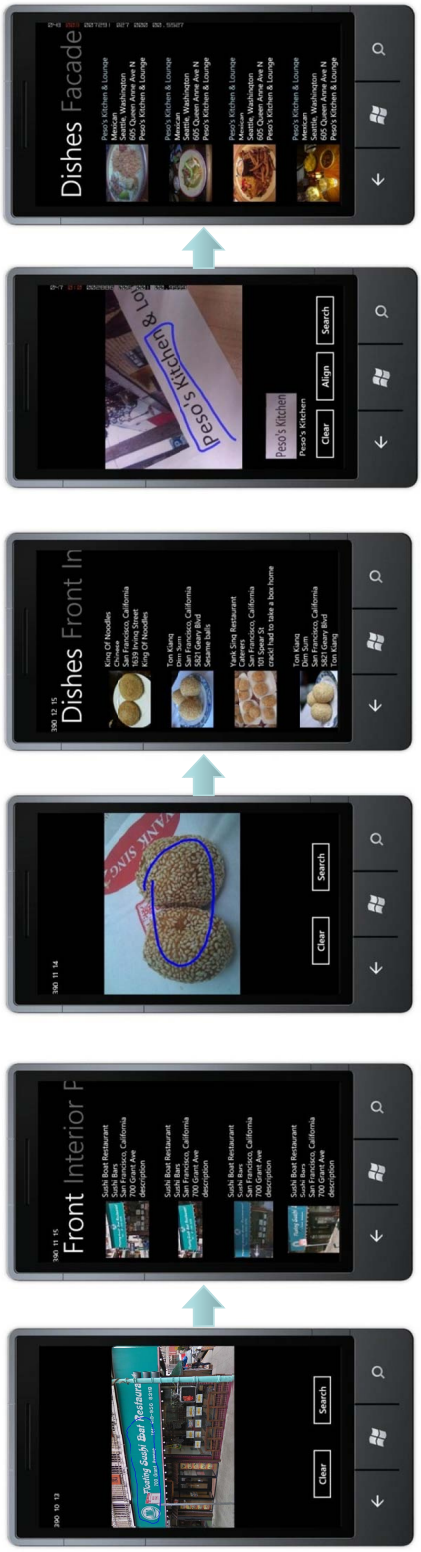
indicate your intended dish or text in the photo. Your visual intent on either the photo or the description of the dish will be analyzed. For example, optical character recognition (OCR) can help you automatically recognize the indicated text, while a visual search can help you identify the dish (which may not be recognized without indication) and recommend nearby restaurants serving a similar dish.

Figure 4.6 shows three corresponding scenarios. The visual intent model consists of two parts: visual recognition by search and social task recommendation. The first problem is to recognize what is captured (e.g., a food image), while the second is to recommend related entities (such as nearby restaurants serving the same food) based on the search-based recognition results. This activity recommendation is a difficult task in general, since visual recognition in the first step still remains challenging. However, the advanced functionalities, such as natural multi-touch interaction and a set of available rich context on the mobile device, bring us opportunities to accomplish this task. For example, although one image usually contains multiple objects, a user can indicate an object or some text of interest through a natural gesture, so that visual recognition can be reduced to search a similar single object. Moreover, the contextual information, such as geo-location, can be used for location-based recommendations.

Since the proposed *visual intent* is an original term, we retrospect the evolution of intent in general and walk the readers through the formation of the *intent* from text, voice, and visual inputs, with both desktop-based and mobile domain-based searches and recognition.

4.3.1 Related Work

For desktop user intent mining, an early study on web search taxonomy is introduced by Broder [139]. In this work, the most searched items belong to an "informational" category, in which it sought for related information to answer certain questions in a user's mind. A later work from Rose and Levinson further categorized the informational class to five sub-categories, where the *locate* of a product or service occupies a large percentage [140]. On the other hand, compared to general web searches, intents derived from mobile information have strong on-the-go characteristics. Church and Smyth conducted a diary study of user behavior of mobile-based text search and summarized a quite different categorization from its general web search counterpart [141]. Besides the informational



(a) building façade (b) dishes (c) OCR

Figure 4.6: Snapshots of *TapTell* with three different scenarios. A user can take a photo, specify the object or text of his/her interest via different gestures (e.g., tap, circle, or line), and then get the search and recommendation results through *TapTell*.

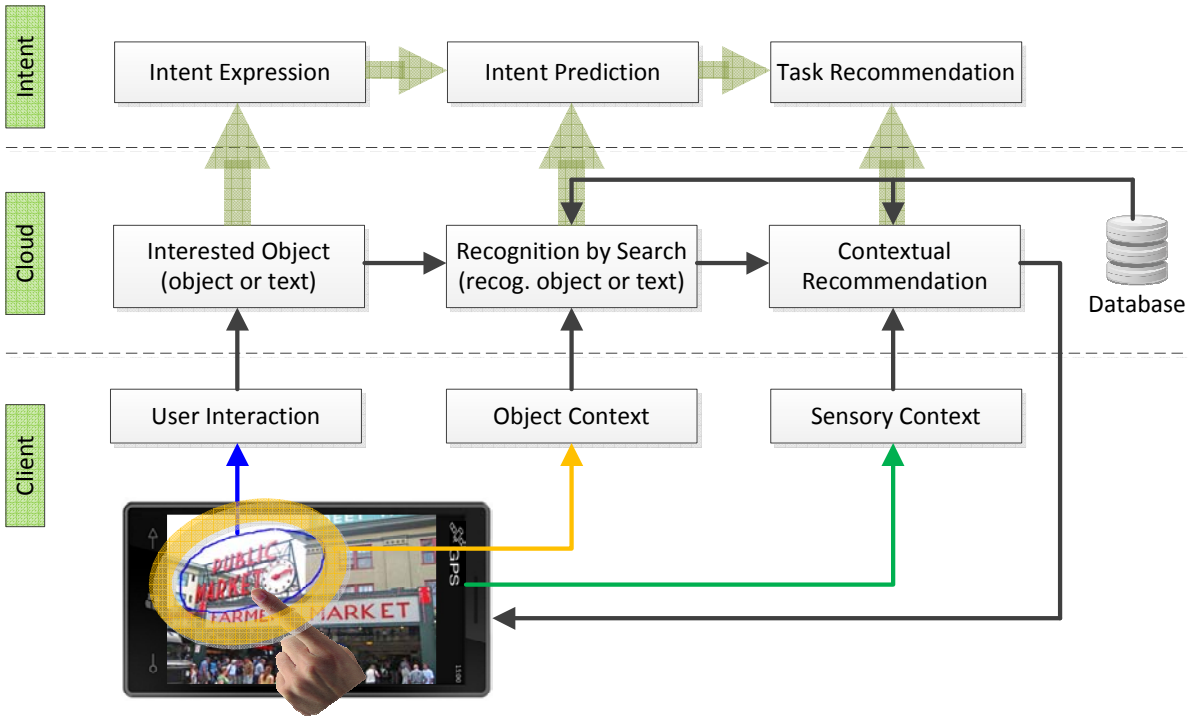


Figure 4.7: The framework of *TapTell*, based on previously introduced visual recognition algorithm in Figure 4.1, incorporates with the visual intents notation.

category at 58.3%, a new geographical category which is highly location dependent takes a share of 31.1% of total search traffic. From a topic perspective, *local services* and *travel & commuting* are the most popular ones out of 17 total topics, with 24.2% and 20.2% entries respectively. It can be concluded that the on-the-go characteristics play an important role for intent discovery and understanding on mobile devices [142].

4.3.2 Overview

Figure 4.7 shows the framework of *TapTell*. It extends Figure 4.1 by including user intent. This illustration can assist readers from an implementation perspective to understand the importance in linking individual intents to final recommendations. Intent expression recognizes the object specified by the user-mobile interaction. Intent prediction formulates intent expression and incorporates image context. Finally, a task recommendation is achieved by taking both the predicted intent, as well as, the sensory context.

In the following, Section 4.3.3 presents a conducted survey and explains why the

(a) tap (b) line (c) O: circle

Figure 4.8: Different gestures for specifying user intent in *TapTell*: (a) "tap"—selection of image segments, (b) "line"—rectangular box, and (c) "O"—circle or lasso.

"O" gesture is chosen as the best solution among several gesture candidates. With the "O" gesture and selected ROI, visual recognition by search is achieved using the algorithm introduced in the previous section. Consequently, Section 4.3.4 describes the recommendation, using text metadata associated with visual recognition to achieve a better re-ranking.

4.3.3 User Interaction for Specifying Visual Intent

It has been studied and suggested that visual interface will improve mobile search experiences [143]. In this section, we have performed a user study to identify the most natural and efficient gesture for specifying the visual intent on mobile devices. By taking advantages of multi-touch interaction on smart-phones, we defined three gestures for specifying visual intents on captured photos as follows:

- **Tap.** A user can "tap" on the pre-determined image segments, in which a captured image is automatically segmented on-the-fly. Then, the tapped segments indicated

by user's gesture will be connected as the region-of-interest (ROI). The ROI will be further used as the visual query, as shown in Figure 4.8 (a).

