Content based Video Retrieval via Spatial-Temporal Information Discovery

Wang, Lei

A thesis submitted in partial fulfilment
of the requirements of
The Robert Gordon University
for the degree of Doctor of Philosophy

December 2013

Abstract

Content based video retrieval (CBVR) has been strongly motivated by a variety of real-world applications. Most state-of-the-art CBVR systems are built based on Bag-of-visual-Words (BovW) framework for visual resources representation and access. The framework, however, ignores spatial and temporal information contained in videos, which plays a fundamental role in unveiling semantic meanings. The information includes not only the spatial layout of visual content on a still frame (image), but also temporal changes across the sequential frames. Specially, spatially and temporally co-occurring visual words, which are extracted under the BovW framework, often tend to collaboratively represent objects, scenes, or events in the videos. The spatial and temporal information discovery would be useful to advance the CBVR technology.

In this thesis, we propose to explore and analyse the spatial and temporal information from a new perspective: i) co-occurrence of the visual words is formulated as a correlation matrix, ii) spatial proximity and temporal coherence are analytically and empirically studied to refine this correlation. Following this, a quantitative spatial and temporal correlation (STC) model is defined. The STC discovered from either the query example (denoted by QC) or the data collection (denoted by DC) are assumed to determine specificity of the visual words in the retrieval model, *i.e.* selected Words-Of-Interest are found more important for certain topics. Based on this hypothesis, we utilized the STC matrix to establish a novel visual content similarity measurement method and a query reformulation scheme for the retrieval model. Additionally, the STC also characterizes the context of the visual words, and accordingly a STC-Based context similarity measurement is proposed to detect the synonymous visual words. The method partially solves an inherent error of visual vocabulary under the BovW framework.

Systematic experimental evaluations on public TRECVID and CC_WEB_VIDEO video collections demonstrate that the proposed methods based on the STC can substantially improve retrieval effectiveness of the BovW framework. The retrieval model based on STC outperforms state-of-the-art CBVR methods on the data collections without storage and computational expense. Furthermore, the rebuilt visual vocabulary in this thesis is more compact and effective. Above methods can be incorporated together for effective and efficient CBVR system implementation. Based on the experimental results, it is concluded that the spatial-temporal correlation effectively approximates the semantical correlation. This discovered correlation approximation can be utilized for both visual content representation and similarity measurement, which are key issues for CBVR technology development.

Keywords: Video Retrieval, Bag-of-visual-Words, Spatial and Temporal Information

Acknowledgements

I am greatly indebted to my supervisors, Prof. Dawei Song and Dr. Eyad Elyan, for their valuable guidance, fruitful discussions and tremendous support throughout my PhD. Special thanks to Dr. Jun Wang, Dr. Peng Zhang, and Dr. Leszek Kaliciak for their collaborations or discussions with my research.

Many thanks to Robert Gordon University for granting me full studentship and to Open University for supporting my visiting at Milton Keynes. Gratitude also goes my colleagues, Ibrahim, Jean-Claude, Sandy, Thiery, David, Malcolm, Guofu, Peter, Sadiq, Amina, Richard, Noura, Micheal, Claire, Ben, Nuka, Olivier, and Yanghui, for making CTC a convenient research environment.

I appreciate especially my wife Shuo Li for her love, encouragement, patience and understanding. Finally, I am also indebted to my parents Mr Xijun Wang and Ms Yuanjing Lei for their love and unconditional support.

Declaration

I declare that all of the work in this thesis was conducted by the author except where otherwise indicated, and this thesis has not been submitted at any other university.

Parts of the work outlined in this thesis have appeared in the following publications.

Chapter 3 & Chapter 4

Lei Wang, Dawei Song, Eyad Elyan, Improving Bag-of-visual-Words model with spatial-temporal correlation for video retrieval. Accepted by *The 21st ACM Conference on Information and Knowledge Management (***CIKM 2012***)*, Hawaii, USA, Pages 1303-1312

Lei Wang, Dawei Song, Eyad Elyan, Words-of-Interest Selection based on Temporal Motion Coherence for Video Retrieval. Accepted by *The 34th Annual ACM SIG Information Retreival Conference (***SIGIR 2011***)*, Beijing, China, Pages 1197-1198

Lei Wang, Dawei Song, Eyad Elyan, Video Retrieval based on Words-of-Interest Selection. Accepted by *The 33rd European Conference on Information Retrieval (***ECIR 2011***)*, Dublin, Ireland, Pages 687-690

Chapter 5 Lei Wang, Eyad Elyan, Dawei Song, Rebuilding Visual Vocabulary via Spatial-Temporal Context Similarity for Video Retrieval. Accepted by *The 20th Anniversary International Conference on MultiMedia Modeling (***MMM 2014***)*, Dublin, Ireland, pp 74-85

Contents

List of Algorithms

List of Figures

List of Tables

Chapter 1

Introduction

This chapter introduces fundamental concepts of content based video retrieval (CBVR) technology, challenges and opportunities encountered in developing the CBVR technology, as well as the research objectives, contributions, and outline of this thesis.

1.1 Content based Video Retrieval

Digital video technology has developed quickly in recent decades and cheap recording instruments, such as cameras and smart phones, have become very popular around the world. A vast number of video clips are continuously produced by not only professional video program broadcasters like the BBC, but also amateur users, and even unmanned recorders like various visual sensors. As a result, there are always billions of hours of videos maintained by broadcasters or commercial video websites. For example, 72 hours of video are uploaded to the commercial video website YouTube every minute, and more than 200 millions of video have been shared on the website (Youtube 2013).

Thus, there is urgent need for advanced information retrieval technology which helps users to access desired videos more efficiently and effectively. The available information contained by a video normally includes (Smeaton 2006) : 1) video metadata, which are textual information such as titles, summary, authors, copyrights and format information; 2) audio information, which is provided by the auditory channel in form of music, background sounds, and speeches; 3) transcripts, some of which are already packaged within the video, and they can also be discovered by optical character recognition technology; 4) visual information contained in temporally sequential images provided by visual channel. Besides, the online videos are always archived within web pages, which also provide some extra textual information which may be related to the videos.

Current video resources organizing and archiving technology widely utilized by commercial websites largely relies on the tagged textual information embedded with the videos. This scheme has an advantage that most existing text information access and manage-

ment technology can be directly utilized to organize the video resources. The text directly covers the semantics and thus the computational perceptual problem is avoided. However, the textual descriptions are not always sufficient, because a dynamic video has very rich content (far more than still images and textual articles) and it is very challenging for video producers to manually generate adequate textual descriptions.

A scheme to overcome the deficiency of textual clues in video data is to leverage speech recognition or optical characters recognition technology (Bakker & Lew 2002) to discover textual description from the audio and transcripts information. Although such technologies have made a significant progress during the previous decade (Hauptmann 2002), many videos (*e.g.* music video, silent movie, amateur videos) lack prior transcripts. Moreover, in other cases, the transcripts can not effectively cover the huge visual content contained in the raw visual data. As a consequence, this scheme can not completely satisfy the requirements of certain real-world applications.

A broad range of applications have motivated novel technologies to directly analyze and understand the visual content (Hu, Xie, Li, Zeng & Maybank 2011) , for example, copyright infringement detection, landmark/object recognition, digital library, harmful video tracing, abnormal event detection, automatic remote surveillance, and so on. Related research topics include content base image/video retrieval, video semantic indexing, hot event detection, and image/video annotation.

These topics have attracted the interests of researchers from all over the world. For example, The National Institute of Standards (NIST) and Technology has organized and sponsored an annual workshop, namely Text Retrieval Conference Video Retrieval Evaluation (TRECVID) since 2001. It provides public and large scale video resources and evaluation tools. A majority of studies in the video retrieval research community , e.g., (Smeaton, Over & Kraaij 2006), (Smeaton, Over & Kraaij 2009), have applied their systems with the TRECVID data and tasks.

The importance and popularity of the visual-content based video access technology have led to an inter-discipline scientific field: content based video retrieval (CBVR), which is a hybrid between Information Retrieval (IR) and Computer Vision. A number of concepts, definitions, theoretical frameworks, and powerful technologies are inherited from these two traditional computer science fields. In the next section, key definitions and terminologies utilized in this thesis are introduced and explained in details.

1.2 Background and Terminologies

The CBVR is a branch of visual information retrieval, which aims to help users automatically obtain visual information which is **relevant** to the **information need** of users from large scale visual **data collection**.

At first, the information need is often described by a **query** given by the users. The

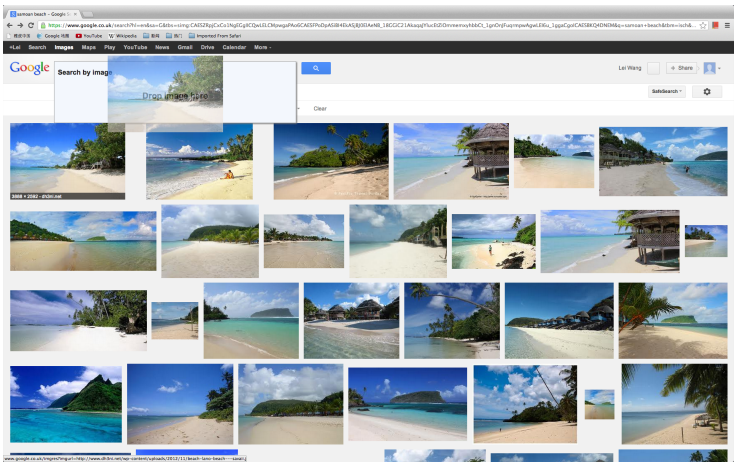


Figure 1.1: Google: Image search by given example

modalities of query vary from textual words to visual example. For example, as shown in Figure 1.1, the well known search engine Google released its new image search service in 2011: users can upload or select an image and search for related images. This case is always referred to as **Query-by-Example** (QBE) in the literatures, which is an important application of CBVR.

Within the information retrieval research community, objectives of retrieval systems are defined as **relevances** to the query. Here, relevance is a subjective criteria. Theoretically, only real world users are able to judge the relevance, and the system can only predict the degree of relevance. The prediction computation is often defined as relevance function. There is no fixed connotation of this concept, because different users may have different preferences. For example, some users may look for visual content with similar color whilst others will search for a specific object. The various purposes of real world information services also ask for different relevant criteria. Figure 1.2 and Figure 1.3 presented a few examples for near duplicate video searching and generic topic searching respectively, and videos on each row are relevant to an identical topic. The relevances of the former application share a large degree of visual similarity, whilst the latter application shows more concern about semantics.

The large amount of raw visual data should be abstracted as a visual content **representation** that enables the system to index and retrieve videos more efficiently. Moreover, the raw data is also of very low level semantic meaning, and the representation extraction could encode higher level semantics into metadata, which would make the retrieval system more effective. Accompanied by progress in computer vision, the visual content of the video is always represented by a bunch of visual features. A **visual feature** is defined as an attribute or aspect of visual information (Smeaton et al. 2009). These features may be global features like color (Carson, Belongie, Greenspan & Malik 2002a), (Hu 2005), shape, texture (Carson, Thomas, Belongie, Hellerstein & Malik 1999) of the image/video,

Figure 1.2: Relevances of Near Duplicate Video Searching

or local features, which are concrete patterns around a few salient points in the video. More information regarding features can be found in the next chapter.

Based on the invented local features, a novel video content representation framework is proposed, namely the **Bag-of-visual-Words (BovW)** model. As firstly introduced in the context of visual object search (Sivic & Zisserman 2006), high dimensional feature descriptors, such as SIFT (Lowe 2004) and SURF (Bay, Ess, Tuytelaars & Van Gool 2008), are extracted to represent the stable and salient regions surrounding the points-of-interest detected by certain local feature detectors. The regional descriptors generated from the collection or a training dataset are clustered and each cluster forms a **visual word**. Given an image (or a keyframe in a video), the region descriptors in the image are then quantized into discrete visual words. Specifically, the quantization function maps a region descriptor onto its closest cluster centroid. The region descriptor is then called an instance of the corresponding visual word (in this thesis, we use the term visual word and "instance of visual word" interchangeably, for convenience, unless explicitly distinguished). As a result, an image can be represented as a bag of visual words. A pair of descriptors mapped onto an identical visual word are considered as a match between their visual contents.

This representation framework has become very powerful for CBVR. It represents the visual content as a number of basic elements: visual words, which play similar role to words in the textual document representation. This characteristic enables a CBVR system to utilize a series of methods, whose retrieval effectiveness has been proven in the textual information retrieval field: Term Frequency and Inverse Document Frequency(TF-IDF) (Sivic & Zisserman 2006), inverted document indexing structure, query expansion(Chum, Mikulik, Perdoch & Matas 2011), vector space model (Zhao, Wu & Ngo 2010), feedback

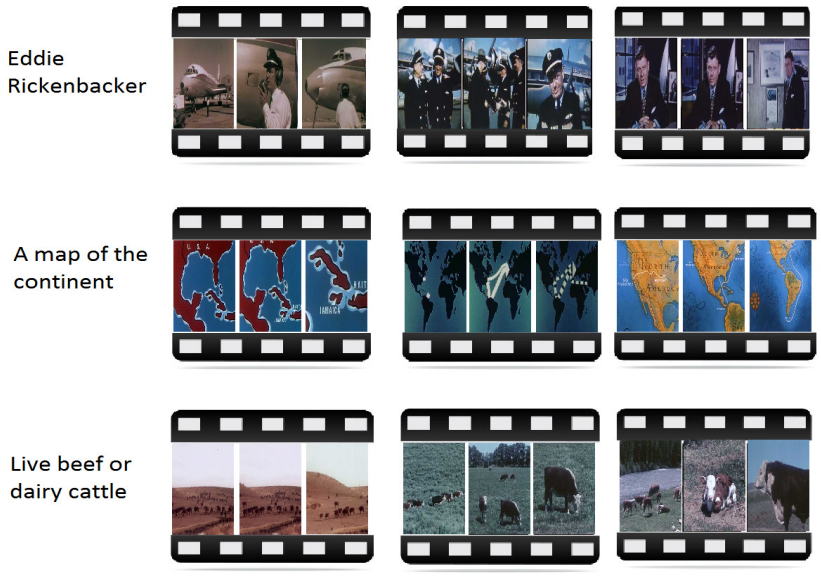


Figure 1.3: Relevances of Generic topics Searching

(Chum, Philbin, Sivic, Isard & Zisserman 2007), etc. Given its simple structure and descriptive ability, the BovW model has succeeded in many related fields such as: object recognition (Zhang, Tian, Hua, Huang & Li 2009) and categorization (Zhang, Liu, Ouyang, Lu & Ma 2009), event detection (Ke, Sukthankar & Hebert 2005), and video event classification (Ballan, Bertini, Bimbo & Serra 2009).

However, the BovW framework has a few inherent drawbacks, and development of CBVR also meets a number of theoretical and practical challenges as outlined in the next section.

1.3 Challenges and Opportunities

The BovW framework has enabled the CBVR system to efficiently search for resources via a similarity measurement based on the visual representations of videos. It has been reported in the previous research that visual words, which are based on low level visual features, are not always as effective as textual words for IR application (Zhang, Jia & Chen 2011) (Zhang, Tian, Hua, Huang & Li 2009). The major obstacle is contributed by the so called "semantic gap" between visual features' similarity and the true relevance of videos (Datta, Joshi, Li & Wang 2008). The gap is partially caused by a few drawbacks of the BovW model, which are summarized as *Ignorance of Spatial-Temporal Information*:

Amongst the fundamental problems of the BovW model is that it always neglects geometrical information associated with the visual words. For example, the representation assumes that each visual word is independent and ignores its spatial position and timing information. Recently, the spatial and temporal information has been demonstrated

to be very important for the visual content representation (Galleguillos, Rabinovich & Belongie 2008) (Wang, Jiang & Ngo 2008). Neglecting spatial-temporal information also leads to losing the context of visual words. The context is not only a clue regarding the inherent connection between the visual words (Zhou, Tian, Yang & Li 2010) , but may also determine their semantics (Su & Jurie 2011) . Without such information, the descriptive ability of the visual words would definitely be harmed (Chum et al. 2011). A model with which to quantitatively discover the spatial-temporal information for video representation is still an open problem for researchers. It is always expensive, in terms of computational and storage to directly embed such information into the video representation and retrieval model. How to effectively and efficiently utilize the spatial-temporal information for CBVR is one of the most important challenging questions.

Different with short queries of a few words normally used in textual information retrieval, the CBVR system can easily enable user query by visual example. As shown by many previous advanced visual information retrieval system (Flickner, Sawhney, Niblack, Ashley, Huang, Dom, Gorkani, Hafner, Lee, Petkovic, Steele & Yanker 1995), (Smeulders, Worring, Santini, Gupta & Jain 2000), (Smeaton, Wilkins, Worring, de Rooij, Chua & Luan 2008), visual example (image/video) is convenient for the users to describe the clue of their desired information. On one hand, users can directly express their information interests precisely without composing good textual query for retrieval. On the other hand, rich information is contained in the query example, which can be considered as clues of relevances. The Query-by-Example technology provides another important opportunity for the system to capture the information need and discover relevant information.

However, at the same time, this opportunity is also a great challenge that rich information is mixed together with other messy information in the query example and videos. Users' desired information and irrelevant information cannot be distinguished without visual content relationship modelling, where spatial-temporal information can play important role. More specifically, the ignorance of spatial-temporal information leads to the following two problems for the retrieval models.

- *Ineffective Visual Content Representation and Relevance Prediction.* The lack of spatial-temporal information would increase the like hood of irrelevant results being retrieved and relevant results being missed by the CBVR system. A single visual words always contains little information, but visual concepts are always composed of multiple visual words. As a consequence, without the spatial-temporal constraints, the system cannot effectively represent the visual contents and accurately measure the similarity.

- *Quantization Errors of the Visual Words.* The quantization process without spatial-temporal context may result in less reliable visual vocabulary and less distinctive visual words. For example, some visual word represents different meanings in various

spatial-temporal contexts (Nister & Stewnius 2006), which are like polysemy words in text. Some others visual words may always appear in the same or similar context and represent an identical visual concept (Zhang, Huang, Hua, Jiang, Gao & Tian 2010), which are like synonymous words. Compared to relatively more concrete textual words, the spatial-temporal context information is even more important for the visual words (Fernando, Fromont, Muselet & Sebban 2012) in retrieval tasks. How to address these errors incurred by the BovW framework is another important challenge faced by researchers.

It should be pointed out that the discovery of relationship between words is also an attractive topic in textual retrieval. For example, Topic Model assembles frequently co-appearing words as topic to describe the content of documents. A number of feature selection or reduction technology are invented to emphasize descriptive words. These methods can be applied under BovW framework for CBVR. But, the spatial-temporal information contained in the video opens different door to develop video retrieval, and this thesis aims to address the above-mentioned challenges by discovering the neglected spatial-temporal information, and efficiently using such information to improve the effectiveness of the BovW framework for CBVR tasks. The research objectives will be introduced in the next section.

1.4 Research Objectives

This work aims to address the above challenges and develop novel CBVR technology via spatial-temporal information discovery. The main research objectives are as follows:

1. To explore existing literatures in the area of Spatial-Temporal Information and build a quantitative spatial-temporal information discovery framework;

2. To establish novel relevance prediction and visual content model, which incorporates the spatial-temporal information discovered;

3. To build more descriptive visual words based on the spatial-temporal information for CBVR;

1.5 Contributions

- **Spatial-Temporal Correlation Model** The use of spatial-temporal information to preserving semantics in video representation and processing has been extensively studied. For example, plenty of works have investigated the embedding of geometric information into the visual words(Cao & Li 2007).