- **Line.** A user can draw straight "lines" to form a rectangular bounding box. The region in the box will be used as the visual query, as shown in Figure 4.8 (b).

- **O (*circle*).** A user can naturally outline an object of irregular shape. The "O" gesture can be also called the *circle* or *lasso*. Note that an "O" is not limited to a circle, but any arbitrary shape, as shown in Figure 4.8 (c).

We performed a user study following the principles of focus group in the field of human-computer interaction [144]. In this study, 10 participants were invited. After being introduced to the basic functions of *TapTell* and getting familiar with the system, they were asked to perform several tasks using different gestures in 30 minutes. From this study, we found that 7 out of 10 subjects thought that "O" is more natural than the other two gestures, and 8 subjects were satisfied with the "O" interaction. Their comments on "tapping" and "line" are: 1) tapping is sometimes too sensitive and image segmentation is not always satisfying, and 2) the "line" is not convenient for selecting an arbitrary object.

Equipped with the "O" gesture and the user interaction platform, mobile search and recognition can be achieved effectively using the proposed method. The next step of *TapTell* is to recommend social activities based on associated metadata and text-based search.

4.3.4 Social Activity Recommendations

Recently, Jain and Sinha proposed to re-examine the fundamental issue between content and context and why researchers should utilize both of them to bridge the semantic gap [145]. From the perspective of visual content analysis, Hua and Tian surveyed the importance of visual features to help text-based searches [146]. Although the aforementioned two studies focused on context and visual contents, respectively, they both advocate on a multi-modality structure to achieve various tasks. On the other hand, Guy *et al.* suggest that while machine learning and human computer interactions play key roles in recommendations, personalization and context-awareness are also crucial in

establishing an efficient recommendation system [147]. We agree with their arguments that it is necessary to connect data and users. We also believe that smart-phones provide perfect platforms for such data-users connection, from human computer interaction, to visual search, and finally, to the recommendation.

In the *TapTell* system, after the visual intent expression and identification, we utilize rich metadata as a better feature to search. We also use powerful context to re-rank metadata-based search result for the final task completion. To be specific, we adopt the metadata associated with the top image search result as our textual query. Then, we obtain the social activity recommendations based on the text retrieval results. The Okapi BM25 ranking function is used to compute a ranking score based on text similarity [148]. We extract the keywords $Q_t = \{q_{t_1}, q_{t_2}, ..., q_{t_n}\}$ by projecting the text query to a quantized text dictionary. Then, we compute the relevance score of query Q_t and database image descriptions D_t. Detailed score computation techniques can be referred to in reference [148]. In the last step, we re-rank the search results based on the GPS distance of the user's current location. Figure 4.9 demonstrates a sample result of the recommendation list and location-based re-ranking.

4.4 Experiments

Experiments on evaluating proposed context-embedded visual recognition, social activity recommendations through the *TapTell* system, subject evaluation, system performance and complex analysis, and OCR performance, are presented in the following.

4.4.1 Data and Settings

The client-end application is developed on a Windows Phone 7 HD7 model with 1GHz processor, 512MB ROM, GPS sensor and 5 megapixel color camera. In the cloud, a total of one million visualwords is built from 100 million sampled local descriptors (SIFT in this experiment). A hierarchal tree structure consisting of six levels of branches is used, where each superior branch has 10 sub-branches or nodes. In constructing the vocabulary tree, each visualword takes up to 168 bytes storage, where 128 bytes are for the clustering vector (same size as SIFT), and 4 bytes for ten subordinate children nodes connection. In total, 170 megabytes of storage is used for the vocabulary tree in cache.

Figure 4.9: Result of recommendation list, which is visualized in a map to help users to picture the distances between the query and the results.

The dataset consists of two parts. One is from Flickr, which includes a total of two million images, with 41,614 landmarks equipped with reliable GPS contextual information. With a further manual labeling effort, 5,981 images were identified as the groundtruth such that the landmark object façade or the outside appearance can be traced from the image. The second part of the database is a crawled commercial local services data, mainly focusing on the restaurant domain. In this part, a total of 332,922 images associated with 16,819 restaurant entities from 12 US cities were crawled with associated metadata.

4.4.2 Evaluation Metrics

We use mean average precision (MAP) for the evaluation, where MAP is the mean value of average precisions (APs). The average precision (AP) formula is presented as

$$AP@n = \frac{1}{min(n,P)} \sum_{k=1}^{min(n,S)} \frac{P_k}{k} \times I_k \tag{4.8}$$

The number of top ranks is represented as n. The size of the dataset is denoted as S, and P is the total number of positive samples. At index k, P_k is the number of positive results in the top n returns, and I_k is described as the result of the k_{th} position.

Another performance metric we adopt is Normalized Discounted Cumulative Gain ($NDCG$). Given a query q, the $NDCG$ at the depth d in the ranked list is defined by:

$$NDCG@d = Z_d \sum_{j=1}^{d} \frac{2^{r^j} - 1}{\log(1+j)} \tag{4.9}$$

where r^j is the rating of the j-th pair, Z_d is a normalization constant and is chosen so that the $NDCG@d$ of a perfect ranking is 1.

4.4.3 Objective Evaluations

Evaluation of location-based recognition

In Figure 4.10, the proposed CVT-based CBIR method with and without location-based GPS filter is evaluated in both MAP and NDCG measurements for different database sizes. In this case, original image query is used without any visual intent regulation.

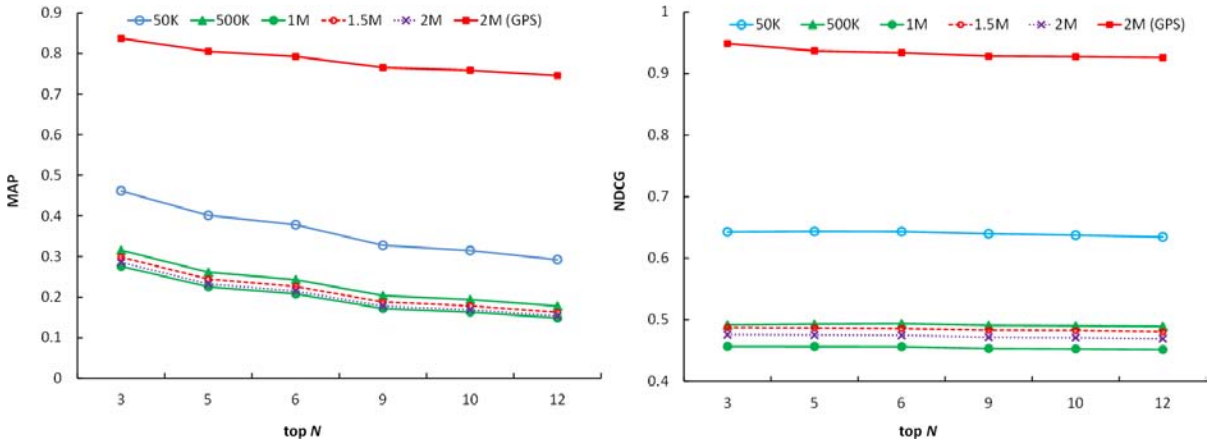


Figure 4.10: Top N returns for both MAP and NDCG evaluations with GPS context, on the whole image itself as query.

The performance suffers a degradation with the increment of database size. For the location-based recognition method, images with related geographical regions have been firstly isolated from irrelevant images, and then, recognition by search algorithm is implemented solely on the filtered dataset. Performance is maintained and demonstrates that the proposed system is applicable for dealing with large-scale databases. For the location-based filter $\phi(\mathbf{q})$, the GPS effective region Q utilizes the Quadkey level 5, which is equivalent to the resolution of 4891 meters in ground. Since landmarks groundtruth includes various object types: from statuaries and buildings, to city skylines and famous mountains, the aforementioned contextual filter will guarantee the inclusion of enough potential image candidates. In summary, such an analysis and investigation demonstrate the usage of location-based filter as an important tool in mobile visual search and recognition.

Evaluation of context-embedded visual recognition

We investigated image contextual information and its effectiveness in recognition by search technique, using the soft weighting scheme. For the bivariate-based function $\Re(x, y)$, we fixed the amplitude A to 1 and tuned two parameters α and β to modulate the standard deviation. We conducted two sets of experimentation with and without GPS context shown in Figure 4.11 and Figure 4.12, respectively. In general, using the soft weighting scheme improves search performance compared to the binary weighting

method. Specifically, in Figure 4.11, $\alpha = 50$ and $\beta = 10$ provide the best performance for both MAP and NDCG measurements. The results of this parameter choice using MAP and NDCG measures outperform the binary weight method by 12% and 15%, respectively.

Similarly, after incorporating the GPS context, the soft weighting method again outperformed the binary one, but in a much higher precision range. This does not surprise us since geolocation is an important feature for differentiating objects and their recognition, and eventually associated visual intent. Different from its counterpart in the non-GPS scenario, Figure 4.12 demonstrates that parameter $\alpha = 5$ and $\beta = 1$ outperforms other parameter choices, as well as the baseline binary weighting scheme. The margin difference from the soft weighting and the binary case has dropped to 2% and less than 1% for MAP and NDCG, respectively. This result demonstrates the importance of the GPS context.

It can be observed that parameter α is higher than parameter β for the best performance in both Figure 4.11 and Figure 4.12. The reason is due to the fact that most images are taken horizontally. Therefore, information is appreciated more and weighted higher by α horizontally than its counterpart β vertically. Similar patterns can also be observed in the following evaluations.

The significance of this image contextual information with soft weighting scheme allows robust user behavior and is seamlessly glued with the "O" gesture, which is spontaneous and natural. The shortcoming of the"O" is that it inevitably suffers from lack of accuracy due to device limitations in outlining the boundary, compared to other gestures, such as segmentation or line-based rectangular shape. However, soft weighting alleviates this deficiency of correctness in object selection and provides a robust method to accommodate behavioral errors when drawing the outlines of the ROI.

Evaluation and comparison with contextual image retrieval model (CIRM)

We also implemented a state-of-the-art contextual image retrieval model (CIRM) [149] and compared its performance to our proposed context-embedded visual recognition. The CIRM has demonstrated a promising result in desktop-based CBIR by applying a rectangular bounding box in highlighting the emphasized region, which can be achieved using mouse control at a desktop platform. The weighting scheme in CIRM model is to

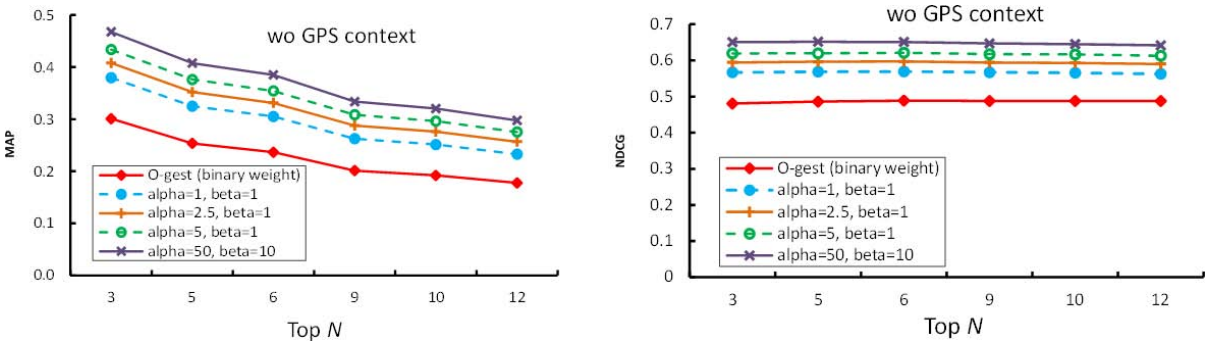


Figure 4.11: Image contextual-based recognition by various parameter α and β, without GPS information.

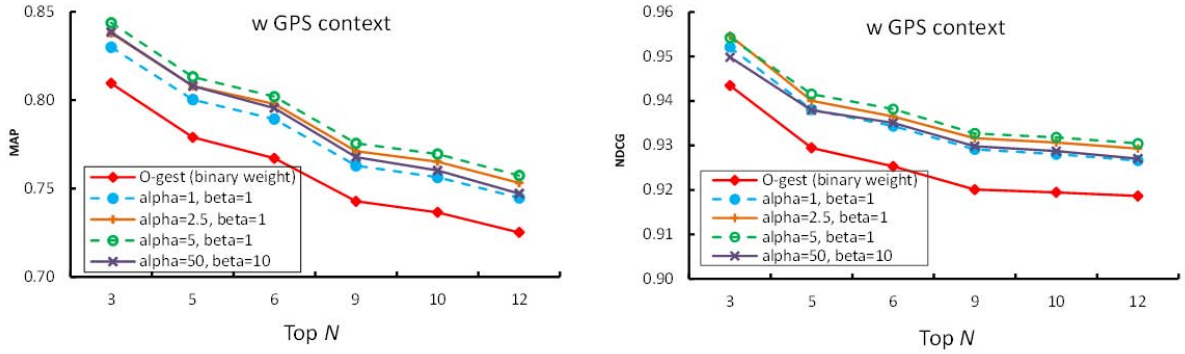


Figure 4.12: Image contextual-based recognition by various parameter α and β, with GPS information.

use two logistic functions joined at the directional (either X or Y) center of the bounding box. Then, the term frequency tf_q is formulated as:

$$\begin{aligned} tf_q \propto \quad & min\left(\frac{1}{1+exp(\delta_X(x_l-x_i))}, \frac{1}{1+exp(\delta_X(x_i-x_r))}\right) \\ * \quad & min\left(\frac{1}{1+exp(\delta_Y(y_t-y_i))}, \frac{1}{1+exp(\delta_Y(y_i-y_b))}\right) \end{aligned} \tag{4.10}$$

where x_l, x_i, x_r represent x pixel values of the left boundary, detected feature point, and the right boundary along the x-axis direction, respectively. Similarly, y_t, y_i, y_b are the y pixel values of the top boundary, detected feature point, and the bottom boundary along the y-axis, respectively. The geometric relations $x_l < x_i < x_r$ and $y_t < y_i < y_b$ hold for this bounding box, such that the tf_q should be approaching the value 0, the

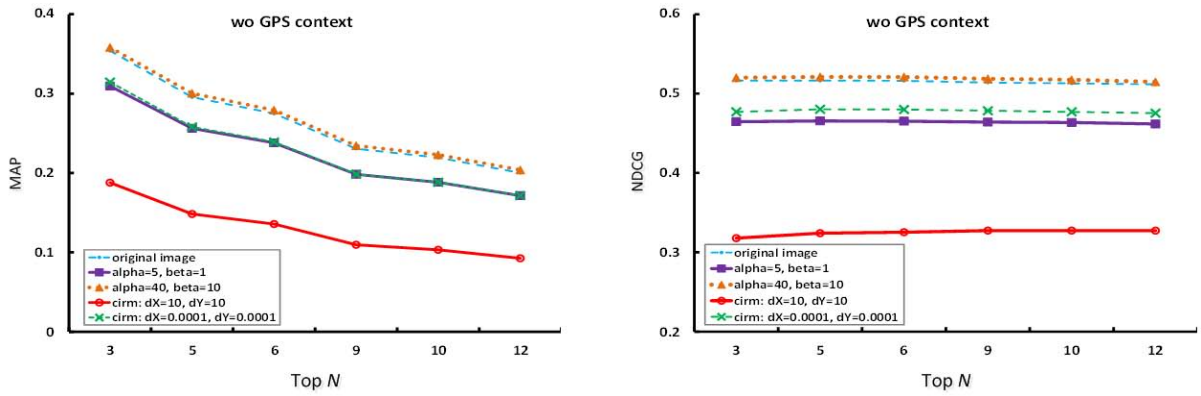


Figure 4.13: Comparison of image contextual-based recognition by various parameter α and β, with the conventional CBIR (original), as well as the CIRM algorithm with parameter dX and dY, without GPS information.

further x_i from the bounding box; while ideally close to value 1 when the feature point is inside the bounding box. δ_X and δ_Y are two tunable parameters for finding the best performance of the bounding box. Detailed explanation of the algorithm can be found in reference [149].

Figure 4.13 shows MAP and NDCG measurements, by comparing the proposed Gaussian-based contextual method with the CIRM model, as well as the CBIR method using the original image. It appears that the proposed method with parameters $\alpha = 40$ and $\beta = 10$ outperformed both CIRM in its best result with parameter $dX = 0.0001$ and $dy = 0.0001$, and the CBIR result of the original image without using contextual model.

Figure 4.14 depicts a similar comparison using the GPS context re-ranking. Again, the proposed method outperformed the CIRM method and the CBIR algorithms. However, the best performance of the CIRM model at $dX = 0.0001$ and $dY = 0.0001$ is close to the performance of our proposed contextual model at $\alpha = 5$ and $\beta = 1$. This result can be explained, such that, by adopting the GPS filtering, the margin of various methods is reduced.

Evaluation of mobile recommendations

For the recommendations, our method is to use the visual photo taken by users as the starting point, and to provide recommendation lists based on text searches associated with the recognized object. First, we identify the object and match it to the database.