Intuitively, co-occurring visual words often tend to collaboratively interpret the layout of visual concept, *i.e.* visual object, person, event, and etc. The co-occurring relationship between these visual words implies that they may be semantically related to each other. This clue has been utilized in recent work to improve the visual content representation in related fields (Galleguillos et al. 2008).

The spatial proximity and temporal motion between co-occurring visual words characterize higher level correlations between them. We assume that it is an approximation of semantic correlation among corresponding visual words. For example, closer visual words are more likely to be related to each other (Chen, Chen & Chien 2008). Based on the paradigm of the textual IR research community, we proposed to define a concept: Spatial-Temporal Correlation (STC) between the visual words. This work has proposed to build a novel and uniform framework to model the co-occurring visual words and their STC.

- **Representation Reformulation Model based on the STC** The BovW framework normally assumes that all the visual words are priorly equal and independent. It distinguishes the visual words with weights defined by term specificity, *e.g.*, Term Frequency. Rich visual informations are always mixed in video, and the traditional term weighting such as TF-IDF is found to be not effective enough to address the important visual content (Chum et al. 2011).

  It is found that the importance of different visual words in the representation are different for CBVR, with respect to the searcher's interest points elected in the query, *i.e.* selected Words-of-Interests are more Important. We propose to re-weight the visual word according to its correlation with others in the video. A stronger correlation is seen as indicating a closer association with the visual topic of the video, which is equivalent to an assumption that noise tends to be singular and unstable. While the STC can be discovered by the proposed method, it is utilized to emphasize the descriptive visual words for visual content representation. This novel term weighting scheme is used to reformulate the query in this thesis, and it is defined as Query Correlation (QC) weights.

  Another idea is that the term weights should also be distinguished by STC discovered from entire video data collection. Inspired by the inverse document frequency, we assume that visual words with higher STC in the data collection are less informative. We defined this term weighting scheme as Inverse Documents Correlation (IDC).

  The present thesis has utilized these term weights to reformulate the query representation, which is equivalent to a novel similarity measurement function for the retrieval model. We have conducted a number of practical experiments to evaluate its effectiveness.

- **Rebuilding the visual vocabulary via the Spatial-Temporal Context** A more compact visual vocabulary would definitely result in a more descriptive visual content representation. We propose to utilize spatial-temporal context, which is characterized by the STC, to reduce redundancy in the visual vocabulary of the BovW model.

  We assume that the semantics of visual words can also be differentiated by their context. Based on this assumption, synonyms are detected based on the context similarity measurement. Afterwards, the detected synonyms are merged to form a more compact visual vocabulary. In this way, the new representation of videos can be more descriptive based on the rebuilt visual vocabulary.

  This thesis also conducts series of experiments to evaluated its effectiveness in CBVR applications.

The above proposed approaches, this work aims to address the challenges and realize the research objectives. The methodology for implementation and evaluation of the approaches in various typical CBVE tasks are summarized in the next section.

1.6 Methodology

To boost the research activities, Information Retrieval research community always produces public data collection (Everingham, Van Gool, Williams, Winn & Zisserman n.d.), which includes a number of documents and a series of standard topics or given queries. The advantages of this methodology is that the researchers can compare their approaches more easily and accurately.

The common evaluation criteria utilized by Information Retrieval research community include precision, recall and mean average precision, which is also used in this thesis to compare our methods with state-of-the-art.

Precision and **Recall** are widely used criteria. Precision presents the ratio of positive results to retrieved documents:

$$\textbf{Precision} = \frac{\textbf{Relevant Retrieved Result}}{\textbf{Number of All Retrieved Results}} \tag{1.1}$$

Recall defines the ratio of positive results to number of all relevant documents in the ground-truth:

$$\textbf{Recall} = \frac{\textbf{Relevant Retrieved Result}}{\textbf{Total Relevant Documents}} \tag{1.2}$$

The precision and recall are both used to evaluate top-N retrieved results. To summarize the performance, **Mean Average Precision** is the mean value of average precisions

Table 1.1: Video Search of TrecVID

Year	Search task	Query Availability	Evaluation	Resources
2001	Video Search	Only Textual Topics	Precision and Recall	10 hours
2002	Video Search	Topics and Images	MAP	68 hours
2003	Video Search	Topics and Images	MAP	120 hours
2004	Video Search	Topics and Images	MAP	70 hours
2005	Video Search	Topics and Images	MAP	80 hours
2006	Video Search	Topics and Images	MAP	82 hours
2007	Video Search	Topics and Images	MAP	50 hours
2008	Video Search	Topics and Images	MAP	100 hours
2009	Video Search	Topics and Images	MAP	80 hours
2010	Instance Search	Instances' Images	Location and Accuracy	200 hours
2011	Instance Search	Instances' Images	Location and Accuracy	200 hours
2012	Instance Search	Instances' Images	Location and Accuracy	200 hours

for all topics, where the average precision of each topic is defined as the average value of precision associated with relevant documents in retrieved rank:

$$\textbf{Average Precision} = \frac{\sum \textbf{Precision}(k)}{\textbf{Total Relevant Documents}} \tag{1.3}$$

where $\textbf{Precision}(k)$ denotes the precision of documents ranked not lower than k^{th} relevant document.

To validate the research proposals of this work, a number of retrieval function modules are designed and implemented to construct an experimental prototype CBVR system. The implementation is programmed with C++ and supported by public computer vision toolsets OpenCV. The OpenCV toolset provides a bunch of standard image processing algorithm and their interfaces. Based on this toolset, videos codec and uncompressing, images preprocessing, and a number of feature detection and extraction methods can be implemented and compared, which is a fundament of CBVR system built for this thesis.

The system is able to complete common QBE-CBVR applications, and this work takes two typical applications as examples: near-duplicate video detection and generic topic video retrieval. To evaluate the effectiveness of our proposals, we tested the proposed methods with two publicly available data collection resources to make our evaluation comparable with other researchers: CC_WEB_VIDEO and TRECVID2002 video dataset. Each data collection includes a list of topics and relevant videos associated with the topics, which consists ground-truth.

TrecVID is hosted by NIST, and it organizes a series of activities related to video understanding, archiving and retrieving technologies development. Specifically, it annually releases retrieval tasks and corresponding video collection and calls for participation of world-wide researchers and their CBVR system. The searching task and video data released by TrecVID are summarized in Table 1.1. Before 2009, the search task normally

includes a set of unannotated videos and a number of topics. The attendants need to compose their own queries manually or semi-automatically, and to use the queries to search for relevant videos of each topic.

After 2010, the video search task has been replaced by instance search, which provides common query example with modal of images for targeting visual objects (Smeaton et al. 2009). This thesis aiming to investigate on development of query by video example technology for CBVR. Because of this purpose, we do not choose these data collections.

In this thesis, the video search task in TRECVID2002 is utilized as evaluation task in this thesis. We must follow its rule to compose our own query, because query in modality of video example is not available to us. To make our results comparable and repeatable by other researchers, we do not use external visual resources to compose queries but randomly select a relevant video from ground truth as a query for every run of video search. As discussed in (Donald & Smeaton 2005), MAP of the video search on this collection can achieve 6.9%, but they fuse both visual information and textual information to achieve this result.

CC_WEB_VIDEO is a near-duplicate video detection data set, which is one of the typical applications of CBVR technology. For example, the search engine engineers may want to reduce the repeated results and diversify the top-ranked documents; the copyright infringement detection may need to find out the visual resources, which contain specific content. The relevant videos are normally very similar to each other, and they only slightly differ from on some small points: the watermark, logo, background, jointed with other video clip, and etc.

As described above, using the two types of public video retrieval tasks, effectiveness of CBVR methods in this thesis can be completely evaluated based on the common criteria used in IR research community.

1.7 Thesis Organization

The remainder of the thesis is organized as follows:

In Chapter 2, we review a variety of visual features that have been developed for visual content understanding, representation and access technology in Section 2.1. Following this, in Section 2.2, we presented the background of the Bag-of-visual-Word framework, including its visual content representation and similarity measurement. We also reviewed recent works on improving the BovW model in this section. In Section 2.3, we then move to recent works to address spatial and temporal information in visual content. The state-of-the-art approaches to utilizing this spatial-temporal information for visual information retrieval technology are also discussed in this section.

Chapter 3 describes the proposed Spatial-Temporal Correlation (STC) quantization model. In Section 3.1, we presented a theoretical analysis and a quantization framework

of the co-occurring relationship across different visual words. In Section 3.2, the formulated spatial correlation discovery function based on proximity is introduced. In Section 3.3, we define the temporal correlation preserved in motion coherences among visual words. A series of discussions about the STC model are presented in Section 3.4.

In Chapter 4, we investigate approaches to improving retrieval performance of the BovW framework based on the STC. We firstly presented a method to address descriptive visual words in Section 4.1. Section 4.2 reveals a query reformulation method to emphasize the descriptive visual words according to the STC in the query example (QC). Following this in Section 4.3, we theoretically analyse the impact of the STC discovered from data collection (IDC). A series of practical experiments are constructed to evaluate the STC based retrieval models, and the experimental results are shown in Section 4.4.1.

Chapter 5 presents a novel visual vocabulary rebuilding scheme. Firstly, causes and results of quantization errors are analyzed theoretically in Section 5.1. We then define the context of visual words based on the STC in Section 5.2. In Section 5.3, we propose to measure the context similarity and detect synonyms in visual vocabulary. Section 5.4 illustrates performance of the constructed visual vocabulary in a group of practical experimental results.

Chapter 6 presents the implementation of a prototype system and the evaluation of our methods in two typical CBVR applications. The architecture of the experimental prototype system is introduced in Section 6.1. Sections 6.2 and 6.3 describe additional experimental results when applying it to typical CBVR tasks including Near-duplicate video detection and generic topic video retrieval respectively. Section 6.4 puts forth discussions about the combination of our approaches.

Finally, in Chapter 7, we conclude this thesis by summarizing our contributions in Section 7.1. We also addressed the limitations of our work and propose future directions in Section 7.2.

Chapter 2

Literature Review

In this chapter, we provide an introduction to and a critical review of the existing theoretical achievement and important invented technologies, which is a background of our research. We focus the visual content analysis and understanding, and its utilization in visual information retrieval. Specifically, we will discuss the opportunities involved by spatial-temporal information discovery and state-of-the-art progress. The scope of current progress in parsing, represention and retrieval of visual content is defined under the following criteria:

Data Scope This thesis aims to develop the technology searching videos against data collection from diverse domains, for example the web video dataset. The query and video data are processed without the domain information. The use of domain knowledge in visual content representation and the retrieval research will not be referred to in this review.

Query Modality This project considers the queries with modality of visual example as most important opportunities for our technology development. In order to concentrate on visual content representation and similarity measurement problems, the utilization of text information is not in the scope of future research.

Core Technique Retrieval will be formalized into two scientific theories: i) A theory to mathematically characterize the visual content. Because raw video data is normally considered to be not concrete enough, CBVR systems are always based on various visual features extraction theory. ii) A theory to assess the similarity between a pair of videos. The common objective of this theory is to predict relevances between visual contents based on the extracted visual features. The two theories are strongly correlated, but for convenience, this thesis reviews the progresses concerning these two topics respectively.

Bag-of-visual-Words Framework Most state-of-the-art retrieval methods are proposed based on the BovW framework. We review the background, structure, and a selection of algorithms of the BovW framework, which is concrete but effective for CBVR.

Spatial-Temporal Information The ignored spatial and temporal information in BovW framework is found to be bottleneck of CBVR technology development. Therefore, a series of attempts to model and utilize spatial-temporal information within the visual resources. The literatures related to these topics is reviewed and discussed in Section 2.3.

2.1 Visual Content and Similarity

Raw video data, for example a streaming of colored pixels, is always not concrete enough for video content understanding and visual information retrieval. Although some very early works directly used simple gray values of pixels for object recognition, most modern visual information analysis technology utilize some method to abstract and combine of the pixels values (Datta et al. 2008). These methods are always called visual feature extraction, which is one of most important backgrounds of CVBR development.

Fundamentally, a video is a series of temporally aligned images: $V = f_1, f_2, f_3, ..., f_M$, where each image f is called a frame of the video. The mathematical characterization of the visual content in a still frame is always identical to the image representation problem in the image processing and understanding research.

At first, we explore and review visual features proposed for image and video representation. In the literature, the video content representation methods are based on the extraction of visual features from images at different levels:

- Low level global feature: color, shape, texture, and etc.;
- Mid-level feature: regional signature, local feature, and etc.;
- Visual concept: visual entity and event.

In addition, we introduce research progress on video shot boundary detection, which divides a sophisticated video program into a sequence of simple and concrete video shots.

Finally, we review image similarity measurement approaches based on different categories of representation formulation, and approaches for measuring the video similarity based on the key frame similarities.

2.1.1 Global Visual Features

In the early years, extensive research efforts have been made to investigate on visual content representation based on the low level global visual feature (Zhang & Petkovic 1996). The global feature is so named because it describes a global property of an image without considering components in the image. For example, color composition of the image is a typical global feature. As shown in the Figure 2.1, the color composition is always modelled as histograms of intensity values associated with all pixels appearing in

the image. Colored pixels in an image can be identified by $I = \{x, y, c_i\}$, where x, y denote coordinates of the pixel on the image. The image representation based on the color feature is normally formulated as follows:

$$C_I(c_i) = \sum_{x,y} \{I(x, y) = c_i\} \tag{2.1}$$

where c_i is i^{th} discrete color value defined. It means that the i^{th} bin of the color histogram shown in Figure fig:color is counting the number of corresponding pixels appearing in the image that fall into it.

Figure 2.1: The color histogram of image

The color composition extraction idea is straightforward, and it has been widely used in many visual content retrieval systems (Pentland, Picard & Sclaroff 1995), (Carson et al. 1999), (Ponceleon, Srinivasan, Amir, Petkovic & Diklic 1998). The color information could be combined with additional information like relationships between the pixels to form new features. For example, visual feature namely color correlograms (Huang, Kumar, Mitra, Zhu & Zabih 1997) incorporate spatial distance between the color pixels into the feature. If there are N types of color appear in the image, the correlogram histogram may include $N \times N \times D$ bins as follows:

$$Correl_I(c_i, c_j, d) = \{\|(I(x, y) = c_i), (I(x, y) = c_j)\| = d\} \tag{2.2}$$

where $\|*, *\|$ denotes the physical distance between the two pixels, which are in color bins c_i and c_j respectively. In this way, the color correlogram describes the visual content with a combination of physical distance and color information.

Another kind of visual perception of the image is texture information. The feature extraction technologies utilized to describe this information include: edge detection (Tu & Zhu 2002) (Jain & Vailaya 1996), Hough transformation (Ballard 1981) (Illingworth & Kittler 1988), color co-occurrence matrix (Palm 2004), and etc. An example is shown in the Figure 2.2, the edge detection is a typical method aiming to characterize the visual shape of objects.

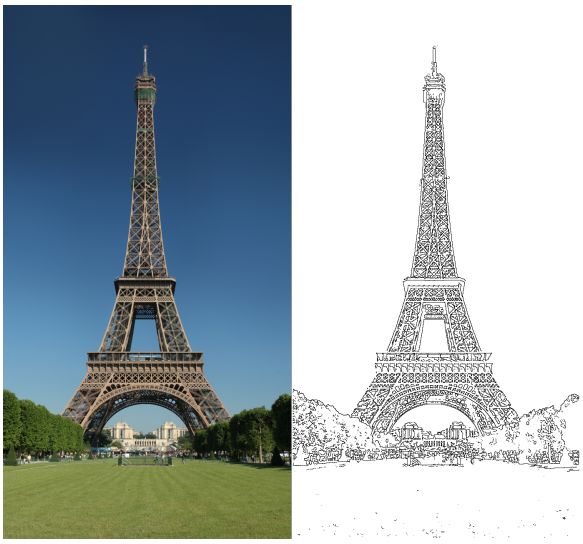


Figure 2.2: The edge detection of image

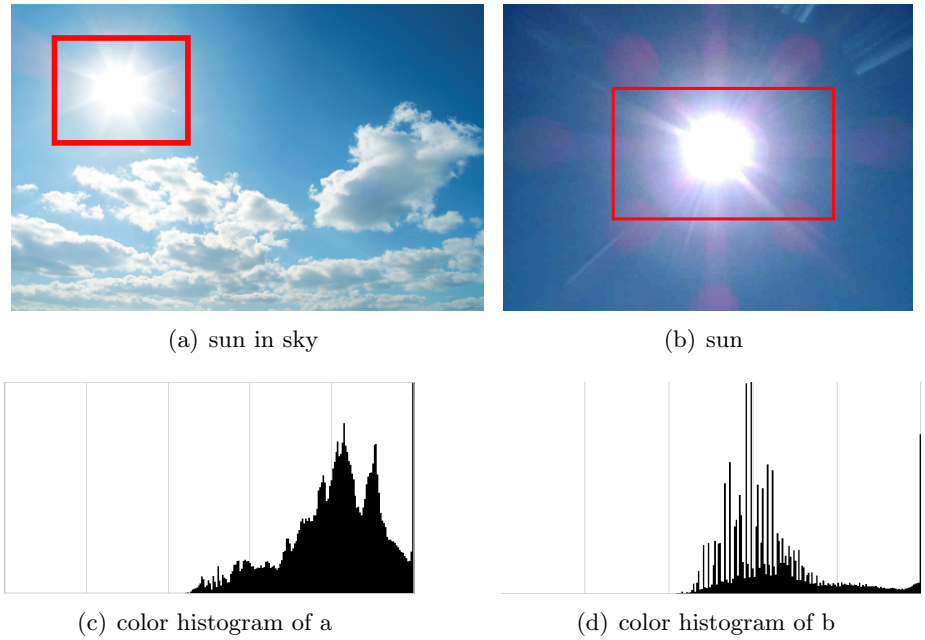
The edge detection normally computes spatial derivation of the pixels' intensities. Most classical detection technologies are often grouped according to their derivation expression as: first-order detector and second-order detector. The first-order detector utilizes the first order deviation and captures strength of the intensity change over the spatial space. The second-order detector utilizes the second order derivation and captures the change rate of the intensity gradient. Recently, advance edge detectors are always designed by incorporating scale space model (Lindeberg 1996). The edge features can also be archived in the form of a histogram (Park, Jeon & Won 2000) for the visual content representation.

The global feature is not always sufficient to describe a particular object or entities in an image to achieve a more accurate retrieval process. Figure 2.3 illustrates one of the inherent problems for the global feature representation. The similarity based on the color feature does not match well with the visual similarity, because sun is on a different scale in the two images and extra information in the two images interferes the match of sun. Other problems of global feature representation include cluttering, change of view point, and geometric distortion. To address these limitations, region-based feature/local feature has been investigated in recent years. It has been shown that a representation model based on the local features has an advantage in object matching and access (Savarese, Winn & Criminisi 2006).

2.1.2 Regional and Local Feature

The regional/local feature describes the visual information within a region or local area of the image.

The first direction is to cluster the nearby pixels as a region, and mark it with the global features extracted from the divided region. The feature associated with each region



(a) sun in sky (b) sun

(c) color histogram of a (d) color histogram of b

Figure 2.3: The rigid objects in images are hard to represented by global feature

is called the regional signature. For example, if the regions around the sun in Figure 2.3(a) and Figure 2.3(b) are extracted from all of the images, the system would be able to match them easily.

An essential process when it comes to extracting the regional feature is image partition. The process could be formulated as follows:

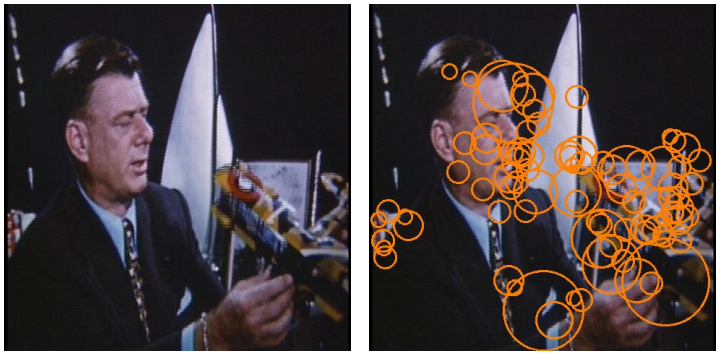
$$I_m = \{(z_1) + (z_2) + (z_3) + \ldots + (z_k)\} \tag{2.3}$$

where each z_k denotes a regional feature, which is a signature of the corresponding sub-region. Intuitively, a very first idea is that an image could be gridded, and it forms several equal sub-images (Wong & Pun 2008). Other similar applied approaches includes using K-means clustering (Chen et al. 2008) to divide original images.

Further development is utilizing the normalized cut criterion (Shi & Malik 2000), and the image partition problem is converted to a weighted graph partition problem. More complex approaches based on Bayesian statistical framework (Tu & Zhu 2002) , Gaussian Mixture model (Carson, Belongie, Greenspan & Malik 2002b) were also proposed recently.

However, computational complexity and reliability of partition remains an open problem (Datta et al. 2008), and the computation complexity always limits their application to real world search engines.

Another direction to overcome the drawback of global features is to use the local feature. Firstly, the system detects salient points in the images, which may be invariant



(a) Original Image (b) Local Feature Extracted

Figure 2.4: An example of local feature extracted from the image

to scale or affine transformation(Mikolajczyk & Schmid 2004). For example, in Figure 2.4, with a number of detected salient points and their nearby local area are denoted by orange circles. The scale of the feature is identified by the size of the circle.

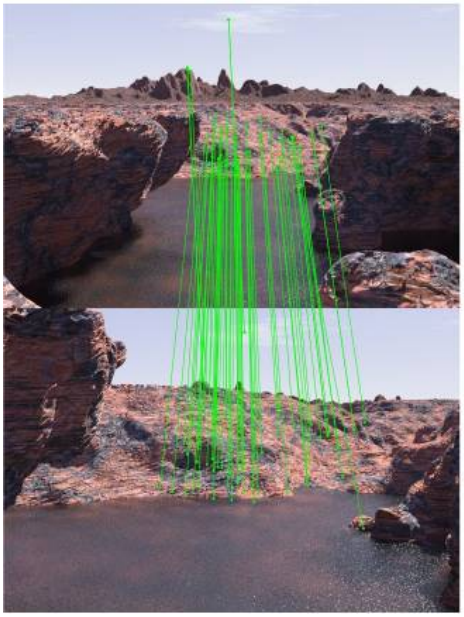


Figure 2.5: Match of local feature descriptors

As shown in Figure 2.5, the feature describes an area near the extracted points. Because the feature is invariant to distortion and rotation (Bay et al. 2008), it could be used to match the visual content of different scale in different images. The detailed progress in reference to the local features will be discussed in the Section 2.2.

The above features are all extracted to represent still images. Several features are specially designed for motion videos. A video can be seen as a visual volume of frames

in time. Various detectors are designed to detect the salient points in volume, which represents salient "spatial-temporal corners" or "sub-volume" of a video. For example, Spatial-Temporal Interest Points (STIP) (Laptev & Lindeberg 2003), (Cao, Tian, Liu, Yao, Zhang & Huang 2010) and Volumetric features (Ke et al. 2005).

The advantages of the local feature have been demonstrated in the recent CBVR development. This thesis also follows this research trend. The visual content representation framework in this thesis is based on the local feature extraction technology.

2.1.3 Visual Concepts

A visual concept is a visual feature on a higher semantic level. A visual concept could be manually annotated by human beings or learned and recognized from the low level visual features by automatic technologies like (semi-)supervised learning, video understanding, and visual object categorization (Oomoto & Tanaka 1993) . Both object entity and event concept, such as car, person, a goal, or airplane setting off, are defined as visual concepts. In this way, a video is represented as a collection of visual concepts. The video retrieval based on these semantic concepts is more straightforward for human perception. There have been many ambitious attempts to introduce the automatic visual concept annotation into the video retrieval (Adali, Candan, C, shing Chen, Erol & Subrahmanian 1996) (Decleir, Hacid & Kouloumdjian 1999) and solve the problem of mapping the query visual example of the user to desired visual concepts (Hauptmann, Christel & Yan 2008) .

Video retrieval based on visual concepts is an important research area. It is also very compatible with queries which are in modality of text description, because the concepts are always identified in the form of textual tags. However, the semantic concept recognition is always as difficult as low level feature based visual information retrieval (Smeaton et al. 2006), and the semantic visual concept extraction is still an open problem in the video understanding and retrieval research community. There is a significant challenge due to the fact that the visual content may have multiple, hidden or suppressed semantic meanings.

It should be pointed out that this thesis does not focus on the visual concept annotation, although the two share a similar motivation: that is , the problem with the low semantic level of the local features. We are aiming to develop the CBVR technology directly based on the low level local features.

2.1.4 Shot Boundary Detection

A typical raw video is always composed of a sequence of video **shots**, and each shot records a meaningful action or event, which is produced by a single camera. To facilitate finer visual content representation and retrieval, the video can be segmented into video shots. This process is a common prerequisite step for the automatic visual content access and retrieval. Moreover, compared to the raw video or other structural levels of video (frame,

scene, etc.), the video shot is more appropriate for indexing, retrieval and management (Cooper 2004). In this thesis, the term "video" is used interchangeably with "video shot", except that they are explicitly distinguished.

As is evident from Figure 2.6, the video can be divided into several intervals, and the shot boundary detection technology is to detect the key frames where shot transition happens.

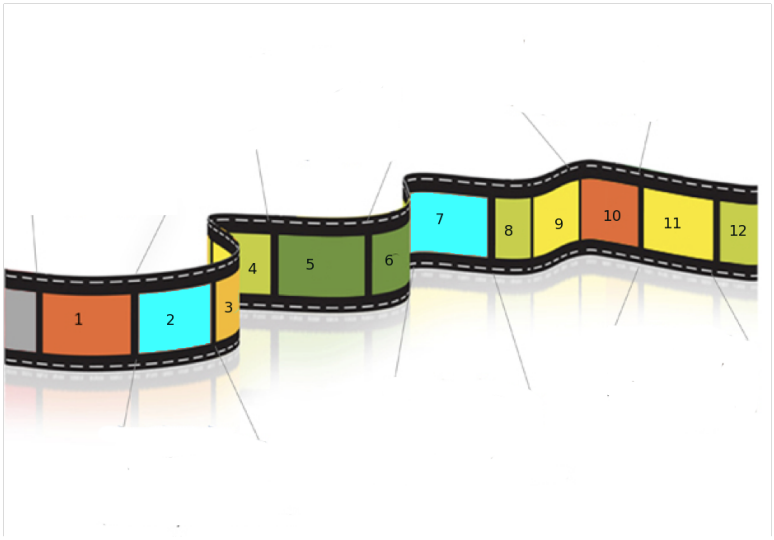


Figure 2.6: The video segmented as video shots

The shot boundary is always classified into two categories according to whether the change of shot is abrupt or gradual: cut and gradual transition (Yuan, Wang, Xiao, Zheng, Li, Lin & Zhang 2007) . Both the global feature (Boreczky & Rowe 1996) (Lienhart 1998) and the local feature (Zheng, Yuan, Wang, Lin & Zhang 2005) could be used to track the transition of visual content between adjacent shots.

Under the representation framework, the shot boundary detection can be transferred to a threshold determination problem. Initially, the standard thresholding model (Lienhart 2001) determines the threshold in a hard decision manner. However, the gradual transition would incurr interferences like dissolve, wipe, fade in/out, and a number of methods are proposed to address this problem. For example, adaptive threshold was proposed based on Bayesian formulation (Vasconcelos & Lippman 2000). A unified model (Bescos, Cisneros, Martinez, Menendez & Cabrera 2005) was proposed to map the inter-frames difference to the decision space and determine the threshold. Furthermore, some detection methods (Qi, Hauptmann & Liu 2003) (Cooper 2004) are based on machine learning approaches like supervised/unsupervised classification.

With the above-described development of the shot boundary detection technology, videos are often divided and stored as video shots in the visual resource. It provides a well prepared foundation for video indexing, query, and access.

2.1.5 Similarity Measurement

Although video representation and similarity measurements could be structured without considering the frame-intervals, the accurate video content similarity measurements are normally constructed as a two-step process: i) to measure the frame level similarity with image similarity measurement function; ii) to calculate videos' similarity based on the frame level similarity.

Recent years have seen a large number of visual similarity frameworks proposed in recent years. The motivations of the visual content similarity measurement design can be summarized as follows (Datta et al. 2008):

- Agreement with visual similarity;
- Robustness to noise;
- Computational efficiency;
- Regional-based query match.

The similarity measurement frameworks differs from each other according to the used visual feature for representation model. Figure 2.7 (Datta et al. 2008) shows some examples of the similarity measurement technologies developed for the corresponding visual feature.

Given the progress that took place in the computer vision field, we could extract the local or regional features to represent the image/video. A key-frame of video could be represented mathematically by a single vector (global feature), multi vectors (feature based on local region), or a summary of vectors, or index of entity. A video could be represented as a number of key-frames. The similarity measurement has been formulated as computing the difference between the representations.

In the Figure 2.7, the distances are calculated as a measurement of "dissimilarity" by these technologies.

The problem could be formulated, in general, as computing the distance between two sets of vectors.

We generally denote a vector of feature to represent an image as $I_m = \{(z_1^{(m)}, t_1^{(m)}), (z_2^{(m)}, t_2^{(m)}), \ldots, (z_k^{(m)}, t_k^{(m)})\}$, where z_i represents a feature vector and t_i represent the weight assigned to the vector in the image. Given two images, $, m = 1, 2$, the first step is to match the components of I_1 and I_2 one by one, and the distances calculated are summed up at the end to calculate the distance between the two images.

Wang et al. (Wang, Li & Wiederhold 2001) proposed an advanced metrics that softly weights the distance between the regional features. It distributes the weight factors $t_i^{(1)}$ and $t_j^{(2)}$ to the significant factor $s_{i,j}$ for a pair of vectors $z_i^{(1)}$ and $z_j^{(2)}$. The distance between two images is aggregated from the pair-wise distance between the vectors as:

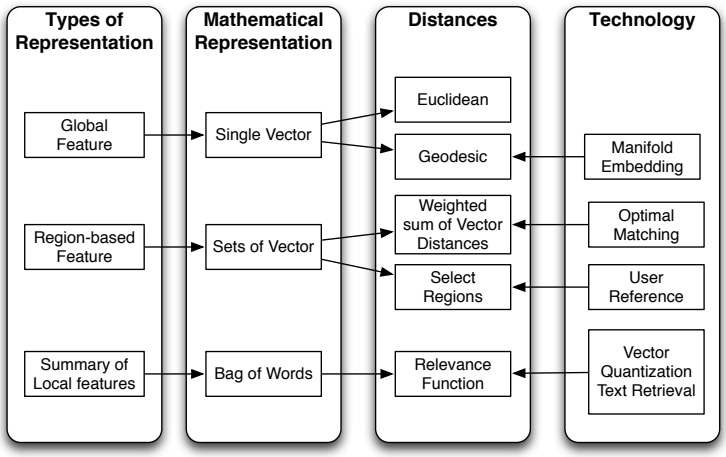


Figure 2.7: The different types of image similarity measurement

$$D(I_1, I_2) = \sum_{i=1}^{k_1} \sum_{j=1}^{k_2} s_{i,j} \cdot d(z_i^{(1)}, z_j^{(2)}) \tag{2.4}$$

where the weights $s_{i,j}$ are determined according to various constraints, for example $\sum_i s_{i,j} = t_j^{(2)}$ and $\sum_j s_{i,j} = t_i^{(1)}$. One special case of Equation 2.4 is proposed by using Hausdorff Distance for image retrieval (Ko & Byun 2002). It matches the $z_i^{(1)}$ to its closest vector within I_2 and vice versa. The distance between images is defined as:

$$D(I_1, I_2) = \max(\max_i \min_j d(z_i^{(1)}, z_j^{(2)}), \max_j \min_i d(z_i^{(1)}, z_j^{(2)}) \tag{2.5}$$

Another approach is to measure the two images by seeking a $s_{i,j}$ with which the distance $D(I_1, I_2)$ is minimized. The definition of the distance based on the above idea is thus:

$$D(I_1, I_2) = \min_{s_{i,j}} \sum_{i=1}^{k_1} \sum_{j=1}^{k_2} s_{i,j} \cdot d(z_i^{(1)}, z_j^{(2)}) \tag{2.6}$$

This distance is identical to the Mallows Distance in the case of discrete distribution (Mallows 1972).

The Earth Mover Distance (EMD) is one of the efficient algorithms to approximately measure the Mallows Distance between a set of visual vectors and another set of visual vectors. One advantage is that EMD allows for modelling of the similarity between the local regions within an image and another images (Rubner, Tomasi & Guibas 1998).

After the frame level similarities have been computed, the video similarity can be structured. We generally identify the frame-level similarity as $sim(I_i, I_j)$, and the video similarity could be computed with the highest similarity score among all possible pairs

of key frames compared. This method is especially suitable for the video shot similarity measurement, which is proposed by Peng et al. (Peng & Ngo 2005a). The similarity measure between two video shots can be formulated as:

$$sim_{v_1,v_2} = \max_{\boldsymbol{f}_1 \in v_1, \boldsymbol{f}_2 \in v_2} sim(\boldsymbol{f}_1, \boldsymbol{f}_2) \tag{2.7}$$

Another method is to compute the average similarity between all possible the key-frames pairs (Shang, Yang, Wang, Chan & Hua 2010) , (Ren, Lin, Zhang, Tang & Gao 2009a). It could be formulated as follows:

$$sim_{v_1,v_2} = \frac{1}{N_1 \times N_2} \sum_{\boldsymbol{f}_1 \in v_1} \sum_{\boldsymbol{f}_2 \in v_2} sim(\boldsymbol{f}_1, \boldsymbol{f}_2) \tag{2.8}$$

where N_1 and N_2 are the number of key-frames sampled from the two videos respectively.

Besides, the frame is as a part of the video, and the video can be naturally formulated as a set of vectors, each of which is the representation of a frame. The EMD algorithm could also be used for the video similarity measurement (Peng & Ngo 2005b). Moreover, the temporal order of frames can also be involved in the video similarity, which will be discussed in Section 2.3.

Under the BovW framework, many similarity measurement approaches used in text retrieval are also applied in the frame level and video level similarity measurement, which will be discussed in the following section. This thesis aims to improve the BovW framework, and the frame level similarity measurement is built upon those methods.

It has been demonstrated that good feature should satisfy the requirement to overcome occlusion, rotation, translation, change of view points and illumination. This thesis utilizes the popular local feature and the videos similarity measurement method proposed by Peng et al. (Peng & Ngo 2005a). However, a major disadvantage of local feature is that it is always of high dimension. Both the visual feature matching and video similarity measurements involve in high computational expense, which strongly harms practical feasibility of CBVR technology. The Bag-of-visual-Word framework was proposed in such background and is objective to solve this problem.

2.2 Bag-of-visual-Words Framework

The CBVR framework based on the BovW model was initially proposed by Sivic et al. (Sivic & Zisserman 2006), whose idea was to approximate the textual information retrieval. The general structure of the framework is shown in Figure 2.8.

The core idea is to quantize visual features to a limited number of clusters, and the clusters are used as basic elements to model the video content. The elements are named as visual words, and the visual word collection is called visual vocabulary.

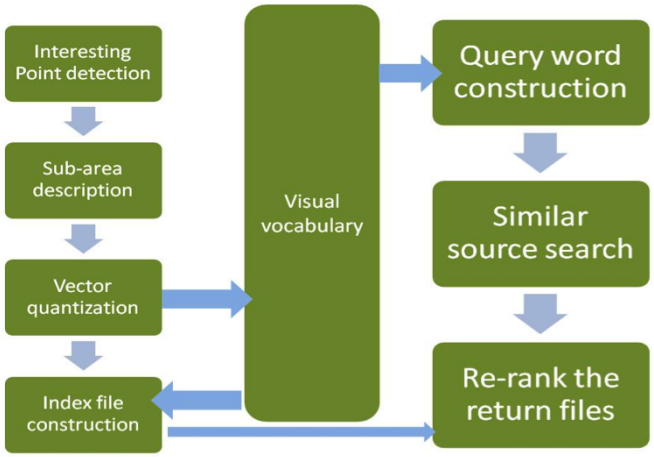


Figure 2.8: The basic architecture of CBVR based on bag of visual word

The basic architecture of query by example video retrieval consists of several function modules: salient content detection, descriptors generation, vocabulary building, video indexing, rank videos according to the similarity.

The technology development associated with the function modules is reviewed in the following sections.

2.2.1 Salient Points Detector

As discussed in the last section, visual content is normally represented by local features describing a number of regions extracted surrounding salient points. The salient points are always called points-of-interest in the research community, and the points-of-interest detection algorithms are always called detectors.

The points-of-interest are expected to be insensitive to local geometric and photometric changes. The main idea is that the extreme points are more insensitive to these distortions. Several types of detecting technology based on edge detectors are developed, generally called Corner-interest detector, because they are originally inspired by the idea to detect the intersection of multiple edges. For example, Harris-Laplacian (Mikolajczyk & Schmid 2004) ascertains the points with maximum dissimilarity to neighborhood through the Harris function.

Another type of points-of-interest detector is often referred to as blob-detector, which aims to detect extreme points in the local area. The Hessian-Laplacian (Mikolajczyk & Schmid 2004) localizes the point-of-interest by selecting the extreme response to the Hessian determinant and local maxima to Laplacian of Gaussian. Laplacian of Gaussian (LoG) was proposed (Bretzner & Lindeberg 1998) to detect local extremes, both in the

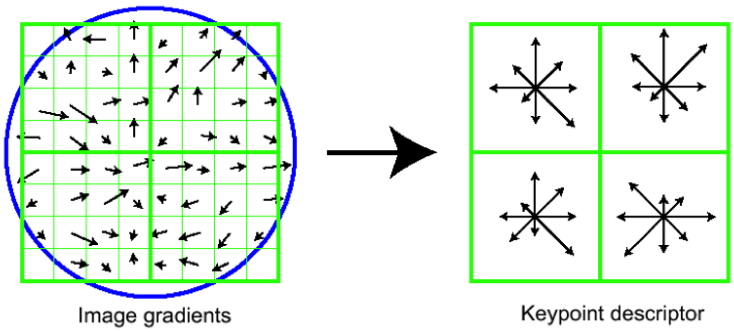


Figure 2.9: SIFT Descriptor

scale space and the pixel space. With this approach, the pixel space is initially smoothed by the Gaussian Kernel to reduce noises. Difference of Gaussian (Lowe 2004) is an effective approximation of LoG.

Several experimental results (Jiang & Ngo 2009) show that the LoG and DoG outperform the Harris-Laplacian and affine Harris in object categorization tasks. It is argued that the blob-like detector is more suitable than corner-like for the interesting object representation.

Another blob-like region detector is the Maximally Stable Extremal Regions (MSER) method (Matas, Chum, Urban & Pajdla 2004), which is designed to find the stable image region in different viewpoints. It has a number of advantages such as invariance to affine transformation and multi-scale detection.