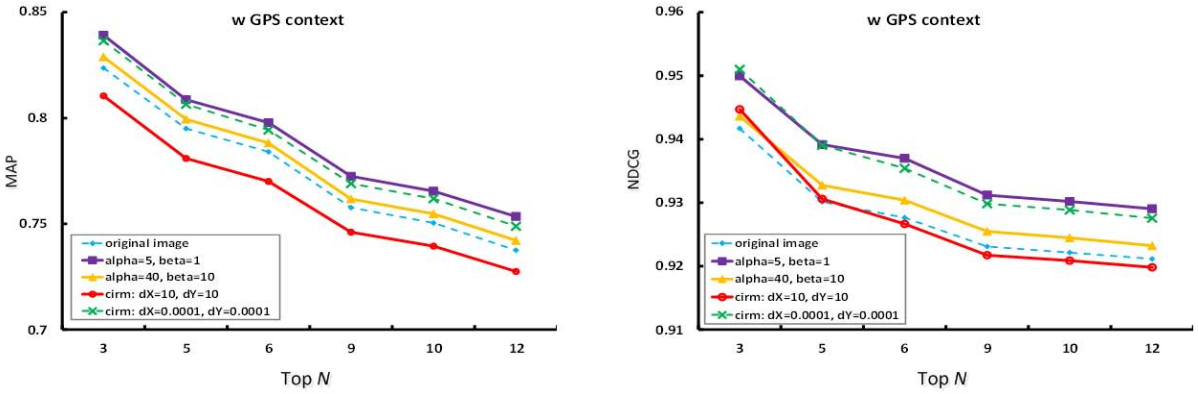


Figure 4.14: Comparison of image contextual-based recognition by various parameter α and β, with the conventional CBIR (original), as well as the CIRM algorithm with parameter dX and dY, with GPS information.

Then, we use the matched metadata as a text query to do a text-based search. The final result is then re-ranked by the relevant GPS distance from the query's image location to the ranked list image locations.

The evaluation was conducted exclusively on a vertical domain of food cuisines. We randomly picked 306 photos and manually labeled and categorized them into 30 featured themes of food dishes, such as beef, soup, burger, etc. We built a 300 word text dictionary by extracting the most frequently used words in the image description.

In order to produce a real restaurant scenario, we printed out dishes in a menu style with both texts and images. We took pictures of the dishes as the visual query and attempted to find the duplicated/near-duplicated images from the dataset. We assumed that the best match of the visual recognition result would be user intent. Such intent was carried by the associated metadata, which were quantized using the prepared 300-word dictionary. The quantized words were searched with a ranked list based on the text similarity. The final step was to re-rank the result list using GPS distance.

Table 4.2 presents the MAP result with the initial visual query and newly formatted text description query after visual recognition. The Table demonstrates that the performance of the text description-based search is much better than the visual-based search. This result is reasonable in the sense that text is a better description than visual content once the ROI is identified and linked with precise textual metadata. However, the merit of the visual input is its role in filling the niche when an individual does not

Table 4.2: MAP evaluation of the visual-based and description-based performance.

MAP	@0	@1	@2	@3	@4
Visual-based	96.08	53.06	37.61	29.60	24.59
Description-based	n/a	75.65	72.66	70.78	65.93

Table 4.3: A summary of the subjective survey.

Q#	**Valid Result**	**Criteria**	**1**	**2**	**3**	**4**	**5**	**Avg.**
1	10	Useful	0	1	2	1	6	4.2
2	10	Satisfied	0	1	1	3	5	4.2
3	10	Satisfied	0	1	1	4	4	4.1
4	10	Satisfied	0	2	2	2	4	3.8
5	9	Useful	0	1	1	3	4	4.11
6	10	Useful	0	1	3	2	4	3.9
7	10	Useful	0	1	1	4	4	4.1
8	10	Useful	0	1	1	4	4	4.1
9	10	Useful	0	1	2	3	4	4.0

Note: A scale of 1 to 5 is used, with 5 indicating the most useful/satisfied level, 1 indicates the least useful/satisfied level, and 3 is the neutral.

have the language tools to express him/herself articulately. We demonstrate that during the initial visual search (@0), the visual-search result is at a high precision rate of 96.08%. Such accuracy provides a solid foundation to utilize associated metadata as a description-based query during the second stage search. In summary, once the visual query is mined accurately, the role of the search query is then shifted from visual content to text metadata for a better result.

4.4.4 Subjective Evaluation

We also conducted a subjective evaluation on user experience with the *TapTell* system. A total of 13 people participated the survey, nine male and four female. Eight out of the total participants had heard of the term content-based image retrieval, and six of them had heard of a natural user interface. During the survey, they were asked about the usefulness of and satisfaction with the proposed system based on their experience using

the prototype. The survey scale is ranked from 1 to 5 for usefulness and satisfaction, where 1 is the least and 5 is the most. Table 4.3 summarizes the survey result.

- Question 1 and 2 are about the usefulness of the "O" gesture compared to segmentation and line-based gestures, and the satisfaction of the "O" interface.

- Question 3 and 4 are about visual search satisfaction on duplication/near-duplication results, as well as semantic similar results. The rate is higher for the former, which is a fair reflection of the algorithm we took. This is because we use salient-based SIFT points, which are more suitable for duplication/near-duplication detection than object recognition.

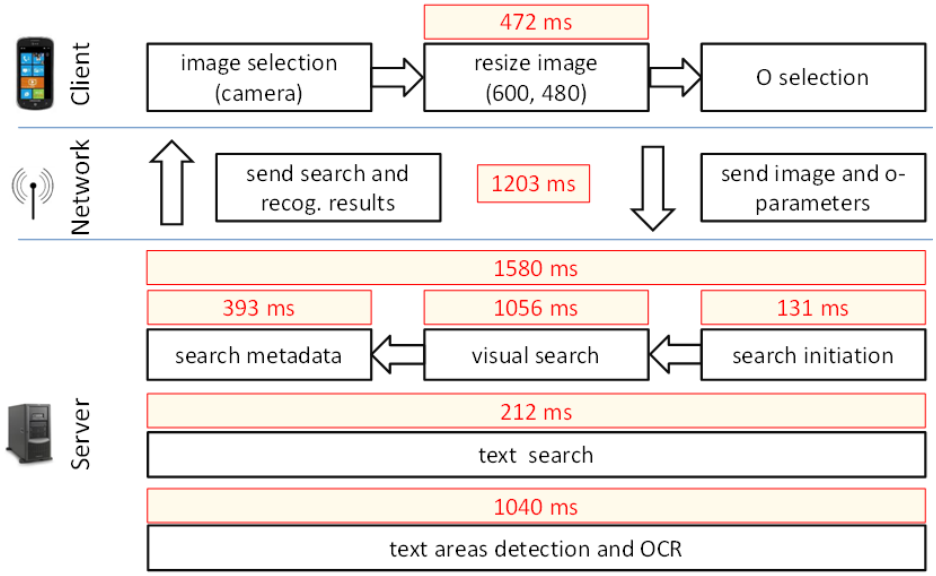
- Question 5 and 6 are the usefulness study on the Optical Character Recognition (OCR) technique and adopted transformation invariant low-rank textures (TILT) for improving the OCR. More people are in favor of the TILT algorithm enhanced OCR method than the OCR itself [150]. (Technical details are presented in Section 4.4.6).

- Question 7 is about the performance of text-based searches. Most people are satisfied with this feature.

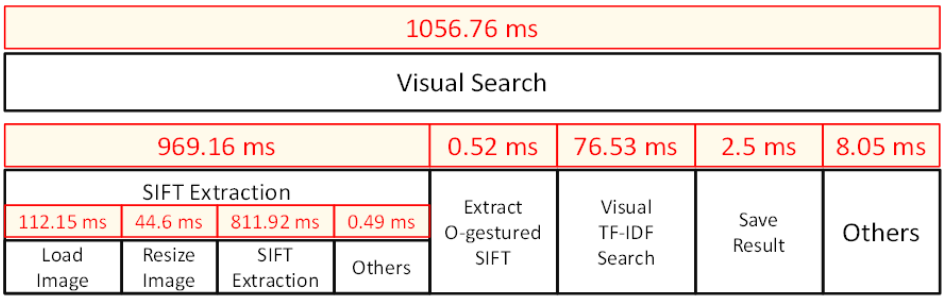
- Question 8 and 9 are about the overall usefulness in terms of a recommendation system and *TapTell* as an application for mobile devices. Most people gave positive response to the usefulness of this system for both recommendations, as well as the application in general.

- The last question asks a price (in USD) they would be willing to pay at the mobile market to obtain this application. Eight out of 10 people prefer a price less than $4.99, where two are not willing to pay anything. The remaining two participants are willing to pay a price above $10.

On average, questionnaire participants were satisfied with the *TapTell* system. Most responses were either 4s or 5s on the 5-point scale. They also provided insightful comments such as



(a) Total time spent of the *TapTell* system.



(b) Visual search time spent.

Figure 4.15: The time analysis of the *TapTell* system as well as the visual search, based on the restaurants dataset.

Quote 1 "Maybe can cooperate with the fashion industry."

Quote 2 "This is quick and natural. Better than pre-segmented based method. The segment results are always confusing."

4.4.5 Time Complexity Analysis

TapTell's efficiency performance of the individual component is evaluated. A detailed analysis is illustrated in Figure 4.15. The total time spent on the server end takes about 1.6 seconds, including initialization, text-based search, visual-based search, and OCR-

based recognition (we also support OCR if the ROI corresponds to a text region). Among the visual search, local descriptor SIFT extraction takes the most time, almost 1 second. The communication time between the server and the client takes about 1.2 seconds, which is the wireless transmission in our experimental set-up.

4.4.6 Improved OCR from "O"

Besides the visual content, Optical Character Recognition (OCR) is another important means to help mobile users to understand their visual intents correctly. It plays a vital role in translating from the visual feature to the text feature. However, most of the OCR techniques are sensitive to the orientation of visual input. If characters are skewed in a certain degree, current OCR techniques cannot successfully recognize the correct characters. However, such a difficulty can be alleviated by using a transform invariant low-rank textures (TILT) algorithm to align the severely tilted characters properly [150].

We found that one of the byproducts from the "O" gesture is that we can achieve better OCR performance if we utilized the estimation results of two principal components by the PCA in Section 4.3.3. Once the original text region is selected by the "O" gesture, those characters are first aligned by performing rotation alignment based on the PCA result, and then, further aligned by the TILT algorithm before the OCR process. Figure 4.16 illustrates a successful OCR detection.

4.4.7 Video Demonstration

We also have uploaded a video demo to showcase the *TapTell* system. The video speed is set to x1.7 more than the original footage to make this video demo more compact and agreeable to watch [5].

4.4.8 Visual Examples

Two visual examples are demonstrated in Figure 4.17 with the visual queries associated location metadata of (a): Bleecker Street Pizza, located at 69 7th Ave S. New York. (b): Beef Marrow and Marmalade, located at 97 Sullivan St. New York.

[5] `http://www.viddler.com/explore/Mm11132/videos/1/`

(a) original image with O (b) OCR fails without O (c) OCR with "O + alignment"

Figure 4.16: Standard OCR failed to recognize multiple lines of skewed characters, but is successful after using the "O + TILT alignment" procedure.

4.5 Summary

A contextual-based mobile visual search utilizing the BoW model is proposed in this Chapter. A viable application, *TapTell*, is implemented to achieve mobile recognition and recommendations. Meaningful social tasks and activities are suggested to users with the assistance of multimedia tools and rich contextual information in the surroundings. We have investigated different gestures from tapping the segments, to drawing the lines of rectangle, to making an "O"-circle via the multi-touch screen. We demonstrated that the "O" behavior is the most natural and agreeable user-mobile interaction. Along with the BoW model, a context-embedded vocabulary tree for soft weighting is proposed by using both "O" object and its surrounding image context to achieve mobile visual intents mining. We evaluated various weighting schemes with and without GPS conditions, and verified that image context outside the "O" region plays a constructive role in improving the recognition. We also compared our method with the state-of-the-art algorithms and it has demonstrated that the proposed method outperformed both the conventional CBIR using original image query and the CIRM algorithm. Moreover, a recommendation system is built upon an initial visual query input, where neither the text nor the voice

(a) Bleecker Street Pizza

(b) Beef Marrow and Marmalade

Figure 4.17: Visual examples based on the recommendation system. The left snapshot shows the visual query. The middle snapshot is the result using metadata-based text search. The right snapshot is the re-ranking based on user's current position and location-based distance.

has the strength in describing the visual intent. Once the context metadata is associated with the intent, more reliable contextual text and GPS features are taken advantage of in searching and re-ranking. Ultimately, interesting and related social activities are recommended to the users.

Chapter 5

Conclusions and Future Work

5.1 Thesis Summary

This thesis focuses on large-scale unlabeled image/video data and proposes multimedia analysis and approaches by integrating the bag-of-words (BoW) model with image based classification and retrieval. In particular, we proposed a systematic video analysis framework and a mobile based visual search with recommendation system, based on image classification and retrieval methodologies, respectively. The BoW model was applied to images and video frames by incorporating local scale-invariant feature descriptors (SIFT). A codebook was built based on uniformly sampled local descriptor data. Each image or video frame was then mapped onto the codebook to form a BoW representation.

In the systematic and generic video analysis framework proposed in chapter 3, the BoW model was used to represent video frames and clips. Codebook generation was achieved by an innovative two-layer bottom-up K-means clustering. In this way, computational efficiency was improved compared to the conventional single K-means clustering. Using the BoW model based representation, three levels of the video analysis were investigated from a large-scale sports video dataset with 23 types. First, an unknown video clip was categorized by its genre using the K-nearest neighboring algorithm. Once the genre was decided, middle level views were learned and classified using an unsupervised PLSA model. Finally, the result of view classification in the form of a labeled sequence of video frames were fed into a HCRF model to achieve final high-level semantic event detection. This proposed framework is generic and requires minimum human input. Therefore, it is

ideal for processing large-scale multimedia data. The experimental performance demonstrates that an unsupervised PLSA model as input is comparable in event detection accuracy, compared to its supervised SVM counterpart. However, labeling work can be saved by PLSA learning. Thus, it makes the proposed framework scalable to an even larger data consortia with more diverse genres.

In chapter 4, visual search, and consequently, social task recommendations were achieved on mobile platforms. User intention in form of visual queries was obtained by an interactive platform provided by the natural user interface and the advance of mobile multi-touch technology. By understanding user input and incorporating that with the BoW model, a context-embedded vocabulary tree (CVT) was built to generate a hierarchical visual codebook. Query images consisting of both the ROI segment selected by the "O" gesture and its surrounding context were mapped onto the codeobook using a Gaussian-distance based weighting scheme. Using this method, the query image was treated with an emphasis on the ROI while also including the background information. Experiments show that the proposed algorithm outperformed both the conventional CBIR using the whole image as query and the single ROI segment as query. It also demonstrated that the proposed CVT, using soft Gaussian-distance weighting, outperformed a desktop CBIR algorithm (called CIRM), which used a logistic function weighting scheme. An implementation called "TapTell" was engineered to achieve mobile visual search and mine users' visual intent for social activity recommendations. GPS information and OCR techniques were also adopted during this implementation process to achieve a better understanding of visual intent.

5.2 Future Work

Despite the advancement and research focus on the BoW model in image classification and retrieval with large-scale multimedia analysis and applications, there are still many opportunities to further extend and improve the BoW model by utilizing more representative local descriptors, generating more robust and extensive codebooks, discovering spatial connections between local descriptors, and incorporating text and audio multimodalities. The following directions are worth further exploration.

- Although various local features have been designed and applied successfully in var-

ious applications, there is still a need of more powerful and robust local descriptors to face ever growing large-scale and complicated visual data.

- Codebook construction is an important stage of the BoW model. Efficient codebook generation and concise representation methods are crucial in accurately mapping visual information, which are highly appreciated in large-scale multimedia applications. In addition, there is also a promising future in developing multiple codebooks, each of which has its own focus on the feature space while maintains connection with each other.

- The BoW model is unordered. This means that no priorities are given to particular visualwords inside the BoW. However, certain developed local descriptors carry more representative information than others. How to prioritize information inside the BoW model is a worthy investigative direction, which may be related to spatial context and semantic labeling.

- The BoW model is a content-based analysis method. Despite some effort in this thesis to link the BoW based image retrieval with indexed metadata, connections between visual contents and modalities, such as audio and text, are worthy further research in order to achieve an optimized multimodal solution.

Bibliography

[1] L. Fei-Fei, "Two bag-of-words classifiers," International Conference on Computer Vision (ICCV) short courses on, 2005.

[2] D. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[3] D. Nister and H. Stewenius, "Scalable recognition with a vocabulary tree," in *Proc. IEEE CVPR*, 2006, pp. 2161–2168.

[4] R. Bohn and J. Short, "How much information? 2009 Report on American Consumers," *University of California at San Diego, Global Information Industry Center*, 2010.

[5] R. Yan and W. Hsu, "Content-based and concept-based analysis for large-scale image/video retrieval," in *Proc. ACM MM*, 2009, pp. 913–914.

[6] A. Ekin and A. Tekalp, "Framework for tracking and analysis of soccer video," in *Proc. SPIE VCIP*, vol. 4671, 2002, pp. 763–774.

[7] L. Duan, M. Xu, and Q. Tian, "Semantic shot classification in sports video," in *Proc. SPIE*, 2003, pp. 300–313.

[8] Y. Rui, T. Huang, and S. Chang, "Image retrieval: Current techniques, promising directions, and open issues," *Journal of visual communication and image representation*, vol. 10, no. 1, pp. 39–62, 1999.