Other than visual content detection, some researchers have utilized the oversampling scheme to capture sufficient information from the visual resource (Chum et al. 2011). Multiple detectors can be applied simultaneously to the image to capture more information. However, these schemes may suffer from expensive computational cost and higher risk of noise information.

2.2.2 Feature Descriptor

After the salient regions are extracted, each region should be modelled as a set of ordered data. This formulation function is called feature descriptor. One of the most widely used feature descriptors for the BovW framework is the SIFT feature.

Scale Invariant Feature Transform (SIFT) was first proposed by Lowe (Lowe 2004). An orientation (gradient) is assigned to each salient point to record the affine transformation. As shown in Figure 2.9 (Lowe 2004), for each point at each scale in the region around the salient point location, the gradient magnitude and the orientation are computed.

Next, an orientation histogram is created. It consists of 36 bins covering a 360 degree range of orientations. The points are weighted by the Gaussian weighting function. Each descriptor consists of four histograms, and each histogram has eight bins. This produces

a vector with 4*4*8=128 elements. The normalization of that vector will increase the invariance to illumination changes. As a result, the signature z of the SIFT feature normally contains a 128-dimensions vector s, a main direction vector d, a scale value σ, and a physical position x, y determined by the detector used.

Other popular local features utilized by the researchers include *PCA-SIFT* (Ke & Sukthankar 2004) and *SURF* (Bay et al. 2008), Gradient Location and Orientation Histogram (Mikolajczyk & Schmid 2005), and etc.

In this thesis, the retrieval system is built upon the most commonly used detectors and features: LoG and DoG, SIFT and SURF, which provide a mathematic description of local visual information. However, directly matching the local visual information via key-point features is neither efficient nor effective . We focus on how to better use it for relevance prediction, rather than improving feature description.

2.2.3 Visual Vocabulary

As discussed in Section 2.1.5, direct similarity measurement between the features is a computationally expensive process, because an advance local feature is always a high dimensional vector and an image may contain many features. The computational cost is one of the most important bottlenecks for large scale CBVR. To simplify the similarity comparison between the features, the solution provided by the BovW framework is to cluster the features to a number of categories. In this way, the feature similarity is simplified as a boolean computation: the categories of the features match or not.

In most previous works (Sivic & Zisserman 2006) , (Niebles, Wang & Fei-fei 2006), the K-Means algorithm (Lloyd 1982) is utilized to cluster the descriptors without supervision. The number of clusters K ranges from 60 (Ren, Lin, Zhang, Tang & Gao 2009b) to 32,357 (Zhang, Tian, Hua, Huang & Li 2009), and K is also the size of vocabulary. It is argued that a larger vocabulary is reasonable for object categorization and semantic retrieval (Jiang & Ngo 2009).

To reduce the computation cost of K-means for large scale clustering, Philbin et al. (Philbin, Chum, Isard, Sivic & Zisserman 2007) have proposed an approximate K-means method to build a larger scale visual vocabulary. Similarly, Nister et al. (Nister & Stewnius 2006) used the hierarchical K-means algorithm to construct a vocabulary tree, which resulted in significant reduction of the computation cost. The method scales the size of vocabulary up to 1,000,000, because it recursively uses the K-Means to quantize the descriptors generated.

Above methods employe the hard mapping scheme, which maps a feature to a single visual word in the original BovW model. Other than the clustering algorithm, the Gaussian Mixture Model (ByJason Farquhar & Shawe-Taylor 2005) can also be adopted to generate the visual vocabulary. The GMM model softly assigns a visual feature to multiple visual words probabilistically. Moreover, supervised learning information (Fernando et al. 2012)

is exploited based on the GMM model to improve the discriminative power of visual words.

A main problem with visual vocabulary in the traditional BovW is that it is generated by an unsupervised cluster algorithm, which may influence the descriptive ability of visual word. The supervised learning method was also adopted to select discriminative visual words and expand the visual vocabulary (Zhang, Tian, Hua, Huang & Li 2009). The visual vocabulary can also combine other information with the visual feature. For example, Zhang et al. (Zhang et al. 2010) proposed to build visual vocabulary incorporating the contextual information into more informative visual words.

The visual vocabulary buildt based on the unsupervised methods may lead to quantization errors. However, the semantic information is not always sufficient to build contextual visual vocabulary for general domain video retrieval. This thesis aims to consider the spatial-temporal context to improve the visual vocabulary building.

2.2.4 Video Representation and Indexing

The video indexing process is to construct a lookup table of the video collection to facilitate efficient retrieval. Similar to other representation frameworks discussed in Section 2.1.5, a frame in a video is often represented by the BovW framework as a weighted vector of the visual words appearing:

$$I = \{(w_1, t_1), (w_2, t_2), (w_3, t_3), \ldots, (w_K, t_K)\} \tag{2.9}$$

where w_i is i^{th} visual word and t_i is a weight assigned to it.

Similar to the textual IR community, the weighting scheme (Baeza-Yates & Ribeiro-Neto 1999) is important for retrieval performance. There are three major term weighting scheme used for CBVR: Binary, Term Frequency (TF), TF-inverse document frequency (TF-IDF). The Binary scheme uses "1" or "0" to identify whether the word appears in the frame. The TF scheme weights a visual word according to how many times it appears in the frame. The TF-IDF takes document frequency into consideration: the more frames in which a visual word appears, the less information that visual word contains.

Recently, a proposed scheme (Jiang & Ngo 2009) weights a single feature to multiply nearby visual words in the descriptor space. Because it differs from 1-to-1 hard matching, it is named as soft matching scheme. A visual word may be correspond to several possible visual meanings, and the soft matching will reduce the risk of losing relevant information. It has been proven that it can improve the object recognition performance, however, the improvement decreases slightly along with the increasing vocabulary size.

In addition to simple linear vocabulary, the vocabulary tree (Nister & Stewnius 2006) provides a scalable and discriminatory method to organize the vocabulary of the visual words rather than a simple list. In this way, the videos are organized in a hierarchical structure. It is also argued that the hierarchical structure can elaborate linguistic and

ontological factors of the visual words (Jiang & Ngo 2009).

If a large visual vocabulary is applied, the video representation becomes very sparse. To speed up index and query, the videos collection can be stored in the inverted file indexing structure, which is inspired by the textual information retrieval.

The above weighting schemes neglected the spatial and temporal information, and the visual words are assumed to be independent in the video representation, indexing, and query, which will affect the retrieval performance of the BovW framework.

2.2.5 Similarity Measurement

The retrieved videos are normally ranked according to their visual similarities to the query example, each of which is normally computed based on the key frame similarity in Section 2.1.5. The key frame similarity under the BovW framework can be computed by cosine function:

$$sim(f_d, f_q) \approx \frac{\sum_{i=1}^{K} f_q(w_i) \times f_d(w_i)}{l(f_d) \times l(f_q)} \tag{2.10}$$

where $l(f)$ is the L^2-norm of a vector.

The linear scan of the videos collection for ranking score computation is computationally expensive for the real-time video detection. Some non-linear methods under the BovW framework are proposed to achieve a higher ranking efficiency. For example, Locality Sensitive Hashing (LSH) (Hu 2005) is proposed to reduce the high-dimension score computation of large scale data collection. Another similar method proposed for image retrieval is based on the min-hash (Chum, Philbin & Zisserman 2008) algorithm. The ranked results retrieved based on the BovW framework can be re-ranked according to other information constraints ignored by the BovW model, for example, spatial constraint (Sivic & Zisserman 2006).

The BovW model is designed to approximate the function of the Bag-of-Word model in the textual information retrieval. Inspired by the success in the textual information retrieval certain types of technology like relevance feedback (Hopfgartner 2007) are also used to develop the CBVR technology.

The visual words is not as effective as textual word for retrieval, because the spatial-temporal relation between visual words has been ignored, despite its obvious importance for video retrieval, mainly due to two reasons. First, the visual words within a frame are assumed independent of each other and the spatial relationship is discarded. Second, the temporal motions of visual words between the sequential frames are neglected. The spatial-temporal relation would definitely link individual visual words as a whole by filling up their blank context and better describing the visual object.

2.3 Improving BovW Framework via Spatial-Temporal Information

As discussed in the previous section, the ignorance of spatial-temporal information is one of the largest drawbacks of the BovW model for CBVR technology. The spatial temporal layout of the visual words is obviously a clue of the visual object, which determines the semantic of the visual content. Intensive research has been investigated to discover the spatial-temporal information hidden in the videos. Furthermore, the spatial-temporal constraint can be seen as context of the visual words, which can be incorporated into the advance similarity function to improve the CBVR technology. The spatial-temporal relationship between the visual words also determines discriminative ability of the visual words, and it is utilized to build a more descriptive visual vocabulary and represent the visual content more effectively. Related works of the above ideas are discussed in the following sections.

2.3.1 Spatial-Temporal Information Modeling

A visual word, which contains little direct information about high-level meaning, and one object always consists of more than one visual words. The spatial-temporal layout of several visual words may form the structure of a visual object, and thus the spatial-temporal connection between the visual words should be modelled to represent the visual object.

Spatial Information A straightforward discovery method to model the spatial relationship which exists between a pair of visual words is to count their co-occurrence frequency. The co-occurrence between visual words can be seen as a simple way to model spatial information in the image (Galleguillos et al. 2008). Like the textual information process and retrieval, the statistical co-occurrence of visual words is utilized to discover the latent topic, for example, using pLSA (Bosch, Zisserman & Muñoz 2006).

The co-occurrence may be too rough, because not all of the co-occurrence on a single image is meaningful. The spatial relationship between the visual words can be refined by K-Nearest-Neighbors or ϵ-Nearest-Neighbors (Yuan & Wu 2008). Only several nearest neighbors or neighbors appearing within a certain range are considered as spatial related visual words.

Another way to refine the spatial information modelling is by partitioning an image into a series of sub-images, following which the co-occurrences are trimmed to the visual words appearing within the identical sub-images. For example, the visual words in the images can be grouped with respect to the color appearance (see Figure 2.10). Cao et al. (Cao & Li 2007) proposed a Spatial Coherence Latent Topic Model (Spatial-LTM) that models the visual words co-occurring within a local area that has similar color appearance. The segmentation method is always computationally expensive, and how to obtain a reliable

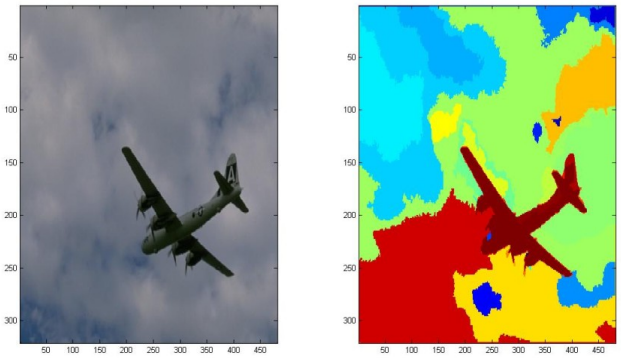


Figure 2.10: Image segmented by color appearance

segmentation remains an open problem (Datta et al. 2008). As a result, the spatial information is some times encoded in a simpler way, for example, the image is equally divided into several grids. The visual words are encoded with the indexed grid to represent the visual content (Zhang et al. 2011).

However, the globally spatial coherence of visual words generally loses the scale information. One method to address the scale problem is to consider the segmentation of the image on different scale levels. Spatial Pyramid Matching (Lazebnik, Schmid & Ponce 2006) is proposed to approximate the global geometric correspondence. It partitions the image into hierarchical sub-images to model the spatial relation of visual words and improve the similarity measurement.

Another direction is to directly compute and encode the distance between the visual words to model the scale information. For example, Correlograms of visual words was proposed by Savarese et al. (Savarese et al. 2006) to capture the spatial correlation on different scales and visual words pairs as well. However, in these methods, the distance (scale information) between visual words is quantized into fixed bins to save the storage and computational cost.

The spatial model of the fully connected visual words will be very complex if there are a large number of visual words within the frame. The complexity will grow exponentially along with the increase of the visual words. As a compromise, several sparser topologies have been proposed. A "part and structure" model named *constellation model* (Fergus, Perona & Zisserman 2003) was proposed to capture the objects with the local features and their connection. However, it suffers from its complexity, and it could only deal with 20-30 regions per image. This means that the method could not be used for BoW framework, because BoW normally generates 400-600 local feature descriptors per image.

Recent years have seen the proposal of the *star topology* (Fergus, Perona & Zisserman 2005),which simplifies the representation model by reducing the connection between the nodes in the model. A hierarchy spatial connection model (Bouchard 2005) was proposed

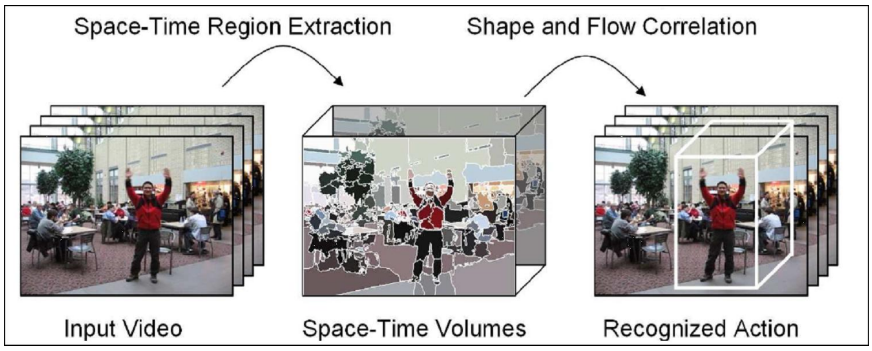


Figure 2.11: Process Video as Spatial-Temporal Volume

as a generative model to characterize the geometry of the visual object categories. The Markov model can significantly simplify the full connections model to a multiplication of the nearest connection. For example, the 2D-HMM model has been proposed (Othman & Aboulnasr 2003) to model the dependency between the co-occuring visual words. However, there is an argument that the assumption made in the Markov model, in which a visual word depends only on its nearest neighbor, is too strong and will lead to the loss of spatial information. Wu et al. (Wu, Hu, Li, Yu & Hua 2009) considered correlations between each pair of co-occurring visual words in a frame to form a visual language model with supervised leaning.

Recent work also utilized the probabilistic model to estimate the relationship between the visual words. Generally, the visual words appearing in the image are treated as observed samples, which are assumed to follow a certain distribution. For example, Liu & Chen (Liu & Chen 2009) have modeled the visual object with a mixture of Gaussian distributions of visual words, and utilized the EM algorithm to determine the parameters of the distribution. This method demonstrates that the spatial proximity may be related to the probability that they belong to an identical visual object.

As demonstrated in previous works, the spatial proximity between each pair of visual words is always important. However, there lacks an efficient way to quantitatively model the statistical spatial relations between the visual words under the BovW framework. This thesis aims to discover the spatial correlation based on the proximity of visual words.

Temporal Information The temporal information is largely contained in the video data. It is extremely important to represent the temporal visual content, such as the human action and other events in videos. For example, it is hard to distinguish between the launching and landing of an airplane without the temporal information.

Based on our critical investigation of the related literatures, the temporal information modeling can be roughly grouped into two types of method. The first one is to treat video as spatial temporal volume (Ke et al. 2005) (see Figure 2.11), and transfer the temporal axis to the third spatial axis. The spatial-temporal information within the video

is modeled with 3-D spatial information.

Another group of methods model the temporal information via capturing temporal motion in the videos. Like common video analysis tools, e.g. Optical Flow method, some methods track the visual words on the continuous frames of video, and the modeled appearing/motion/dissolving of visual words represents the temporal information of video (Brand, Oliver & Pentland 1997) (Niebles et al. 2006).

Wang et. al. (Wang et al. 2008) have argued that the relative motion of the visual features records the temporal pattern of the video content. They have proposed a method which assigns the visual features with relative motion directions and models the visual event by the visual features and their relative motion directions. This work has shown the descriptive ability of the temporal information to represent the visual event in the videos. However, it directly embeds the temporally relative motion into the video representation, which raises up the storage expense.

Kovashka & Grauman (Kovashka & Grauman 2010) have proposed to select a series of areas, in which the visual features and their orientations with respect to the central visual feature are recorded. The temporal actions are modeled by the organized visual words based on their relative motion to the central visual word. This method was proposed to recognize the human action.

For key frames based videos indexing and access system, modelling the motions is always a more convenient method to discover the temporal information. However, few methods have explored the possibility of building a uniform formulation to model the temporal and spatial correlation between the visual words. The present thesis will focus on this topic and propose a novel method.

2.3.2 Enriching Content Representation and Indexing with Geometric Information

The modeled spatial-temporal information using the previous methods could be utilized to enrich the information contained in the BovW based visual content representation. Firstly, the additional geometric information is added into the visual words vectors, which will be beyond the histogram. Secondly, the spatial-temporal information will be used to address the descriptive ability of visual words, and indirectly improve the content representation.

As shown in Equation 2.9, the classical BovW based image representation is based on a term-weighting scheme. Taking SIFT/SURF as an example, the additional geometric information, *e.g.* physical position, main direction, and scale factor, associated with the feature, can be stored with the visual word (Zhao et al. 2010). The new formulation of visual representation is as follows:

$$I = \{(w_1, \boldsymbol{z}_1), (w_2, \boldsymbol{z}_2), (w_3, \boldsymbol{z}_3), \ldots, (w_n, \boldsymbol{z}_k)\} \tag{2.11}$$

where $\boldsymbol{z}_i$ is a package of geometric information: $\boldsymbol{z}_i = \{x, y, d, \sigma\}$, and n is the number of visual words appearing in the image. As previously stated, the storage cost of full spatial information as previous is high, and calculating the relative spatial relation is computationally expensive. As shown in Section 2.3.1, the physical position $\{x, y\}$ is proposed to be quantized into discrete values (Jégou, Douze & Schmid 2010), which relatively saves the storage cost of representation, especially when multiple visual words are concatenated together.

Similar to spatial information, the additional temporal information can also be utilized to improve the retrieval performance. Qu et al. (Qu, Bashir, Graupe, Khokhar & Schonfeld 2005) introduced a series of methods to detect and represent the view-point invariant motion trajectory of an object. The visual words motion trajectories are extracted and incorporated into visual content representation, based on which the video classification, retrieval and recognition could be performed.

The storage expense normally increases if we directly enrich the representation with geometric information. To avoid the extra storage cost, researchers have proposed approaches utilizing the spatial relationship as the constraint condition to improve visual content representation. Liu & Chen (Liu & Chen 2009) proposed to emphasise the Objects-of-Interest in the visual content representation. The object extraction highlights certain visual content but ignores other contents. Temporal information was also used by Sivic & Zisserman (Sivic & Zisserman 2006) to select stable features, which continuously appear in several neighboring frames.

In this thesis, we would like to keep the simple structure of BovW, and so it also aims to construct an efficient spatial-temporal constraint to improve the video representation. But unlike other existing technologies, here we aim to identify and emphasize more descriptive visual words rather than interesting visual content via the spatial-temporal constraints.

2.3.3 Descriptive Visual Vocabulary

As discussed in Section 2.2.3, in the earliest attempt to build the BovW representation, Sivic et al. (Sivic & Zisserman 2006) proposed to use an unsupervised algorithm, *e.g.* K-Means, to cluster the SIFT descriptors and generate the visual vocabulary. The semantics of the visual words are at a low level, owing to the unsupervised clustering.

A method proposed to promote the semantic level of visual words is to discover the semantic meaning of visual words through supervised learning (Zhang, Tian, Hua, Huang & Li 2009). The descriptive visual words for a pre-defined category of visual objects are selected by the algorithm. However, the practical utility of the supervised learning algorithm is always limited by the lack of manual tags. Some unsupervised statistical learning methods have been proposed to discover the latent topic (Cao & Li 2007) of visual content, where the topic consists of a number of frequently co-occurring visual words.

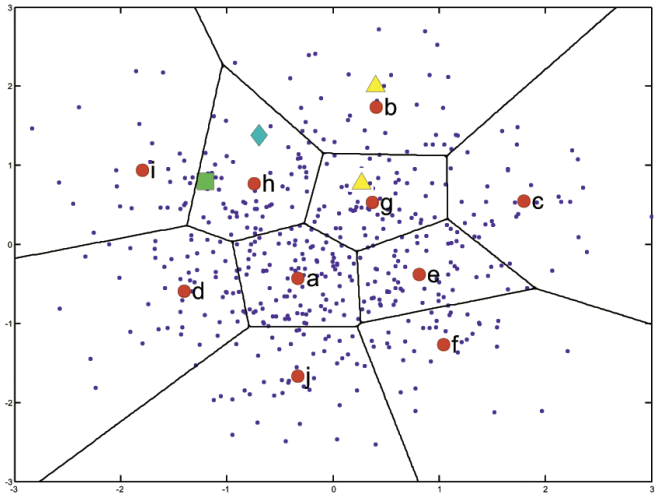


Figure 2.12: Each red round represents a center of a visual word. The ambiguity: visual word h has multiple meanings (green rectangle and blue diamond); The synonyms: visual word b and g has identical meaning (yellow triangles)

Furthermore, the spatial-temporally co-occurring visual words are proposed to be concatenated into the combination of visual words, which is termed as a visual phrase or visual sentence (Zheng & Gao 2008). Such combination tends to be more informative than the single visual word, and thus it promotes the descriptive ability of visual vocabulary. The pattern of concatenated visual words can either be discovered by supervised (Zhang, Tian, Hua, Huang & Li 2009) or unsupervised learning (Zhang et al. 2011) techniques, which aim to discover stable and meaningful spatial-temporal connection between visual words. Other than simple visual words combination, Zhang et al. (Zhang et al. 2010) have proposed to build a semantic visual vocabulary by concatenating spatial contextual information to groups of visual words. Visual word correlogram (Savarese et al. 2006) and correlation were proposed to represent the visual content for objects recognition.

It has been demonstrated by these works that the co-occurrence and spatial-temporal correlation between visual words are clues of descriptive ability, which inspired the present work to utilize this information to address the descriptive visual words.

The descriptive ability of visual vocabulary is also interfered the problem of ambiguity and synonyms. In Figure 2.12, the black points is the samples of visual features, and red rounds denotes the centers of visual words. Rectangle, diamonds and triangles, represents a few visual features, whose meaning is denoted by its color and shape. The ambiguity is that features of different meanings (blue and green ones) are mapped into identical visual words. Most of the aforementioned approaches optimize the visual vocabulary by addressing the ambiguity problem with the idea that additional information helps to distinguish the blue and green feature in order to reduce false positive matching.

The synonyms problem means that some feature descriptors of the same meaning are

mapped onto different visual words, for example, the yellow triangles are mapped into visual words g and h separately in Figure 2.12. This problem has also attracted significant attention in the research community (Perronnin, Dance, Csurka & Bressan 2006). A major idea to tackle the problem is to softly map a feature to multiple visual words (Winn, Criminisi & Minka 2005) to reduce probability to lose the relevant information. Jiang et al. (Jiang & Ngo 2009) built an ontology that specifies the relationship between visual words in the feature space, and softly assigns a feature into nearby visual words. For the application of image/video retrieval, Chum et al. (Chum et al. 2011) proposed to learn a generative model from data collections to expand the query with originally missing visual words.

This thesis aims to address the semantics of visual words and solve the synonyms problem in a novel perspective: using the spatial-temporal context of visual words (part of related works will be reviewed in Section 2.3.5). In real world applications, the visual vocabulary could also be constructed with multiple features, either by concatenating multiple features into a long vector before the quantization (Hsu & Chang 2005) or combining the visual words after quantizing the different features separately (Zhang, Liu, Ouyang, Lu & Ma 2009). However, this thesis only exploits vocabulary based on single feature (SIFT) to demonstrate our approach to improve the CBVR technology. Moreover, it can be easily applied to the visual vocabulary based on other single or combined features.

2.3.4 Ranking via Spatial-Temporal Constraints

The BovW framework ranks videos according to visual similarity between the query and videos, which is normally computed by the overlapped visual words contained in the two representation vectors. However, there may exists false matches in the content comparison. This problem has been noticed in the earliest BovW work (Sivic & Zisserman 2006). It has been proposed to rank the videos via a spatial consistency check, which is shown in Figure 2.13. The idea is motivated by the fact that the visual object has local spatial-temporal consistency, and the correctly matched visual object does not simply rely on matched visual words independently. In addition, the correctly matched visual words must be supported by the surrounding matched visual words. However, the local spatial consistency check is computationally expensive, and can only be used in the re-ranking of a small number of retrieval results (Sivic & Zisserman 2006).

The additional geometric information, *e.g.* scale, main direction, can also be used as similarity measurement constraint to reduce the false matched visual content. For example, Jegou et al. (Jégou et al. 2010) proposed a visual words matching framework based on the so-called geometric weak consistency constraints (WGC), which assumed that main direction and scales of matched visual words should have some degree of consistency. This method and its variant Tight Geometric Constraint (TGC) achieved promising performance for applications such as near duplicate image/video search and video automatic

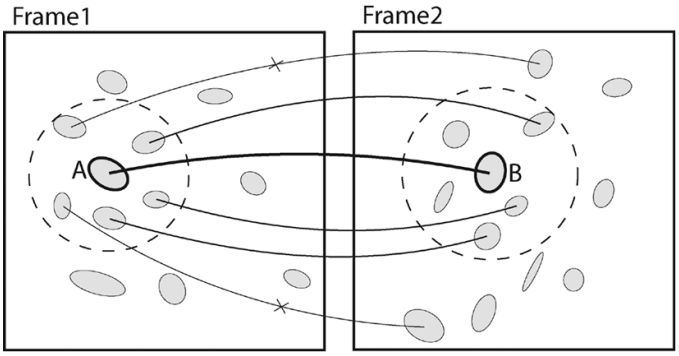


Figure 2.13: Spatial Consistency of Visual Words

annotation (Zhao et al. 2010). Spatial Pyramid Matching and its variants (Lazebnik et al. 2006) address the similarity comparison problem on a different scale, and generally promote the accuracy.

A topic model (Cao & Li 2007) enhanced image similarity is always formulated as follows:

$$sim(I_d, I_q) \approx \lambda \times sim_w(I_d, I_q) + (1 - \lambda) \times sim_t(l_d, l_q) \tag{2.12}$$

where sim_w and sim_t identify the similarity based on the statistical visual words and latent topic respectively, and λ is a combination factor.

Other than latent topic discovery, Chum et al. (Chum et al. 2007) proposed a generative model based the query expansion model, which is built based on the first round retrieval results, for image retrieval. The spatial dependency between visual words in the pseudo feedback set and the original query is used to address the visual words related to the query. The retrieval performance is improved by the query expansion, and some relevant information missing from by the initial results are recovered.

Unlike these approaches, this thesis utilizes the spatial-temporal correlation between visual words rather than extra geometric or feedback information to construct the constraints for similarity scoring function to rank the retrieved results.

2.3.5 Context of Visual Words

The context of textual word is an important topic in the field of nature language processing to better understand the text resources. Similarly, the context of visual word is also important for visual content accessing and retrieval. In the computer vision field, the context of visual word is usually defined by the interaction between the visual word and other pixels, object, region, and information in other modalities (text, audio, and etc.).

The most commonly defined context of visual words is shaped by its interaction with other words in the nearby region. For example, Bolovinou and et al. (Bolovinou, Pratikakis & Perantonis 2013) defined the spatial correlogram between visual words as context and encoded it into visual content representation. The spatial context has also been incor-

porated into the contextual visual vocabulary construction in other works (Zhou, Wang, Wang & Feng 2010) (Zhang et al. 2010).

Except for the intra-frame context, the inter-frame context of visual words is also utilized for visual information matching, which is a typical method in textual information retrieval (Skov, Larsen & Ingwersen 2008). In CBVR, the inter-images correlation of the visual words was explored as the context (Zhou, Tian, Yang & Li 2010) to discover the latent semantic connection.

In recent work, the multi-modal information contained in the visual documents has attracted an increasing attentions. The multiple modality context is incurred by the multi-modal information. The information other than visual content can also be seen as context. For example, an image link graph is analyzed for image context exploration (Zhou, Tian, Yang & Li 2010). Textual information associated with visual content is defined as the semantic contexts in (Su & Jurie 2011) to address the disambiguation problem of visual words.

The semantical context of visual information is not always available for large scale CBVR. In most cases, we must sufficiently utilize the intra-image or intra-video context to develop the CBVR technology. Visual words correlation can be seen as a type of this context, and we will propose novel methods defining and utilizing this context to enhance the video retrieval model in this thesis.

As demonstrated in this section, utilizing spatial-temporal information would definitely be a key to further improve state-of-the-art BovW framework. However, existing works lack a uniform model for both spatial and temporal information, especially the relation between visual words. And queries in modality of visual example normally contain rich such information, but few theoretical investigation on how to effectively discover and efficiently use this information to develop CBVR technology can be found in the reviewed literatures. These challenges motivate us to propose our methods presented in next chapters.

2.4 Summary

The previous sections have reviewed the recent research progress in the CBVR and its related topics. The descriptive ability of global features suffers from the geometric inconsistency of visual content like scale variant, clutter and distortion. The quality of regional feature is heavily reliant on the reliable partition approach, and robust partition is still an open problem for researchers. It has been shown that the local feature based visual content describing methods have recently begun to dominate the CBVR research. A typical CBVR framework utilizing the local features is the BovW model.

The BoW framework is an attractive topic. It provides a compact and efficient framework and has shown a promising performance on visual content modeling for a number of research topics: object recognition, image retrieval, image annotation, and so on. Fur-

thermore, it has been found that extensive researches have been undertaken to enhance the BovW model with additional spatial-temporal information which is generally ignored by the classical BovW model, and some progress has been made.

However, there are still many challenges facing the application of the spatial-temporal information for BovW based CBVR technology.

Firstly, the traditional spatial-temporal information modeling relies on region based methods or neighborhood based methods, *e.g.* K-nearest-neighbors, ϵ-nearest-neighbors. More compact, quantitative, flexible, and uniform spatial-temporal information discovery and representation models should be investigated .

Secondly, the existing works revealed the possibility of utilizing the additional spatial-temporal constraint to measure the similarity. However, the relationship between the spatial-temporal constraint and term specificity of the visual words has not been fully discussed. Is it effective to address the descriptive visual words for the visual content representation via appropriate spatial-temporal constraints? This is still an open question.

Finally, the spatial and temporal information has been used to build more informative and contextual visual vocabulary for the BovW model. With this said however, it lacks theoretical analysis to compensate the quantization errors via the spatial-temporal context characterization.

Chapter 3

Spatial-Temporal Correlation Modeling

In the previous chapter, we have reviewed the literatures relating to retrieval models for CBVR. As we pointed out, one of the major limitations of the state-of-the-art BovW framework based CBVR model is the ignorance of spatial-temporal information. We reviewed a number of different methods for spatial-temporal information discovery under the classical BovW framework. Various modelling methods which utilizes this relation from a new perspective are presented in this Chapter.

First of all, we introduce our co-occurrence model for the visual words in a individual frame/image. It is defined according to the layout of a visual entity, which is always composed of co-occurring visual words. We formulate the co-occurrence model as a form of correlation matrix, and each element of the matrix is weighted according to pair-wise co-occurring visual words. Furthermore, the video level and videos collection level visual words co-occurrences are formulated as an accumulation function based on the frame level co-occurrence.

Although it has been assumed that the co-occurring visual words describe the identical visual entity, the fact is that an image/frame always contains rich information and more than one visual entities. This assumption could be refined to better model the correlation between co-occurring visual words. We would consider physical distance between the visual words and make a principled assumption that the spatial proximity represents the degree of relation. Based on a series of theoretical analyses, we select the Gaussian Function to model the spatial correlation.

Other than the spatial information, we also discover temporal information to model the correlation between the visual words. Here, we propose to tracked the temporal motion of visual words on the continuous frames, and define a motion vector for each visual word as its temporal action. In this way, another important assumption is made that the correlated visual words should move more coherently, because a visual object always tends to move

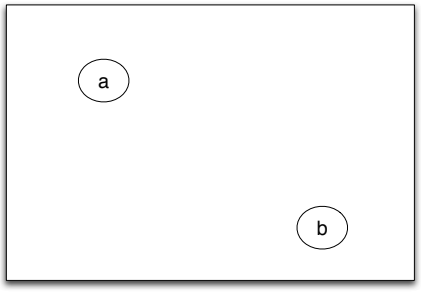


Figure 3.1: Co-occurring instances a and b of visual words w_i and w_j within a frame

as an integrated component. The motion coherent quantization function is also assumed to follow the Gaussian Distribution. The correlation matrix is then formulated based on a spatial and temporal correlation function.

Finally, we defined a concept, namely the Spatial-Temporal-Correlation (STC), to summarize the above correlations, and we present an approach to combine the spatial and temporal correlation. We will also analyze and discuss possible applications of the STC.

3.1 Co-Occurrence Model

In this section, we make an assumption that visual words appearing simultaneously in a single frame are actually semantically related to each other, and the co-occurrence indicates the relation degree. To model this co-occurrence between the visual words, we calculate the frequency with which they co-appear in a frame.

SIFT feature and approximate K-Means clustering are very popular visual content pre-processing algorithm and visual words generation methods in recent researches. To make our methods comparable with other work, we also use these two methods to pre-process the videos and generate visual words. It should be point out that our STC generation method can be applied for all types of visual words generated by different methods.

As discussed in Chapter 2, we want a quantization model to figure out which visual words are more closely correlated using spatial and temporal information. So we count the times they co-occur in the videos. Because this idea is quite straightforward that correlated words should more likely co-occur with each other. Matrix-like form is naturally utilized in this section to model the one-to-one relation between visual word. In this way, the relation is clearly quantized and easily to be used in future application.

3.1.1 Co-occurring Visual Words in A Frame

As defined in the previous chapter, the BovW framework describes a frame as $f = \{(z_1, t_1), (z_2, t_2), \ldots, (z_k, t_k)\}$, where z_i represents a feature vector and $t_i \in \{w\}$ represent the index of visual word assigned to the i^{th} feature vector according to the visual vocabulary. Each z_i

is called an instance of the corresponding visual word. For example, as shown in Figure 3.1, the local features a and b are instances of visual words w_i and w_j respectively.

The co-occurrence between a pair of visual words is defined by counting the instances co-occurring in the current frame. The computation function is defined as follows:

$$c^c(i,j) = \sum_{t=w_i} \sum_{t=w_j} 1 \tag{3.1}$$

In practice, to save the storage expense, the frame representation is always indexed in a simpler form. Normally, the frame can also be represented by a term frequency histogram vector $\boldsymbol{f}$ of the visual words as:

$$\boldsymbol{f} = \begin{pmatrix} tf_1 & tf_2 & tf_3 & \cdots & tf_K \end{pmatrix} \tag{3.2}$$

where an element of vector $\boldsymbol{f}(i)$ represents the term frequency of i^{th} visual words in the current frame, and K is the scale of visual vocabulary. It should be pointed out that the term frequency is here defined as raw frequency as an example, which represents the number of terms in the current frame. Other types of term frequency can also be utilized. For example, the binary term frequency can also be utilized in the vector, the frame level co-occurrence computed is then binarized.

In this histogram form of the frame representation, the number of each appearing visual words has been given in the vector, and the co-occurrence computation in Equation 3.1 can be redefined as follows:

$$\begin{aligned} c^c(i,j) &= \sum_{a=1}^{\boldsymbol{f}(i)} \sum_{b=1}^{\boldsymbol{f}(j)} 1 \\ &= \boldsymbol{f}(i) \times \boldsymbol{f}(j) \end{aligned} \tag{3.3}$$

where we can align the computed co-occurrence associated with i^{th} visual word to construct a row vector as follows:

$$\boldsymbol{c}_i^c = \begin{pmatrix} c^c(i,1) & c^c(i,2) & c^c(i,3) & \cdots & c^c(i,K) \end{pmatrix} \tag{3.4}$$

This vector is defined as **co-occurrence vector** of i^{th} visual word in the current frame, which actually represents the histogram of its co-occurring visual words. An example of the vector is shown in the Figure 3.2. Furthermore, the **co-occurrence matrix** of this frame can be easily constructed by aligning the row vectors associated with all visual words. The form of co-occurrence matrix is represented as follows:

$$\boldsymbol{C}^f = \begin{pmatrix} \boldsymbol{c}_1^c \\ \boldsymbol{c}_2^c \\ \vdots \\ \boldsymbol{c}_K^c \end{pmatrix} \tag{3.5}$$

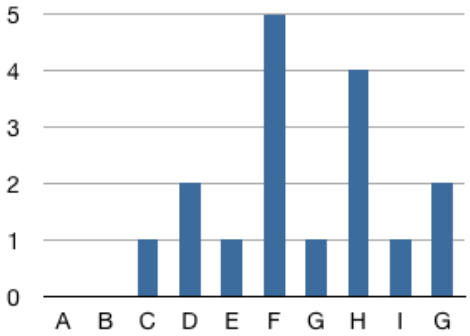


Figure 3.2: Co-occurrence Vector: The histogram of co-occurring visual words of A in a frame

where each entry of the co-occurrence matrix $\boldsymbol{C}^f$ is computed by the Equation 3.3: $\boldsymbol{C}^f(i,j) = \boldsymbol{f}(i) \times \boldsymbol{f}(j)$. Therefore, the computation of $\boldsymbol{C}^f$ is equivalently formulated as the **tensor product** of the term frequency vector $\boldsymbol{f}$ and itself. The formulated function is as follows:

$$\begin{aligned} \boldsymbol{C}^f &= \boldsymbol{f} \otimes \boldsymbol{f} \\ &= \boldsymbol{f}^T \times \boldsymbol{f} \end{aligned} \tag{3.6}$$

The co-occurrence matrix only considers the pair-wise co-occurrence, which is like a bi-gram model. The co-occurrence matrix can be easily expanded to N-order tensor to model the n-gram co-occurrence, which is its tensor product with the term frequency vector $\boldsymbol{f}$:

$$\boldsymbol{C}^f_{\boldsymbol{n}} = \boldsymbol{C}^f_{\boldsymbol{n-1}} \otimes \boldsymbol{f} \tag{3.7}$$

where $\boldsymbol{n}$ identifies the order of tensor $\boldsymbol{C}$.

However, the storage and computation expense will increase exponentially along with increasing n. In the remaining part of this thesis, the discussion and application only focuses on the bi-gram model, which is in the form of above co-occurrence matrix.

3.1.2 Co-occurrence Matrix for Videos

The co-occurrence matrix for a video could be in a similar form to the function for a frame presented in Section 3.1.1. If we represented a video with term frequency vector, in the same form as term frequency of a frame:

$$\boldsymbol{v} := \begin{pmatrix} tf_1 & tf_2 & tf_3 & \cdots & tf_K \end{pmatrix} \tag{3.8}$$

The co-occurrence computation for a video can be in exactly the same form as the frame level computation function:

$$\boldsymbol{C}^v = \boldsymbol{v} \otimes \boldsymbol{v} \tag{3.9}$$

However, this representation normally does not describe the content of video sufficiently, even for well segmented video shots. In CBVR, the frame based video representation is utilized more often. Typically, a video is represented by a sequence of frames:

$$\boldsymbol{v} := \begin{pmatrix} \boldsymbol{f}_1 \\ \boldsymbol{f}_2 \\ \boldsymbol{f}_3 \\ \vdots \\ \boldsymbol{f}_N \end{pmatrix} \tag{3.10}$$

where N identifies the number of frames contained in the video and each $\boldsymbol{f}$ is a vector representing the content of the corresponding frame. Compared to the representation in Equation 3.8, this representation sometimes has too much redundant information. Because many neighboring frames are actually very similar, the researchers often sample key frames to save the storage and indexing cost. For example, only the I-frames in MPEG videos are utilized.