[9] R. Datta, D. Joshi, J. Li, and J. Wang, "Image retrieval: Ideas, influences, and trends of the new age," *ACM Computing Surveys (CSUR)*, vol. 40, no. 2, p. 5, 2008.

BIBLIOGRAPHY

[10] T. Joachims, "A probabilistic analysis of the rocchio algorithm with tfidf for text categorization." DTIC Document, Tech. Rep., 1996.

[11] D. Lewis, "Naive (bayes) at forty: The independence assumption in information retrieval," *Machine Learning: ECML-98*, pp. 4–15, 1998.

[12] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic *et al.*, "Query by image and video content: The qbic system," *Computer*, vol. 28, no. 9, pp. 23–32, 1995.

[13] C. Faloutsos, R. Barber, M. Flickner, J. Hafner, W. Niblack, D. Petkovic, and W. Equitz, "Efficient and effective querying by image content," *Journal of intelligent information systems*, vol. 3, no. 3, pp. 231–262, 1994.

[14] J. Smith and S. Chang, "Visualseek: a fully automated content-based image query system," in *Proceedings of the fourth ACM international conference on Multimedia*. ACM, 1997, pp. 87–98.

[15] A. Pentland, R. Picard, and S. Sclaroff, "Photobook: Content-based manipulation of image databases," *International Journal of Computer Vision*, vol. 18, no. 3, pp. 233–254, 1996.

[16] J. Bach, C. Fuller, A. Gupta, A. Hampapur, B. Horowitz, R. Humphrey, R. Jain, and C. Shu, "Virage image search engine: an open framework for image management," in *Proceedings of SPIE*, vol. 2670, 1996, p. 76.

[17] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *Proc. ICCV*, vol. 2, 2003, pp. 1470–1477.

[18] ——, "Efficient visual search of videos cast as text retrieval," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 4, pp. 591–606, 2009.

[19] O. Cula and K. Dana, "Compact representation of bidirectional texture functions," in *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, vol. 1. IEEE, 2001, pp. I–1041.

BIBLIOGRAPHY

[20] T. Leung and J. Malik, "Representing and recognizing the visual appearance of materials using three-dimensional textons," *International Journal of Computer Vision*, vol. 43, no. 1, pp. 29–44, 2001.

[21] M. Varma and A. Zisserman, "Classifying images of materials: Achieving viewpoint and illumination independence," *Computer VisionECCV 2002*, pp. 255–271, 2002.

[22] S. Lazebnik, C. Schmid, and J. Ponce, "A sparse texture representation using affine-invariant regions," in *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, vol. 2. IEEE, 2003, pp. II–319.

[23] L. Fei-Fei and P. Perona, "A bayesian hierarchical model for learning natural scene categories," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 2. Ieee, 2005, pp. 524–531.

[24] A. Bosch, A. Zisserman, and X. Muoz, "Scene classification using a hybrid generative/discriminative approach," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 30, no. 4, pp. 712–727, 2008.

[25] J. Sivic, B. Russell, A. Efros, A. Zisserman, and W. Freeman, "Discovering object categories in image collections," 2005.

[26] L. Cao and L. Fei-Fei, "Spatially coherent latent topic model for concurrent segmentation and classification of objects and scenes," in *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on.* Ieee, 2007, pp. 1–8.

[27] J. Shotton, M. Johnson, and R. Cipolla, "Semantic texton forests for image categorization and segmentation," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on.* IEEE, 2008, pp. 1–8.

[28] D. Ziou and S. Tabbone, "Edge detection techniques-an overview," in *International Journal of Pattern Recognition and Image Analysis.* Citeseer, 1998.

[29] T. Lindeberg, "Feature detection with automatic scale selection," *International journal of computer vision*, vol. 30, no. 2, pp. 79–116, 1998.

[30] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," *Lecture notes in computer science*, vol. 3951, p. 404, 2006.

BIBLIOGRAPHY

[31] K. Mikolajczyk and C. Schmid, "Indexing based on scale invariant interest points," in *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*, vol. 1. IEEE, 2001, pp. 525–531.

[32] ——, "Scale & affine invariant interest point detectors," *International journal of computer vision*, vol. 60, no. 1, pp. 63–86, 2004.

[33] M. Berger and M. Cole, *Geometry*. Springer Verlag, 1987, vol. 1.

[34] K. Mikolajczyk and C. Schmid, "An affine invariant interest point detector," *Computer VisionECCV 2002*, pp. 128–142, 2002.

[35] F. Schaffalitzky and A. Zisserman, "Multi-view matching for unordered image sets, or how do i organize my holiday snaps?," *Computer VisionECCV 2002*, pp. 414–431, 2002.

[36] T. Tuytelaars and L. Van Gool, "Content-based image retrieval based on local affinely invariant regions," in *Visual Information and Information Systems*. Springer, 1999, pp. 656–656.

[37] ——, "Wide baseline stereo matching based on local, affinely invariant regions," in *british Machine vision conference*, 2000, pp. 412–425.

[38] J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust wide-baseline stereo from maximally stable extremal regions," *Image and Vision Computing*, vol. 22, no. 10, pp. 761–767, 2004.

[39] T. Kadir and M. Brady, "Saliency, scale and image description," *International Journal of Computer Vision*, vol. 45, no. 2, pp. 83–105, 2001.

[40] T. Kadir, A. Zisserman, and M. Brady, "An affine invariant salient region detector," *Computer Vision-ECCV 2004*, pp. 228–241, 2004.

[41] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Gool, "A comparison of affine region detectors," *International journal of computer vision*, vol. 65, no. 1, pp. 43–72, 2005.

BIBLIOGRAPHY

[42] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 27, no. 10, pp. 1615–1630, 2005.

[43] Y. Ke and R. Sukthankar, "Pca-sift: A more distinctive representation for local image descriptors," in *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, vol. 2. Ieee, 2004, pp. II–506.

[44] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1. Ieee, 2005, pp. 886–893.

[45] V. Chandrasekhar, G. Takacs, D. Chen, S. Tsai, R. Grzeszczuk, and B. Girod, "CHoG: Compressed histogram of gradients A low bit-rate feature descriptor," pp. 2504–2511.

[46] D. Lowe, "Object recognition from local scale-invariant features," in *Proc. IEEE ICCV*, vol. 2, 1999, pp. 1150–1157.

[47] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," in *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, vol. 2. IEEE, 2003, pp. II–257.

[48] 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*, vol. 1. IEEE, 2001, pp. I–511.

[49] B. Girod, V. Chandrasekhar, D. Chen, N. Cheung, R. Grzeszczuk, Y. Reznik, G. Takacs, S. Tsai, and R. Vedantham, "Mobile visual search," *Signal Processing Magazine*, vol. 28, no. 4, pp. 61–76, 2011.

[50] V. Chandrasekhar, S. Tsai, Y. Reznik, G. Takacs, D. Chen, and B. Girod, "Compressing a set of chog features," in *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, vol. 8135, 2011, p. 39.

BIBLIOGRAPHY

[51] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," in *Workshop on statistical learning in computer vision, ECCV*, vol. 1, 2004, p. 22.

[52] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on.* Ieee, 2009, pp. 248–255.

[53] J. Deng, A. Berg, K. Li, and L. Fei-Fei, "What does classifying more than 10,000 image categories tell us?" *Computer Vision–ECCV 2010*, pp. 71–84, 2010.

[54] X. Zhou, K. Yu, T. Zhang, and T. Huang, "Image classification using super-vector coding of local image descriptors," *Computer Vision–ECCV 2010*, pp. 141–154, 2010.

[55] F. Perronnin, J. Senchez *et al.*, "Large-scale image categorization with explicit data embedding," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on.* IEEE, 2010, pp. 2297–2304.

[56] F. Perronnin, J. Sánchez, and T. Mensink, "Improving the fisher kernel for large-scale image classification," *Computer Vision–ECCV 2010*, pp. 143–156, 2010.

[57] F. Perronnin and C. Dance, "Fisher kernels on visual vocabularies for image categorization," in *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on.* IEEE, 2007, pp. 1–8.

[58] Y. Jiang, J. Yang, C. Ngo, and A. Hauptmann, "Representations of keypoint-based semantic concept detection: A comprehensive study," *IEEE Transactions on Multimedia*, vol. 12, no. 1, pp. 42–53, 2010.

[59] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Lost in quantization: Improving particular object retrieval in large scale image databases," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on.* IEEE, 2008, pp. 1–8.

[60] H. Jégou, M. Douze, C. Schmid, and P. Pérez, "Aggregating local descriptors into a compact image representation," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on.* IEEE, 2010, pp. 3304–3311.

BIBLIOGRAPHY

[61] H. Jégou, M. Douze, and C. Schmid, "Improving bag-of-features for large scale image search," *International Journal of Computer Vision*, vol. 87, no. 3, pp. 316–336, 2010.

[62] J. Sivic and A. Zisserman, "Video google: Efficient visual search of videos," *Toward Category-Level Object Recognition*, pp. 127–144, 2006.

[63] ——, "Video data mining using configurations of viewpoint invariant regions," in *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, vol. 1. IEEE, 2004, pp. I–488.

[64] T. Quack, V. Ferrari, and L. Van Gool, "Video mining with frequent itemset configurations," *Image and Video Retrieval*, pp. 360–369, 2006.

[65] J. Sivic and A. Zisserman, "Efficient visual search for objects in videos," *Proceedings of the IEEE*, vol. 96, no. 4, pp. 548–566, 2008.

[66] J. Sivic, F. Schaffalitzky, and A. Zisserman, "Object level grouping for video shots," *Computer Vision-ECCV 2004*, pp. 85–98, 2004.

[67] Y. Jiang, C. Ngo, and J. Yang, "Towards optimal bag-of-features for object categorization and semantic video retrieval," in *Proc. ACM CIVR*, 2007, p. 501.

[68] A. Basharat, Y. Zhai, and M. Shah, "Content based video matching using spatiotemporal volumes," *Computer Vision and Image Understanding*, vol. 110, no. 3, pp. 360–377, 2008.

[69] J. Law-To, O. Buisson, V. Gouet-Brunet, and N. Boujemaa, "Robust voting algorithm based on labels of behavior for video copy detection," in *Proceedings of the 14th annual ACM international conference on Multimedia*. ACM, 2006, pp. 835–844.

[70] J. Sivic, M. Everingham, and A. Zisserman, "Person spotting: video shot retrieval for face sets," *Image and Video Retrieval*, pp. 592–592, 2005.

[71] X. Zhou, X. Zhuang, S. Yan, S. Chang, M. Hasegawa-Johnson, and T. Huang, "Sift-bag kernel for video event analysis," in *Proceedings of the 16th ACM international conference on Multimedia*. ACM, 2008, pp. 229–238.

BIBLIOGRAPHY

[72] D. Xu and S. Chang, "Video event recognition using kernel methods with multilevel temporal alignment," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 30, no. 11, pp. 1985–1997, 2008.

[73] P. Xu, L. Xie, S. Chang, A. Divakaran, A. Vetro, and H. Sun, "Algorithms and system for segmentation and structure analysis in soccer video," in *Proc. IEEE ICME*, 2001, pp. 928–931.

[74] L. Xu and Y. Li, "Video classification using spatial-temporal features and PCA," in *Proc. IEEE ICME*, vol. 3, 2003, pp. 485–488.

[75] S. Nepal, U. Srinivasan, and G. Reynolds, "Automatic detection of'Goal'segments in basketball videos," in *Proc. ACM MM*, 2001, pp. 261–269.

[76] G. Zhu, C. Xu, Q. Huang, Y. Rui, S. Jiang, W. Gao, and H. Yao, "Event tactic analysis based on broadcast sports video," *IEEE Transactions on Multimedia*, vol. 11, no. 1, pp. 49–67, 2009.

[77] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *Proc. IEEE CVPR*, vol. 3613, 2007, pp. 1575–1589.

[78] J. Yang, Y. Jiang, A. Hauptmann, and C. Ngo, "Evaluating bag-of-visual-words representations in scene classification," in *Proc. ACM MIR*, 2007, pp. 197–206.

[79] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories," in *Proc. IEEE CVPR*, vol. 2, 2006, pp. 2169–2178.

[80] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid, "Local features and kernels for classification of texture and object categories: A comprehensive study," *International Journal of Computer Vision*, vol. 73, no. 2, pp. 213–238, 2007.

[81] L. Li, N. Zhang, L. Duan, Q. Huang, J. Du, and L. Guan, "Automatic sports genre categorization and view-type classification over large-scale dataset," in *Proc. ACM MM*, 2009, pp. 653–656.

BIBLIOGRAPHY

[82] S. Fischer, R. Lienhart, and W. Effelsberg, "Automatic recognition of film genres," in *Proc. ACM MM*, vol. 95, 1995, pp. 295–304.

[83] D. Brezeale and D. Cook, "Automatic video classification: A survey of the literature," *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, vol. 38, no. 3, pp. 416–430, 2008.

[84] B. Truong, C. Dorai, and S. Venkatesh, "Automatic genre identification for content-based video categorization," in *Proc. ICPR*, vol. 15, 2000, pp. 230–233.

[85] S. Takagi, S. Hattori, K. Yokoyama, A. Kodate, and H. Tominaga, "Sports video categorizing method using camera motion parameters," in *Proc. IEEE ICME*, vol. 2, 2003, pp. 461–464.

[86] E. Jaser, J. Kittler, and W. Christmas, "Hierarchical decision making scheme for sports video categorisation with temporal post-processing," in *Proc. IEEE CVPR*, vol. 2, 2004, pp. 908–913.

[87] J. Wang, C. Xu, and E. Chng, "Automatic sports video genre classification using pseudo-2d-hmm," in *Proc. ICPR*, 2006, pp. 778–781.

[88] X. Yuan, W. Lai, T. Mei, X. Hua, X. Wu, and S. Li, "Automatic video genre categorization using hierarchical svm," in *Proc. IEEE ICIP*, 2006, pp. 2905–2908.

[89] R. Glasberg, S. Schmiedeke, M. Mocigemba, and T. Sikora, "New Real-Time Approaches for Video-Genre-Classification Using High-Level Descriptors and a Set of Classifiers," in *Proc. IEEE ICSC*, 2008, pp. 120–127.

[90] M. Montagnuolo and A. Messina, "Parallel neural networks for multimodal video genre classification," *Journal of Multimedia Tools and Applications*, vol. 41, no. 1, pp. 125–159, 2009.

[91] A. Ekin, A. M. Teklap, and R. Mehrotra, "Automatic soccer video analysis and summarization," *IEEE Transactions on Image Processing*, vol. 12, no. 7, pp. 796–807, 2003.

BIBLIOGRAPHY

[92] P. Wang, Z. Liu, and S. Yang, "Investigation on unsupervised clustering algorithms for video shot categorization," *Journal of Soft Computing-A Fusion of Foundations, Methodologies and Applications*, vol. 11, no. 4, pp. 355–360, 2007.

[93] L. Zhong, C. Li, H. Li, and Z. Xiong, "Unsupervised Clustering Algorithm for Video Shots Using Spectral Division," in *Proc. ISVC*. Springer, 2008, pp. 782–792.

[94] X. Tong, Q. Liu, H. Lu, and H. Jin, "Shot classification in sports video," in *Proc. ICSP*, vol. 2, 2004, pp. 1364–1367.

[95] J. Wang, E. Chng, and C. Xu, "Soccer replay detection using scene transition structure analysis," in *Proc. IEEE ICASSP*, 2005, pp. 433–437.

[96] M. Kolekar and K. Palaniappan, "Semantic concept mining based on hierarchical event detection for soccer video indexing," *Journal of Multimedia*, vol. 4, no. 5, pp. 298–312, 2009.

[97] R. Benmokhtar, B. Huet, and S. Berrani, "Low-level feature fusion models for soccer scene classification," in *Proc. IEEE ICME*, 2008, pp. 1329–1332.

[98] T. Hofmann, "Learning the similarity of documents: An information-geometric approach to document retrieval and categorization," *NIPS*, vol. 12, pp. 914–920, 2000.

[99] ——, "Probabilistic latent semantic indexing," in *Proc. ACM SIGIR*, 1999, pp. 50–57.

[100] C. Chang and C. Lin, "LIBSVM: a library for support vector machines," 2001.

[101] L. Duan, M. Xu, T. Chua, Q. Tian, and C. Xu, "A mid-level representation framework for semantic sports video analysis," in *Proc. ACM MM*, 2003, pp. 33–44.

[102] G. Lavee, E. Rivlin, and M. Rudzsky, "Understanding video events: A survey of methods for automatic interpretation of semantic occurrences in video," *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, vol. 39, no. 5, pp. 489–504, Sept. 2009.

BIBLIOGRAPHY

[103] D. Sadlier and N. O'Connor, "Event detection in field sports video using audio-visual features and a support vector machine," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 15, no. 10, pp. 1225–1233, 2005.

[104] M. Xu, L. Duan, C. Xu, and Q. Tian, "A fusion scheme of visual and auditory modalities for event detection in sports video," in *Proc. IEEE ICASSP*, vol. 3, 2003, pp. 189–192.

[105] Q. Ye, Q. Huang, W. Gao, and S. Jiang, "Exciting event detection in broadcast soccer video with mid-level description and incremental learning," in *Proc. ACM MM*, 2005, pp. 455–458.

[106] L. Li, Y. Chen, W. Hu, W. Li, and X. Zhang, "Recognition of Semantic Basketball Events Based on Optical Flow Patterns," in *Proc. ISVC*. Springer, 2009, pp. 480–488.