If each $\boldsymbol{f}$ is assumed to be a row vector, this representation of video is equivalent to a term-frame matrix. The general form of the matrix V can be seen in Figure 3.3, where each entry of the matrix is a weight of the corresponding visual word in the relevant frame.

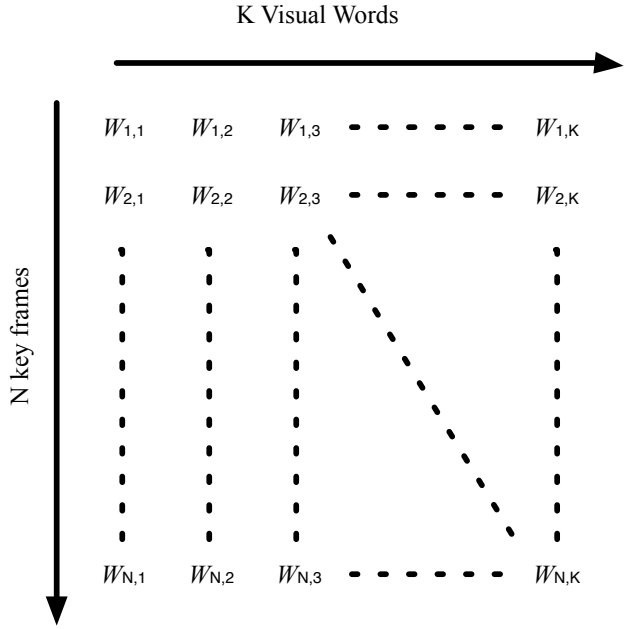


Figure 3.3: The term-key frames representation of video

This term-frame matrix is in the same form as the term-documents matrix, which is widely utilized in the textual information retrieval research domain. Similar to the method used in the textual information retrieval research (Kaliciak, Song, Wiratunga & Pan 2012), the co-occurrence matrix of a video is computed by the visual word-frame matrix V as follows:

$$\boldsymbol{C}^v = \boldsymbol{V}^T \times \boldsymbol{V} \tag{3.11}$$

Equation 3.11 is a multiplication between two very large but sparse matrixes, and as such it is not easy or necessary to directly compute the multiplications. Following Kaliciak et al. (Kaliciak et al. 2012), the computation could be simplified as follows:

$$\boldsymbol{C}^v = \sum_{\boldsymbol{f} \in v} \boldsymbol{f}^T \times \boldsymbol{f} \tag{3.12}$$

Equation 3.12 can be understood as an accumulation of frame level co-occurrences, because Equations 3.12 and 3.6 can be incorporated as follows:

$$\boldsymbol{C}^v = \sum_{\boldsymbol{f} \in v} \boldsymbol{C}^f \tag{3.13}$$

This shows that the video level co-occurrence has been modeled as a summarization of the frame level co-occurrences. For convenience, a video collection is also considered a collection of the key frames of the individual videos. Here, the representation of a videos collection d is defined as:

$$\boldsymbol{d} := \begin{pmatrix} \boldsymbol{v}_1 \\ \boldsymbol{v}_2 \\ \boldsymbol{v}_3 \\ \vdots \\ \boldsymbol{v}_{N_v} \end{pmatrix} \tag{3.14}$$

where N_v indicates the number of videos in the collection. Because the representation is in a similar form to Equation 3.10, the co-occurrence matrix can be computed in the similar way. It could be computed by accumulating the video level co-occurrence:

$$\begin{aligned} \boldsymbol{C}^d &= \textstyle\sum_{\boldsymbol{v} \in d} \boldsymbol{C}^v \\ &= \textstyle\sum_{\boldsymbol{f} \in d} \boldsymbol{C}^f \end{aligned} \tag{3.15}$$

In general, both a video or a video collection can be represented as a collection of key frames, whilst the co-occurrence matrix is formulated as the summarization of the frame level co-occurrence matrixes.

3.1.3 Co-occurrence and Visual Words Correlation

The true correlation of visual words is actually the semantic relationship between visual words, for example, the visual word representing "tyre" is related to "windscreen". In visual information understanding, this correlation is unfortunately unknown to the system without considering additional semantic information. In this thesis, we utilize the co-occurrence relationship to approximate the semantic correlation, and the co-occurrence is here defined as a co-occurring correlation.

Generally, a group of correlated visual words can be defined as:

$$ph_{1,2,3,\cdots,n} := \{w_{i_1}, w_{i_2}, w_{i_3}, \cdots, w_{i_n}\} \tag{3.16}$$

where n indicates the order of the group and i_j indicates that the $(i_j)^{th}$ visual word is correlated to this group. This visual words group can be named as a n-order joint term. In this thesis, we discuss the 2-order correlated visual words $ph(i_1, i_2)$ as an example.

As shown in the previous section, the appearance of pair-wise visual words can be obtained from the co-occurrence matrix. The probability of the joint term ph can be approximated according to the proportion:

$$p(ph_{i_1,i_2}) \approx \frac{\boldsymbol{C}(i_1, i_2)}{2 * sum(\boldsymbol{C})} \tag{3.17}$$

where sum denotes the computation which sums up all the elements of a matrix. Actually, if we define the normalized co-occurrence matrix as follows:

$$\boldsymbol{C}_{norm} := \frac{\boldsymbol{C}}{sum(\boldsymbol{C})}, \boldsymbol{f}_{norm} := \frac{\boldsymbol{f}}{sum(\boldsymbol{f})} \tag{3.18}$$

Then, the $p(ph)$ can be approximated by an entry of $0.5 * \boldsymbol{C}_{norm}$. This proportion roughly indicates the significance of the corresponding joint term. When the samples are large enough, the proportion can also be seen as an approximation of probability of its appearance. As a result, it provides a reasonable quantization of the correlation.

The co-occurrence has been largely influenced by the term frequency of visual words. We propose to capture an accurate correlation by the conditional probability. For example, the conditional probability $p(w_{i_1}|w_{i_2})$ can be approximated by a generalization process:

$$\begin{aligned} p(w_{i_2}|w_{i_1}) &= \frac{p(ph)}{p(w_{i_1})} \\ &\approx \frac{\boldsymbol{C}_{norm}(i_1,i_2)}{\boldsymbol{f}_{norm}(i_1)} \end{aligned} \tag{3.19}$$

where $\boldsymbol{f}$ can be replaced by term frequency representation of video $\boldsymbol{v}$ to formulate the video level conditional probability:

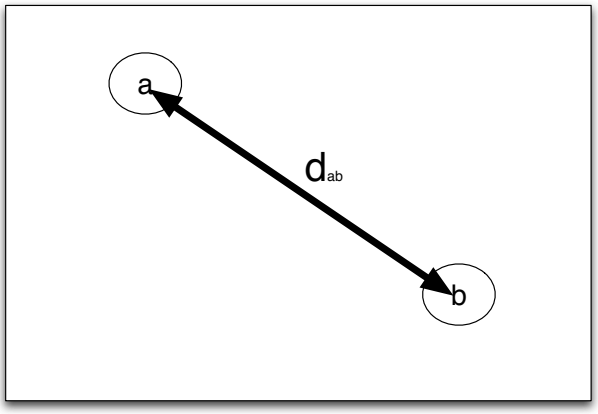


Figure 3.4: Spatial distance between the co-occurring instances a and b of visual words w_i and w_j within a frame

$$p^v(w_{i_2}|w_{i_1}) \approx \frac{\boldsymbol{C}^v_{norm}(i_1, i_2)}{\frac{1}{N}\sum \boldsymbol{f}_{norm}(i_1)} \tag{3.20}$$

If the prior visual words appearance distribution were uniform or the videos collection was large enough, the denominator of Equation 3.20 could be dropped:

$$p^d(w_{i_2}|w_{i_1}) \approx K * \boldsymbol{C}^d_{norm}(i_1, i_2) \tag{3.21}$$

where K denotes the number of visual words.

Following this, we assume that the co-occurring correlation is proportional to the computed conditional probability $p(w_{i_2}|w_{i_1})$. For each visual word, we construct a column vector to define its co-occurring correlation with all other words. Naturally, a co-occurring correlation matrix would be constructed by aligning the column vectors as:

$$\boldsymbol{Corr} := \begin{pmatrix} p(w_1|w_1) & p(w_1|w_2) & p(w_1|w_3) & \cdots & p(w_1|w_K) \\ p(w_2|w_1) & p(w_2|w_2) & p(w_2|w_3) & \cdots & p(w_2|w_K) \\ p(w_3|w_1) & p(w_3|w_2) & p(w_3|w_3) & \cdots & p(w_3|w_K) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ p(w_K|w_1) & p(w_K|w_2) & p(w_K|w_3) & \cdots & p(w_K|w_K) \end{pmatrix} \tag{3.22}$$

where each entry of the matrix can be p^f, p^v, and p^d, which represents the correlation discovered on different levels respectively. It should be noted that the conditional probability is only an estimation of the correlation based on frequency, and it does not necessarily fulfills a specificy distribution.

In summary, the co-occurrence model can be used to discover some meaningful correlation information. The discovered correlation is defined as the conditional probability of visual word co-appearing with another.

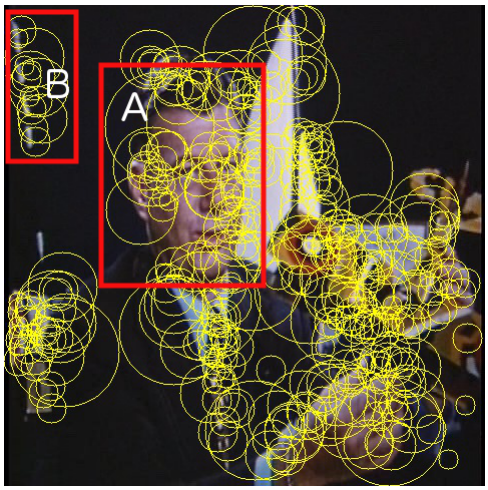


Figure 3.5: The visual features located in A and B respectively are not related to each other.

3.2 Spatial Proximity

As shown in the Equation 3.3, the co-occurring correlation model is constructed based on a trivial assumption: any instance of co-occurring visual words equally determines the correlation between them. This assumption neglects a great deal of information, *i.e.*, the spatial proximity. Actually, the spatial proximity is strongly related to the semantic correlation between the visuals words, for example, visual informations associated with an entity tend to appear closely.

While common visual entity segmentation technology introduces expensive computational cost and may not be reliable, we formulate a proximity based function to discover the spatial information. Based on the discovered spatial information, a finer co-occurring model is constructed. In this section, we present the expanded spatial correlation matrix and discuss how we use it to approximate the semantic correlation.

3.2.1 Spatial Layout of Visual Information

As shown in the Figure 3.4, the co-occurring instances of visual words are located on different positions on the frame. In the previous section, all these co-occurrences are assumed to be equal. A normal frame always contains many different visual objects. Some co-occurring visual words are not truly related to each other, but others are related when they belongs to an identical object layout.

An example is shown in Figure 3.5. Here, a number of features are located in the areas A and B, which are surrounded by two red rectangle respectively. The features in the area A represents the face of the person and those in B represents light spots. Semantically, if two features are located in A and B respectively, they should not be considered as correlated. Inspired by this observation, we can make finer definition of the

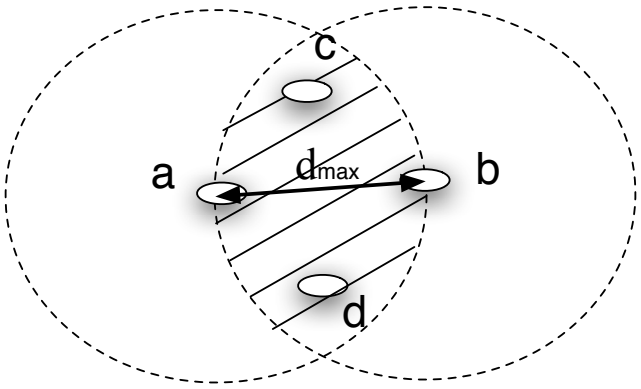


Figure 3.6: Visual objects larger than the shadow area are likely to cover all the instances

co-occurrence, which is aiming to more precisely model the assumption that only instances of visual words representing identical visual objects are correlated with each other. The assumption is defined as follows:

$$c^s_{a,b,c,\cdots} := \begin{cases} 1 & a,b,c,\cdots \in O_i \\ 0 & otherwise \end{cases} \tag{3.23}$$

where O_i is a set of instances of visual words representing a visual object. To get the defined correlation between visual words, all visual entities must be firstly precisely extracted. As a result, the modelling of true co-occurrence is not always practical because of the huge computational expense of the reliable object extraction methods. We can only approximate the true correlation with the spatial proximity information.

To achieve this objective, the hard decided co-occurrence is changed to soft weighting according to the probability that these instances belongs to an identical visual object. This scheme is here defined as spatial co-occurrence. The definition is adapted to:

$$c^s_{a,b,c,\cdots} := p(a,b,c,\cdots \in O_i) \tag{3.24}$$

This probability is obviously unknown to the system, and we can only make an estimation based on additional informations such as shape and scale of the visual object. But these informations are generally also unknown. However, we can estimate this probability according to the area of the object. We can assume that if the area of the visual object is larger enough, its layout will more likely cover all the instances. In other words, this group of instances correlate with each other. As shown in Figure 3.6, if the layout area of the visual object associated with instance a is larger than the shadow, it is likely to cover all the instances in the group. This area threshold can be estimated according to the maximum distances between the instances. If we let the maximum distance between the instances be $\max_{a_1,a_2\in\{a,b,c,\cdots\}}(d_{a_1,a_2})$, then the probability of spatial co-occurrence is

formulated as follows:

$$\begin{aligned} p(a,b,c,\cdots \in O_i) &\approx p(a(O_i) > (\tfrac{2\pi}{3} - 1) * (\max_{a_1,a_2 \in \{a,b,c,\cdots\}}(d_{a_1,a_2}))^2) \\ &\approx p(a(O_i) > (\max_{a_1,a_2 \in \{a,b,c,\cdots\}}(d_{a_1,a_2}))^2) \end{aligned} \tag{3.25}$$

where $a(O)$ denotes the area of the object O_i, and d_{a_1,a_2} denotes the Euclidean Distance between a pair of instances. When the area of the visual object is totally unknown, the probability of spatial co-occurrence should be approximately proportional to the inverse of the maximum distance's square between the instances group. This assumption could be formulated as follows:

$$c^s_{a,b,c,\cdots} \propto \frac{1}{\max(d_{a_1,a_2})^2} \tag{3.26}$$

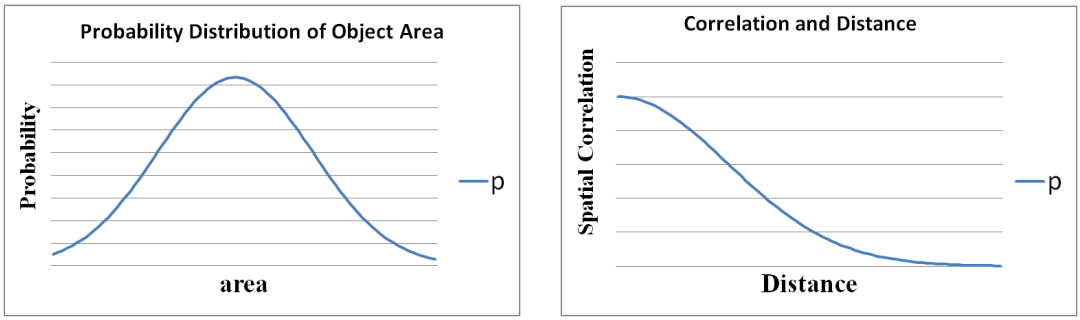
This assumption means that the more closely the instances appear, the greater is the probability of spatial co-occurrence between them. This assumption matches well with human intuition, and the proximity has been shown to be utilized to discover useful information(Zhang, Marszalek, Lazebnik & Schmid 2006, Liu & Chen 2009).

This is all due to the fact that the distribution of visual object area is often not uniform. In practice, important visual objects in a video, which is easily perceived by normal human, can not be too large or too small in a fixed size frame. Intuitively, the shape of distribution of the visual object is generally similar to the curve in Figure 3.7(a).

Based on Equation 3.25, the spatial co-occurrence equals the cumulative probability that an area is larger than d^2. Then, the spatial co-occurrence function should monotonically decrease with the variable d^2, although it should not be linearly proportional to d^2. A reasonable shape of co-occurrence function should look like the shape shown in Figure 3.7(b). A typical function which matches this shape is similar to Gaussian function. Inspired by the previous research (Liu & Chen 2009, Carson et al. 2002a) which found that the visual objects' spatial layout can be modeled by a Gaussian Mixture Model, the spatial co-occurrence between a group of instances in a frame can be quantized in a similar form:

$$c^s_{a,b,c,\cdots} \propto e^{-\kappa \max(d_{a_1,a_2})^2} \tag{3.27}$$

where κ is a parameter to control the width of the function. Theoretically, κ should be inversely proportional to the expectation of the area of the visual object. If the object appears in the frames which are often small, then κ should be larger and fewer instances are assumed to be spatially correlated. Otherwise, κ should be smaller and more instances are correlated.



(a) The distribution of object area (b) Spatial correlation to distance

Figure 3.7: The spatial co-occurrence is cumulative probability of the visual object area.

The pair-wise instances spatial correlation should be formulated as:

$$c^s_{a,b} \approx e^{-\kappa d^2_{a,b}} \tag{3.28}$$

In summary, we propose to utilize the proximity between the instances to refine the co-occurrence between visual words, and the proximity is quantized according to physical distance between the instances. In this thesis, Euclidean Distance is utilized as an example, and other distance metrics can also be used.