[107] N. Babaguchi, Y. Kawai, and T. Kitahashi, "Event based indexing of broadcasted sports video by intermodal collaboration," *IEEE Transactions on Multimedia*, vol. 4, no. 1, pp. 68–75, 2002.

[108] D. Zhang and S. Chang, "Event detection in baseball video using superimposed caption recognition," in *Proc. ACM MM*, 2002, pp. 315–318.

[109] M. Tien, Y. Wang, C. Chou, K. Hsieh, W. Chu, and J. Wu, "Event detection in tennis matches based on video data mining," in *Proc. IEEE ICME*, 2008, pp. 1477–1480.

[110] Y. Zhang, C. Xu, Y. Rui, J. Wang, and H. Lu, "Semantic event extraction from basketball games using multi-modal analysis," in *Proc. IEEE ICME*, 2007, pp. 2190–2193.

[111] X. Tong, H. Lu, and Q. Liu, "A three-layer event detection framework and its application in soccer video," in *Proc. IEEE ICME*, vol. 3, 2004, pp. 1551–1554.

[112] T. Mei and X. Hua, "Structure and event mining in sports video with efficient mosaic," *Multimedia Tools and Applications*, vol. 40, no. 1, pp. 89–110, 2008.

BIBLIOGRAPHY

[113] T. Wang, J. Li, Q. Diao, W. Hu, Y. Zhang, and C. Dulong, "Semantic event detection using conditional random fields," in *Computer Vision and Pattern Recognition Workshop, 2006. CVPRW'06. Conference on.* IEEE, 2006, pp. 109–115.

[114] C. Xu, Y. Zhang, G. Zhu, Y. Rui, H. Lu, and Q. Huang, "Using webcast text for semantic event detection in broadcast sports video," *IEEE Transactions on Multimedia*, vol. 10, no. 7, pp. 1342–1355, 2008.

[115] G. Miao, G. Zhu, S. Jiang, Q. Huang, C. Xu, and W. Gao, "A Real-Time Score Detection and Recognition Approach for Broadcast Basketball Video," in *Proc. IEEE ICME*, 2007, pp. 1691–1694.

[116] J. Dai, L. Duan, X. Tong, C. Xu, Q. Tian, H. Lu, and J. Jin, "Replay scene classification in soccer video using web broadcast text," in *Proc. IEEE ICME*, 2005, pp. 1098–1101.

[117] C. Xu, J. Wang, K. Wan, Y. Li, and L. Duan, "Live sports event detection based on broadcast video and web-casting text," in *Proc. ACM MM*, 2006, p. 230.

[118] A. Quattoni, S. Wang, L. Morency, M. Collins, T. Darrell, and M. Csail, "Hidden-state conditional random fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 10, pp. 1848–1852, 2007.

[119] S. Wang, A. Quattoni, L. Morency, D. Demirdjian, and T. Darrell, "Hidden conditional random fields for gesture recognition," in *Proc. IEEE CVPR*, 2006, pp. 1521–1527.

[120] A. Gunawardana, M. Mahajan, A. Acero, and J. Platt, "Hidden conditional random fields for phone classification," in *Proc. Interspeech*, 2005, pp. 1117–1120.

[121] Y. Tan, D. Saur, S. Kulkarni, and P. Ramadge, "Rapid estimation of camera motion from compressed video with application to video annotation," *IEEE Transactions on circuits and systems for video technology*, vol. 10, no. 1, pp. 133–146, 2000.

[122] L. Morency, A. Quattoni, C. Christoudias, and S. Wang, "Hidden-state Conditional Random Field Library," 2008.

BIBLIOGRAPHY

[123] F. Sha and F. Pereira, "Shallow parsing with conditional random fields," in *Proc. of HLT-NAACL*, 2003, pp. 213–220.

[124] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proc. ICML*, 2001, pp. 282–289.

[125] Y. Rubner, C. Tomasi, and L. Guibas, "The earth mover's distance as a metric for image retrieval," *International Journal of Computer Vision*, vol. 40, no. 2, pp. 99–121, 2000.

[126] R. Duda, P. Hart, and D. Stork, *Pattern classification*, 2001.

[127] A. Jain, M. Murty, and P. Flynn, "Data clustering: a review," *ACM computing surveys (CSUR)*, vol. 31, no. 3, pp. 264–323, 1999.

[128] J. Smith, "Clicking on Things," *IEEE MultiMedia*, vol. 17, no. 4, pp. 2–3, 2010.

[129] *1st WORKSHOP ON MOBILE VISUAL SEARCH*, Dec 2009, http://scien.stanford.edu/pages/conferences/mvs/.

[130] S. Tsai, D. Chen, G. Takacs, V. Chandrasekhar, J. Singh, and B. Girod, "Location coding for mobile image retrieval," in *Proceedings of the 5th International ICST Mobile Multimedia Communications Conference*, 2009, pp. 1–7.

[131] G. Schroth, R. Huitl, D. Chen, M. Abu-Alqumsan, A. Al-Nuaimi, and E. Steinbach, "Mobile visual location recognition," *Signal Processing Magazine*, vol. 28, no. 4, pp. 77–89, 2011.

[132] L.-Y. Duan and W. Gao, "Side Discriminative Mobile Visual Search," in *2nd WORKSHOP ON MOBILE VISUAL SEARCH*, 2011.

[133] G. Takacs, V. Chandrasekhar, N. Gelfand, Y. Xiong, W. Chen, T. Bismpigiannis, R. Grzeszczuk, K. Pulli, and B. Girod, "Outdoors augmented reality on mobile phone using loxel-based visual feature organization," in *Proc. ACM MIR*, 2008, pp. 427–434.

BIBLIOGRAPHY

[134] D. Chen, S. Tsai, B. Girod, C. Hsu, K. Kim, and J. Singh, "Building book inventories using smartphones," in *Proc. ACM Multimedia*, 2010, pp. 651–654.

[135] S. Tsai, D. Chen, V. Chandrasekhar, G. Takacs, N. Cheung, R. Vedantham, R. Grzeszczuk, and B. Girod, "Mobile product recognition," in *Proc. ACM Multimedia*, 2010, pp. 1587–1590.

[136] J. Polifroni, I. Kiss, and M. Adler, "Bootstrapping named entity extraction for the creation of mobile services," in *Proceedings of LREC*, 2010, pp. 1515–1520.

[137] X. Yin and S. Shah, "Building taxonomy of web search intents for name entity queries," in *Proceedings of WWW*, 2010, pp. 1001–1010.

[138] *Merriam-Webster Dictionary.* Merriam-Webster, 2002.

[139] A. Broder, "A taxonomy of web search," in *ACM SIGIR.* ACM, 2002, pp. 3–10.

[140] D. Rose and D. Levinson, "Understanding user goals in web search," in *Proc. of the ACM WWW*, 2004, pp. 13–19.

[141] K. Church and B. Smyth, "Understanding the intent behind mobile information needs," in *Proc. of ACM International Conference on Intelligent User Interfaces*, 2009, pp. 247–256.

[142] J. Zhuang, T. Mei, S. C. H. Choi, Y.-Q. Xu, and S. Li, "When recommendation meets mobile: Contextual and personalized recommendation on the go," in *Proc. of ACM International Conference on Ubiquitous Computing*, Beijing, China, Sept. 2011, pp. 153–162.

[143] K. Church, B. Smyth, and N. Oliver, "Visual interfaces for improved mobile search," in *Workshop on Visual Interfaces to the Social and the Semantic Web*, 2009.

[144] A. Dix, J. Finlay, G. Abowd, and R. Beale, *Human Computer Interaction (Third Edition).* Prentice Hall, 2004.

[145] R. Jain and P. Sinha, "Content Without Context is Meaningless," in *Proc. ACM Multimedia*, 2010, pp. 1259–1268.

BIBLIOGRAPHY

[146] G. Hua and Q. Tian, "What can visual content analysis do for text based image search?" in *Proc. IEEE ICME*, 2009, pp. 1480–1483.

[147] I. Guy, A. Jaimes, P. Agulló, P. Moore, P. Nandy, C. Nastar, and H. Schinzel, "Will recommenders kill search?: recommender systems-an industry perspective," in *Proc. of ACM conference on Recommender systems*, 2010, pp. 7–12.

[148] S. Robertson and H. Zaragoza, "The probabilistic relevance framework: BM25 and beyond," *Information Retrieval*, vol. 3, no. 4, pp. 333–389, 2009.

[149] L. Yang, B. Geng, A. Hanjalic, and X. Hua, "Contextual image retrieval model," in *Proceedings of the ACM International Conference on Image and Video Retrieval*. ACM, 2010, pp. 406–413.

[150] Z. Zhang, X. Liang, A. Ganesh, and Y. Ma, "TILT: transform invariant low-rank textures," *Proc. of ACCV*, pp. 314–328, 2010.