3.2.2 Spatial Correlation Matrix

According to the frame level co-occurrence defined in Equation 3.3, we calculate the spatial level visual words co-occurrence based on the spatial co-occurrence modeled in the previous section. The function is formulated as follows:

$$c^s_{i,j} = \begin{cases} \sum_{a=1}^{\boldsymbol{f}_q(w_i)} \sum_{b=1}^{\boldsymbol{f}_q(w_j)} e^{-\kappa d^2_{a,b}} & \boldsymbol{f}_q(w_i) \& \boldsymbol{f}_q(w_j) \neq 0 \\ 0 & otherwise \end{cases} \tag{3.29}$$

where $d_{a,b} = \sqrt{(x_a - x_b)^2 + (y_a - y_b)^2}$ utilize the Euclidean Distance. The spatial vector of i^{th} visual word in the frame is formulated as:

$$\boldsymbol{c}^s_i = \begin{pmatrix} c^s(i,1) & c^s(i,2) & c^s(i,3) & \cdots & c^s(i,K) \end{pmatrix} \tag{3.30}$$

and the spatial matrix of the frame is constructed as:

$$\boldsymbol{S}^f = \begin{pmatrix} \boldsymbol{c}^s_1 \\ \boldsymbol{c}^s_2 \\ \vdots \\ \boldsymbol{c}^s_K \end{pmatrix} \tag{3.31}$$

where each entry of the $\boldsymbol{S}^f$ is calculated using Equation 3.29. Similarly, video level and videos collection level spatial matrix can be accumulated as:

$$\boldsymbol{S}^v = \sum_{\boldsymbol{f} \in v} \boldsymbol{S}^f \tag{3.32}$$

and

$$\boldsymbol{S}^d = \sum_{\boldsymbol{f} \in d} \boldsymbol{S}^f \tag{3.33}$$

where v denotes the set of frames in the video and d denotes the frame set constructing the videos collection. The spatial matrix adapts each co-occurrence according to the spatial proximity, and then quantitatively represents term correlation. Similar to the idea that the visual words correlation is approximated by the probability of the appearing group in the previous section, here we construct the normalized spatial matrix to compute the joint visual words probability:

$$\boldsymbol{S}_{norm} = \frac{\boldsymbol{S}}{sum(\boldsymbol{S})} \tag{3.34}$$

where each entry of $\boldsymbol{S}_{norm}$ is an adjusted joint probability $p(w_i, w_j)$. Furthermore, the conditional probability of $p(w_j|w_i)$ can be computed accordingly by adapting Equations 3.19, 3.20 and 3.21.

$$p_s^f(w_j|w_i) \approx \frac{\boldsymbol{S}_{norm}^f(i,j)}{\boldsymbol{f}_{norm}(i)} \tag{3.35}$$

and

$$p_s^v(w_j|w_i) \approx \frac{\boldsymbol{S}_{norm}^v(i,j)}{\frac{1}{N}\sum \boldsymbol{f}_{norm}(i)} \tag{3.36}$$

and

$$p_s^d(w_j|w_i) \approx K * \boldsymbol{S}_{norm}^d(i,j) \tag{3.37}$$

Finally, a $K \times K$ spatial correlation matrix is constructed by assigning corresponding conditional probability to elements:

$$\boldsymbol{Corr}_s := \begin{pmatrix} p_s(w_1|w_1) & p_s(w_1|w_2) & p_s(w_1|w_3) & \cdots & p_(w_1|w_K) \\ p_s(w_2|w_1) & p_s(w_2|w_2) & p_s(w_2|w_3) & \cdots & p_s(w_2|w_K) \\ p_s(w_3|w_1) & p_s(w_3|w_2) & p_s(w_3|w_3) & \cdots & p_s(w_3|w_K) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ p_s(w_K|w_1) & p_s(w_K|w_2) & p_s(w_K|w_3) & \cdots & p_s(w_K|w_K) \end{pmatrix} \tag{3.38}$$

where each p_s can be the value computed by either p_s^f, p_s^v, or p_s^d, which construct the spatial correlation matrices on different levels.

In this section, we formulate the estimation function for visual word correlation based spatial proximity information, which is defined as spatial correlation. The discovery of temporal information for correlation estimation will be demonstrated in the next section.

3.3 Temporal Correlation

The temporal information is another important characteristic of the video content, and it makes the videos different from still images. As discussed in the literature review chapter, the discovery of temporal information will help to understand the content of video. We are aiming to quantitatively model the temporal information to approximate a finer visual words correlation.

The temporal information utilized in this thesis is described as the motion of the visual words. If an instance of a visual word can be tracked between the continuous frames, the instance will look as if it is moving from an older position to a new position, which is described by the term "motion". Without geometric distortions or changes of viewpoint, the instances representing identical visual objects tends to move coherently. The assumption will help us to refine the co-occurrence model, and better approximate the visual words correlation.

3.3.1 The Temporal Motion of Visual Word

As presented in previous chapters, a video can be represented as an ordered sequence of frames. The order of frames, which contains temporal information, often plays a key role in visual entity representation. Firstly, the sequenced frames representation of a video is formulated as follows:

$$\boldsymbol{v} := \{\boldsymbol{f}_1, \boldsymbol{f}_2, \boldsymbol{f}_3, \cdots, \boldsymbol{f}_n, \boldsymbol{f}_{n+1}, \cdots, \boldsymbol{f}_N\} \tag{3.39}$$

where N is the total number of the frames, and $\boldsymbol{f}_n$ and $\boldsymbol{f}_{n+1}$ denotes the continuous frames. Each $\boldsymbol{f}$ contains a number of instances of visual words. In this thesis, the temporal motion is defined between the continuous frames. As shown in Figure 3.8, in between the continuous frames the instances look like they are moving from one position to another. If we capture this movement, we can utilize the motion to discover the temporal information hidden in the video content.

Let us assume that the physical layout of the frame as an Euclidean Space, and the position of an instance a on frame $\boldsymbol{f}_n$ is identified by an ordered pair (x_a, y_a). The visual word, to which the instance a is mapped, is denoted by w_a.

We need to track the position of the instance in the next frame to capture its temporal motion. The neighboring key frames are always very similar to each other, and we utilize a simple scheme to track the visual words. The corresponding visual feature in the next

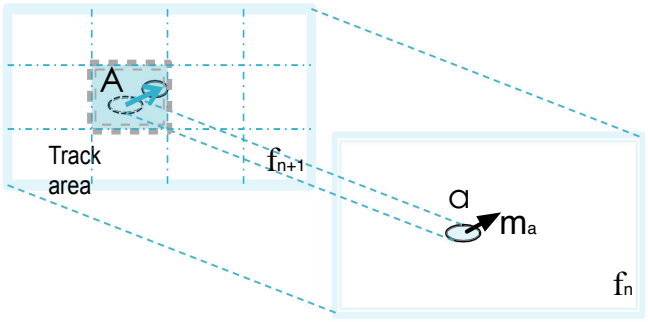


Figure 3.8: The captured motion of instance between the continuous frames.

frame should also be mapped to visual word w_a. The target instance should be one element of the set:

$$\boldsymbol{z}_a = \{z | t_z = w_a\}. \tag{3.40}$$

where t_z denotes the visual word instance z is mapped to. The number of the instances set $\boldsymbol{z}_a$ could be identified by $\boldsymbol{f}_{n+1}(w_a)$.

In this section, the nearest instance of visual word $w_a$to the older position (x_a, y_a) is assumed to be the tracked a' corresponding to L^2 Norm:

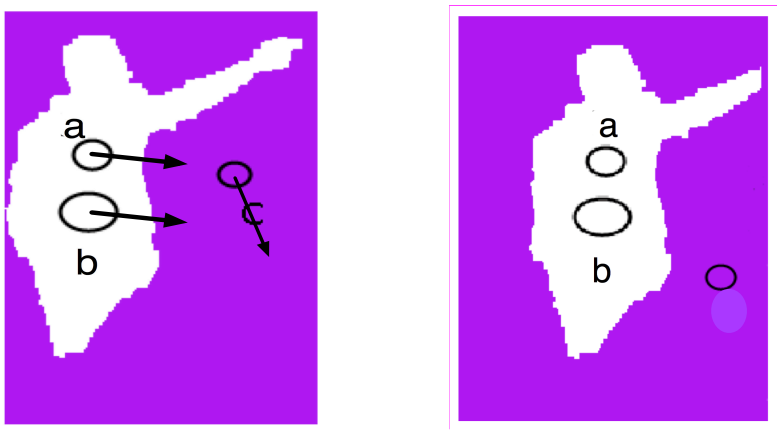
$$a' := \arg\min_{z \in \boldsymbol{z}_a} d(z, a) \tag{3.41}$$

where $d(z, a)$ is the distance between the older position (x_a, y_a) and newer position(x_z, y_z), for example, the Euclidean Distance $\sqrt{(x_z - x_a)^2 + (y_z - y_a)^2}$. Other distance metrics can also be used here. To avoid the falsely tracked instance, we set a tracking area A to track a. While the neighboring frames should not be very different, and the movement of instance normally does not exceed a limit. The area of A is empirically determined by the key frame sampling ratio. A smaller tracking area A should be utilized for a higher sampling ratio, which means that more key frames are sampled. Otherwise, a bigger tracking area A should be used to relax the tracking limit. In this way, the final tracking function is formulated as:

$$a' := \begin{cases} \arg\max_{z \in \boldsymbol{z}_a} d(z, a) & (x_z, y_z) \quad in \quad A \\ Null & otherwise \end{cases} \tag{3.42}$$

If the instances can not be tracked, the inconsistent instances will be assumed as noise and reduced from the frame representation, which is similar to the method proposed by Sivic & Zisserman (Sivic & Zisserman 2006) to select the visual words surviving at least 3 key frames.

Afterwards, each instance is associated with a tracked instance at the next frame. It is assumed to move from (x_a, y_a) to $(x_{a'}, y_{a'})$, and a motion vector is contracted to describe



(a) Key Frame 1 (b) Key Frame 2

Figure 3.9: The relative motion can be a clue of the visual object.

its motion as follows:

$$\begin{aligned} \boldsymbol{m}_a : &= \begin{pmatrix} x_{a'} - x_a & y_{a'} - y_a \end{pmatrix} \\ &= \begin{pmatrix} \Delta x_a & \Delta y_a \end{pmatrix} \end{aligned} \tag{3.43}$$

where $\boldsymbol{m}_a$ is a 2-order vector.

3.3.2 Relative Motion Modeling and Temporal Correlation

As discussed in the previous section, the motion of each instance is captured by its associated motion vector. According to the previous definition of instance correlation, the correlated instances should tend to describe the identical object. The temporal information, especially the motion associated with the instances, could be utilized to address this problem, and the instances representing identical visual objects tend to move in a coherent manner.

An example is demonstrated in Figure 3.9, where a person moves from the old position in frame 1 shown in Figure 3.9(a) to the newer position shown in Figure 3.9(b). Although the position has a little visual difference, the visual person keeps its shape and structure. As a result, the relative positions of instances describing the person tend to be temporally consistent. These instances tend to move in a coherent manner, *e.g.*, instances a and b associated with the person move to the same direction in Figure 3.9(a), but the other instance outside the person tends to move in a different direction.

As the motion vector captures the motion of instance in continuous frames, the relative motion between pair-wise instances can be captured by the relative motion vector. As illustrated in Figure 3.10, instances a and b are assigned with the modelled motion vectors $\boldsymbol{m}_a$ and $\boldsymbol{m}_b$. Based on these two vectors, the relative motion vector can be formulated as:

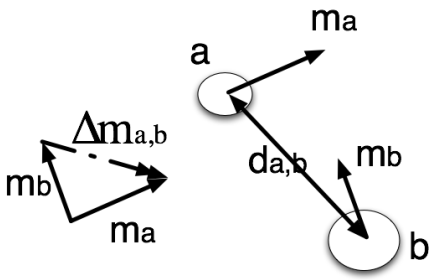
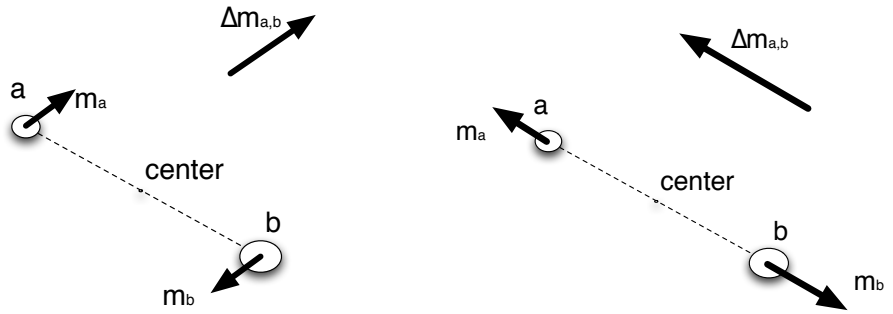


Figure 3.10: Relative Motion between a pair of instances of visual word



(a) Rotation (b) Scaling

Figure 3.11: The relative motion can be caused by the affine transformation of visual object.

$$\begin{aligned} \Delta \boldsymbol{m}_{a,b} &= \boldsymbol{m}_a - \boldsymbol{m}_b \\ &= \begin{pmatrix} \Delta x_a - \Delta x_b & \Delta y_a - \Delta y_b \end{pmatrix} \end{aligned} \tag{3.44}$$

The relative motion vector simultaneously describes the direction and scale difference between two motion vectors. The L^2-Norm of Δm quantizes the motion difference, and it is inversely proportional to the motion coherence between the instances.

According to the above discussion, the motion coherence of the instances is assumed to be a clue regarding the connection to an identical visual object, and co-occurring instances representing identical visual objects have been defined to approximate the semantic correlation. The relative motion vector could be used to construct a finer quantitative co-occurrence model to approximate the semantic correlation.

However, related instances may also move relatively in the neighboring frames. For example, although a few visual words represent identical visual object, a number of affine transformations of the visual object would lead to relative motion between the visual words.

For example, the rotation and scaling of object would result in the relative motion between visual words, which is shown in Figure 3.11. This relative motion, caused by affine

transformation of visual object, can be named as positive relative motion, and otherwise negative relative motion in this section. It is hard to distinguish between the positive and negative relative motion without additional information, for example, the center of the visual object. The probability of pair-wise instances with a certain degree of relative motion being correlated to identical visual objects could only be roughly estimated.

We assume that the prior probability distribution of visual object O and instance pair $\{a, b\}$ are both uniform, and the maximum likelihood function and posterior distribution should be in a similar form:

$$p(\{a, b\} \in O|\Delta \boldsymbol{m}_{a,b}) = k * p(\Delta \boldsymbol{m}_{a,b}|\{a, b\} \in O) \tag{3.45}$$

Because the motion are tracked in between neighboring and very similar frames, other projective transformations are not considered in this temporal information discovery model. The probability of visual object O generating two instances $\{a, b\}$ with a relative motion $\Delta m_{a,b}$ is determined by two independent factors: the degree of affine transformation and the average distance $\bar{d}^c_{a,b}$ between object center and the the two instances. The larger the affine transformation or $d^c_{a,b}$ is, the larger is the generative probability. In other words, the degree of freedom here is two.

We make an assumption that the probability of $\{a, b\}$ being correlated to each other is inversely proportional to square of L^2 Norm of the relative motion vector:

$$p(\{a, b\} \in O|\Delta \boldsymbol{m}_{a,b}) \propto \frac{1}{\|\Delta \boldsymbol{m}_{a,b}\|^2} \tag{3.46}$$

where $\|\boldsymbol{m}\|$ represents the L^2 Norm of the vector $\boldsymbol{m}$. Note that other norm metrics can also be used in this arrangement.

Thus, the probability $p(\Delta \boldsymbol{m}_{a,b}|\{a, b\} \in O)$ is assumed to be in an accumulation of that the $\bar{d}^c_{a,b}$ and affine transformation is large enough, which is like this:

$$\begin{array}{rcl} p(\Delta \boldsymbol{m}_{a,b}|\{a, b\} \in O) & = & \int p(\Delta \boldsymbol{m}_{a,b}|\bar{d}^c_{a,b})p(\bar{d}^c_{a,b}) + \int p(\Delta \boldsymbol{m}_{a,b}|f(O))p(f(O)) \\ & & where \quad \bar{d}^c_{a,b} > \epsilon_d(\Delta \boldsymbol{m}_{a,b}) \quad and \quad f(O) > \epsilon_f(\Delta \boldsymbol{m}_{a,b}) \end{array} \tag{3.47}$$

where $f(O)$ denotes the affine transformation of the visual object, and ϵ_d and ϵ_d denotes the thresholds of the average distance and affine transformation, which should be determined by the relative motion.

If the prior distribution of $f(O)$ and $\bar{d}^c_{a,b}$ is more likely not to be very small and big, then function of temporal correlation of instances and the relative motion could be roughly depicted as Figure 3.12.

Note that this function is similar to the function in Figure 3.7, and thus we utilize a similar form function with the spatial co-occurrence to quantize the temporal information

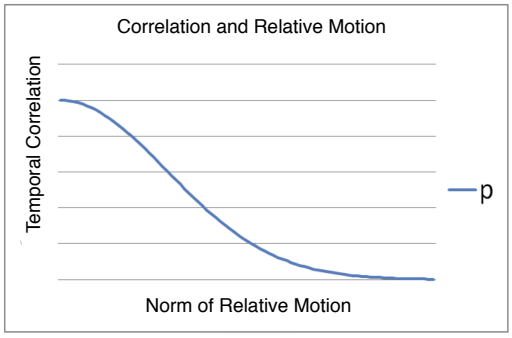


Figure 3.12: The shape of temporal correlation and relative motion function

and in turn enhance the co-occurrence of instances, known as temporal co-occurrence. The formulated function is as follows:

$$c^s_{a,b} \approx k * e^{-\gamma \|\Delta m_{a,b}\|^2} \tag{3.48}$$

where parameter k indicates the scale coefficient of temporal correlation, and parameter γ controls the decreasing rate of temporal correlation along with the expansion of the relative motion. The γ is normally determined by the sampling ratio, and a higher sampling ratio normally leads to a larger γ, which means fewer instances are assumed to be correlated to each other.

In summary, the co-occurrence between the instances is quantized according to the norm of the relative motion vector constructed. In the following section, the temporal correlation between visual words will be approximated based on the temporal co-occurrence quantized.

3.3.3 Temporal Correlation Matrix

The original frame level co-occurrence is defined by Equation 3.3. We replace the normal co-occurrence between a pair of instances with the temporal co-occurrence based on the relative motion, and the frame level visual words temporal co-occurrence is modeled by the quantization function Equation 3.48. The function is formulated as follows:

$$c^t_{i,j} = \begin{cases} \sum_{a=1}^{\boldsymbol{f}_q(w_i)} \sum_{b=1}^{\boldsymbol{f}_q(w_j)} e^{-\gamma \|\Delta m_{a,b}\|^2} & \boldsymbol{f}_q(w_i) \& \boldsymbol{f}_q(w_j) \neq 0 \\ 0 & otherwise \end{cases} \tag{3.49}$$

where $m_{a,b}$ utilizes the relative motion quantized in the previous section as an example, and $c^t_{i,j}$ identifies that it is quantized for i^{th} and j^{th} visual words. The temporal vector of i^{th} visual word in the frame is formulated by assigning all the possible temporal co-occurrences to the corresponding position to form a vector:

$$\boldsymbol{c}_i^t = \begin{pmatrix} c^t(i,1) & c^t(i,2) & c^t(i,3) & \cdots & c^t(i,K) \end{pmatrix} \tag{3.50}$$

Based on this arrangement, a temporal matrix of a frame is defined as follows:

$$\boldsymbol{T}^f = \begin{pmatrix} \boldsymbol{c}_1^t \\ \boldsymbol{c}_2^t \\ \vdots \\ \boldsymbol{c}_K^t \end{pmatrix} \tag{3.51}$$

where each entry of the $\boldsymbol{T}^f$ is calculated by Equation 3.49. Similar to the video level and videos collection level co-occurrence matrix and **spatial matrix**, the temporal matrix can be formulated by an accumulative function:

$$\boldsymbol{T}^v = \sum_{\boldsymbol{f} \in v} \boldsymbol{T}^f \tag{3.52}$$

and

$$\boldsymbol{T}^d = \sum_{\boldsymbol{f} \in d} \boldsymbol{T}^f \tag{3.53}$$

where v denotes the set of frames representing the video and d denotes the frame set composing the videos collection representation. The temporal matrix modifies each co-occurrence according to the temporal motion coherence, and it can be seen as representing the temporal coherence based term correlation. Similar to the definition made that the term correlation is approximated by the probability, we normalize the *temporalmatrix* to compute the probability of a jointed pair-wise visual words group:

$$\boldsymbol{T}_{norm} = \frac{\boldsymbol{T}}{sum(\boldsymbol{T})} \tag{3.54}$$

where each entry of $\boldsymbol{T}_{norm}$ is an estimated probability $p(w_i, w_j)$ modified by a temporal coherence constraint. The modified conditional probability of $p(w_j|w_i)$ can be computed accordingly by adjusting Equations 3.19, 3.20 and 3.21.

$$p_t^f(w_j|w_i) \approx \frac{\boldsymbol{T}_{norm}^f(i,j)}{\boldsymbol{f}_{norm}(i)} \tag{3.55}$$

and

$$p_t^v(w_j|w_i) \approx \frac{\boldsymbol{T}_{norm}^v(i,j)}{\frac{1}{N}\sum \boldsymbol{f}_{norm}(i)} \tag{3.56}$$

and

$$p_t^d(w_j|w_i) \approx K * \boldsymbol{T}_{norm}^d(i,j) \tag{3.57}$$

where the distribution of visual words term frequency in the videos collection is also assumed to be uniform.

Finally, a $K \times K$ **temporal correlation matrix** is constructed by locating the conditional probability to the corresponding position:

$$\boldsymbol{Corr}_t := \begin{pmatrix} p_t(w_1|w_1) & p_t(w_1|w_2) & p_t(w_1|w_3) & \cdots & p_t(w_1|w_K) \\ p_t(w_2|w_1) & p_t(w_2|w_2) & p_t(w_2|w_3) & \cdots & p_t(w_2|w_K) \\ p_t(w_3|w_1) & p_t(w_3|w_2) & p_t(w_3|w_3) & \cdots & p_t(w_3|w_K) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ p_t(w_K|w_1) & p_t(w_K|w_2) & p_t(w_K|w_3) & \cdots & p_t(w_K|w_K) \end{pmatrix} \tag{3.58}$$

where each p_t can be either p_t^f, p_t^v, or p_t^d, with which the temporal correlation matrices on different levels are constructed respectively.

In this section, we formulate the estimation function for visual word correlation with temporal motion coherence information, which is defined as temporal correlation. The discussion of the combination of spatial and temporal correlation will be presented in the next section.

3.4 Spatial-Temporal Correlation and Discussion

3.4.1 Spatial-Temporal Correlation

In Sections 3.1, 3.2, and 3.3, the co-occurring correlation, spatial correlation, and temporal correlation are formulated respectively. The correlation function is actually in a similar form as that shown by Equations 3.22, 3.38, and 3.58. The only difference is that the probability estimation methods are based on different information.

Both Equation 3.38 and 3.58 expand the co-occurring correlation matrix with additional spatial and temporal information respectively. A natural idea is to explore the possibility of expanding the co-occurring relation with combined spatial and temporal information to formulate the spatial-temporal correlation between the visual words.

The first possibility is to directly combine the two constraints, i.e., spatial proximity and temporal motion coherence to a spatial-temporal constraint to modify the weights of each co-occurring instances group. However, the spatial proximity and relative motion are in different dimensions. The unit of spatial proximity is "pixels" and relative motion is "pixels per key frames". It is not appropriate to directly combine the two quantization constraints.

A solution is to utilize the similar formulation of Equations 3.38, and 3.58. The value of elements of spatial and temporal correlation matrices is given by $p_s(w_i|w_j)$ and $p_t(w_i|w_j)$, which represents a conditional probability between two visual words estimated based on different information. We could fuse the two estimated probabilities, rather than fusing the information. The fusion function of probability is assumed to be a linear

addition, and the function is formulated as follows:

$$p_{st}(w_i|w_j) = \frac{p_s(w_i|w_j) + k * p_t(w_i|w_j)}{1+k} \tag{3.59}$$

where k denotes the parameter to balance the spatial and temporal constraints in the combined probability.

The fused probability is used to approximate the semantic correlation, defined as Spatial-Temporal Correlation (STC). The correlation matrices within a frame, a video and a videos collection are calculated by fusing the spatial matrix and temporal matrix:

$$\boldsymbol{Corr}_{st} = \frac{1}{1+k_{st}} \times (\boldsymbol{Corr}_s + k_{st} * \boldsymbol{Corr}_t) \tag{3.60}$$

where, $\boldsymbol{Corr}_{st}$ is a $K \times K$ matrix and each entry of it is computed by Equation 3.59 for the corresponding pair of visual words. The balancing parameter k_{st} is used to adjust the relative weights of discovered temporal and spatial information. A larger k_{st} means that the temporal information plays a bigger role. In this thesis, the value of k_{st} is empirically selected.

In summary, an STC matrix $Corr$ is generated based on information which is maintained in spatially or temporally co-occurring visual words. The way in which the STC is incorporated to improve the BovW framework based information retrieval model will be described in the next chapters.

3.4.2 Discussions on the Correlations

We have formulated the (spatial/temporal) co-occurrence matrix and the derived (co-occurring/spatial/temporal) correlation matrix. This section is aiming to discuss potential utilization of the matrices in the information retrieval model.

The co-occurrence matrix, spatial matrix, and temporal matrix are all symmetric matrices, which describe the information of the two co-occurring visual words, which can be seen as a joint term in the model:

$$\begin{aligned} \boldsymbol{C} &= \boldsymbol{C}^T \\ \boldsymbol{S} &= \boldsymbol{S}^T \\ \boldsymbol{T} &= \boldsymbol{T}^T \end{aligned} \tag{3.61}$$

where $\boldsymbol{M}^T$ represents the transpose matrix of $\boldsymbol{M}$. As shown in the Equation 3.61, $c(i,j)$ always equals $c(j,i)$, and it means that the order of w_i and w_j in the joint term has not been taken into consideration. If $c(i,j)$ is larger than $c(k,m)$, the joint term $\{w_i, w_j\}$ (spatial/temporal adjusted) co-occurs more frequently than $\{w_k, w_m\}$ within the corresponding videos collection.

Unlike the co-occurrence matrix, the correlation matrix formulated is not symmetric.

Each entry describes the degree to which a visual word depends on another visual word as follows:

$$\begin{array}{rcl} \boldsymbol{Corr} & \neq & \boldsymbol{Corr}^T \\ \boldsymbol{Corr}(i,j) & \approx & \frac{p(w_i,w_j)}{p(w_j)} \neq \boldsymbol{Corr}(j,i) \end{array} \tag{3.62}$$

If we defined the k^{th} column vector of a matrix $\boldsymbol{M}$ as $\boldsymbol{M}(i_k)$ and n^{th} row vector of $\boldsymbol{M}$ as $\boldsymbol{M}(j_n)$, the relationship between the co-occurrence matrix and correlation matrix can be summarized as follows:

$$\boldsymbol{Corr}(i_k) = K_k \times \boldsymbol{C}(j_k) \tag{3.63}$$

because $\boldsymbol{C}$ is a symmetric matrix, the relation function can be re-formulated as:

$$\boldsymbol{Corr}(i_k) = K_k \times \boldsymbol{C}(i_k) \tag{3.64}$$

where K_k is a term frequency coefficient associated with k^{th} visual word, which has been discussed in previous sections.

Theoretically, according to Equation 3.64, each column vector $\boldsymbol{Corr}(i_k)$ represents the correlation of all other visual words with visual word w_k, and each row vector $\boldsymbol{Corr}(j_k)$ describes the correlation degree of w_k with all other words.

The column vector can also be seen as the quantized context of the visual word w_k, because it describes how other visual words depend on w_k. A row vector describes the degree to which instances of w_k are related to surrounding instances.

The above characteristics of the correlation matrix motivates our methods to improve the BovW framework for the CBVR model, and our methods will be discussed in the next chapters. In summary, the correlation matrix quantitatively models the spatial-temporal relation existing between a pair of visual words. Noted that the correlation matrix can be easily expanded to a higher order tensor to describe the correlation between more visual words.

3.5 Summary

We have proposed a quantitative analysis framework to discover the statistical co-occurrence between the visual words. The co-occurrence is modeled by a constructed co-occurrence matrix. Furthermore, the co-occurrence can be refined according to the spatial and temporal constraints, which are based on the information discovered between them.

The spatial-temporal constraints aim to approximate the probability that co-occurring visual words describe an identical visual object. To achieve this objective, we have made a couple of assumptions based on the spatial and temporal relationship between the visual word instances. Firstly, visual word instances which appear spatially closely to each other are more likely to describe an identical visual object. Secondly, the instances which move coherently on continuous frames of video are more likely to truly co-occur.

The spatial and temporal co-occurrence is quantized according to the physical distance and relative motion. Considering a couple of practical factors, the models of spatial and temporal co-occurrence both utilize a Gaussian-like function. The co-occurrence matrix is computed by re-weighting the each co-occurrence based on the quantized spatial-temporal information.

The co-occurrence matrix is refined to construct a spatial/temporal correlation matrix. The entry of the correlation matrix represents the degree of correlation between corresponding visual words, and it can also be seen the context of the visual word.

We also attempted to combine the spatial and temporal constraints and fuse the formulated spatial and temporal correlation matrix. The concept of STC is defined to summarize the modeled correlations.

Chapter 4

STC-based Representation Reformulation

In Chapter 3, the spatial-temporal correlation has been quantitatively modelled. In this Chapter, the STC extracted from the video will be firstly utilized to distinguish the more descriptive visual words in the visual content. The descriptive visual words, which are defined as Words-of-Interest, are assumed to be strongly correlated with each other. This assumption is established based on the intuition that meaningful visual content tends to be collaboratively represented by several correlated visual words.

We expect that emphasizing the descriptive visual words will improve the retrieval model. The proposed approach aims at assigning higher weights to the more descriptive visual words, based on the STC incurred by the query video. We define the modified visual words weighting scheme as Query Correlation (QC). The utilization of QC in the retrieval model would reformulate the query representation, which is equivalent to enhancing the similarity measurement with additional spatial-temporal information discovered from the query.

In addition to the STC discovered from query video level, the STC discovered from the whole video collection is also assumed to be related to the descriptive ability of visual words. Similar to Inverse Document Frequency, we assume that visual words co-occurring with others in less number of videos within the collection are more descriptive. Based on this idea, we define an Inverse Document STC (IDC) as another visual words weighting scheme. The IDC weighting scheme will also be used to reformulate the video representation for the retrieval model.

It should be pointed out that, in this section, we focus primarily on evaluating the approach in Query-by-Example (QBE) retrieval (Weng, Li, Cai, Zhang, Zhou, Yang & Zhang 2011). A series of experiments are performed against two widely used video collections which are publically available for the video retrieval research community. Nevertheless, the proposed approach can be applied to many other practical tasks such as

near-duplicate video search, copyright infringement detection, instance search, etc.

4.1 Words-of-Interest

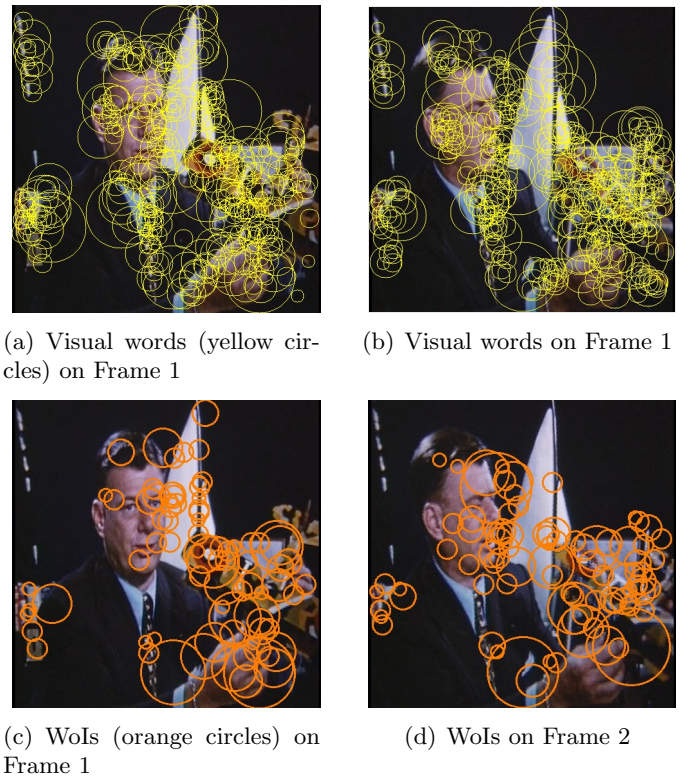
In videos, the information of interesting to the users, is constantly mixed up with redundant information. One idea is to utilized the spatial and temporal information to identify and emphasise the interesting information. In this section, we aim to established a couple of hypothesis for selecting descriptive visual words representing the information. Effectiveness of the hypothesiss and selection schemes for video retrieval is evaluated with some preliminary experiments.

4.1.1 Words-of-Interest Selection based on Spatial Proximity

As discussed in Chapter 2, a drawback of the BoW framework is that it ignores inter-word relationships and assumes that visual words are independent. This framework always characterize the descriptive ability of the visual words only according to the appearing frequency. However, there is a problem with this technology. When a large scale visual vocabulary is utilized, most visual words only appear a few times. The descriptive ability of visual word can not be effectively distinguished by term frequency. In this section, we refer to the descriptive visual word that are of user's interest, as Words-of-Interest (WoI), and the two terms descriptive visual words and WoI are used interchangeably. The mixture of WoI and other words would lead to irrelevant results for CBVR technology.

To address this problem, various supervised or unsupervised learning technologies are proposed to indicate the interesting visual content within the video. For example, Liu & Chen (Liu & Chen 2009) proposed a method to represent a video by extracting regional characteristics, namely Object-of-Interests (OoI). However, the offline OoI extraction may exclude some relevant information, because the interests of users are very hard to determine prior to an online search. Zhang et al. (Zhang, Tian, Hua, Huang & Li 2009) proposed to select Descriptive Visual Words (DVP) through supervised training to improve the performance of image retrieval and object recognition. These methods are proposed based on the spatial co-occurring information. We aim to construct an unsupervised discovery method to discover the words-of-interest for the CBVR model.

In this section, we propose a novel approach based on the selected WoI according to the spatial proximity imposed by a given video. The WoI selection is based on assumptions that a salient visual word tends to co-occur with and is close to the other important ones. Accordingly, we rank the importance of the visual words and select the WoI without supervised learning and training data. The proposed WoI selection algorithm ranks the visual words based on two criteria: i) the WoIs co-occur more frequently in the video, and ii) the Wols are of a greater spatial proximity with each other. The two criteria can be quantified by the spatial correlation matrix proposed in the last chapter.



(a) Visual words (yellow circles) on Frame 1

(b) Visual words on Frame 1

(c) WoIs (orange circles) on Frame 1

(d) WoIs on Frame 2

Figure 4.1: An example of Word-of-Interest Selection based on the Spatial Proximity.

The spatial proximity expectation of the i^{th} visual word can be formulated based on the quantized spatial correlation:

$$\bar{s}_i = \sum \boldsymbol{corr}^s(i,j) * p(w_j) \tag{4.1}$$

where $p(w_j)$ is the probability that visual word w_j occurs in the frame. The probability can normally be estimated by the normalized term frequency of the visual word w_j. If the number of instances of the visual words in the frame is denoted by N_f, the spatial proximity expectation of w_i can be computed as:

$$\bar{s}_i \approx \frac{1}{N_f} \sum \boldsymbol{corr}(i,j) * \boldsymbol{f}(w_j) \tag{4.2}$$

Furthermore, we define a proximity rank vector for all the visual words:

$$\boldsymbol{R}_w = \{r_i\} \tag{4.3}$$

where $i = 1, 2, 3, \cdots, K$ and K is the size of visual vocabulary, and the initial value of r_i is given by $r_i = \bar{s}_i$.

If a visual word is of greater proximity, the visual words co-occurring closely with it tend to be more descriptive. Accordingly, we propose a recursive ranking algorithm in

Algorithm 4.1, where the visual words ranked on top of $\boldsymbol{R}_w$ are selected as WoI:

Algorithm 4.1 Calculate The Spatial Proximity Ranking

Require: Spatial Correlation Matrix $\boldsymbol{Corr}_s$, Term frequency vector $\boldsymbol{f}$, maxiter(maximum iteration steps)
Ensure: Visual word rank $\boldsymbol{R}_w$
  $\boldsymbol{R}_w^0 = \boldsymbol{r}$
  for k=0 **to** k=maxiter **do**
    $\boldsymbol{R}_w^{k+1} = \boldsymbol{Corr}_s \times \boldsymbol{R}_w^k$;
    Normalize($\boldsymbol{R}_w$);
    if $\sum(\boldsymbol{R}_w^{k+1}(i) - \boldsymbol{R}_w^k(i)) < \epsilon$ **then**
      break;
    end if
    i++;
  end for
  Sort($\boldsymbol{R}_w$);

$$WoI = \{w_j\}, \quad j = id_1, id_2, \cdots, id_{N_{woi}} \tag{4.4}$$

where id_k is the index of k^{th} WoI in the visual vocabulary, and N_{woi} is the number of the selected WoI.

In this way, the representation of a frame is divided into two parts: visual words and WoI. The WoI representation is formulated as:

$$\acute{\boldsymbol{f}} = \Big(\boldsymbol{f}(id_1) \quad \boldsymbol{f}(id_2) \quad \boldsymbol{f}(id_3) \quad \cdots \quad \boldsymbol{f}(id_{N_{woi}})\Big) \tag{4.5}$$

An example of WoI representation of the frame is shown in Figure 4.1. Indeed, a number of instances of visual words, which are close to each other, are selected to represent the visual content.

4.1.2 Word-of-Interest Selection based on Temporal Coherence

In the previous section, WoIs were selected according to the spatial proximity from the visual vocabulary. As discussed in Chapter 3 the temporal correlation is quantized in a similar formulation to that of the spatial correlation matrix. We can adjust the assumption made to select the WoI based on temporal constraint in addition to the spatial information.

Similarly, we formulate a hypothesis that the WoIs would appear and move in a relatively coherent manner across neighboring framesin a video, while non-WoI occur more singularly and randomly. Here, the definition temporal motion coherence is similar to the one in Chapter 3, whereby it is the degree to which a visual word moves coherently with other words on the temporally aligned frames in the video. The visual vocabulary can be ranked according to the average temporal coherence $\bar{t}$ of individual words, which is calculated as:

$$\bar{t}_i = \sum \boldsymbol{corr}^t(i,j) * p(w_j) \tag{4.6}$$

In Section 4.1, the WoI with high spatial proximity can be selected, when the iteratively computed $\bar{s}_i$ is lower than an empirical selected threshold. The temporal correlation matrix can be done in the same form, and a similar computation method can be used to select the temporal coherence based WoI. However, a pre-fixed empirical threshold may not be suitable for temporal coherence based WoI selection. The sizes of frames tend to be uniform in different videos, although the temporal scale varies significantly across different visual content. As a result, the visual words motion in frames covers a wide range of values, and a uniform threshold may not be suitable in this selection.

To tackle this problem, the EM algorithm (Dempster, Laird & Rubin 1977) is used to adaptively classify the visual words into WoI and non-WoI based on temporal motion coherence.

More formally, each visual word is associated with a hidden variable $z \in \{z_+, z_-\}$. Here, z_+ indicates that the visual word is WoI, while z_- indicates that it is not. Naturally, $p(z_+)$ represents the probability of a visual word belonging to WoI. The appearing probability of a visual word with a certain motion coherence is denoted by $p(\bar{t}|z)$. We assume that $p(\bar{t}|z_+)$ and $p(\bar{t}|z_-)$ are both Gaussian distributions. From these definitions, the joint distribution of $p(\bar{t}, z)$ is defined as $p(z)(\bar{t}|z)$, and we simplify the problem by assuming that z and $\bar{t}$ are independent variables. All distributions are unknown as yet, and the parameters should be estimated using the EM algorithm.

The steps of the EM algorithm for estimating the unknown distribution are given as follows:

Algorithm 4.2 EM algorithm for WoI selection

E-step:

$p(z|\bar{t}) = c_1 p(z) p(\bar{t}|z)$;

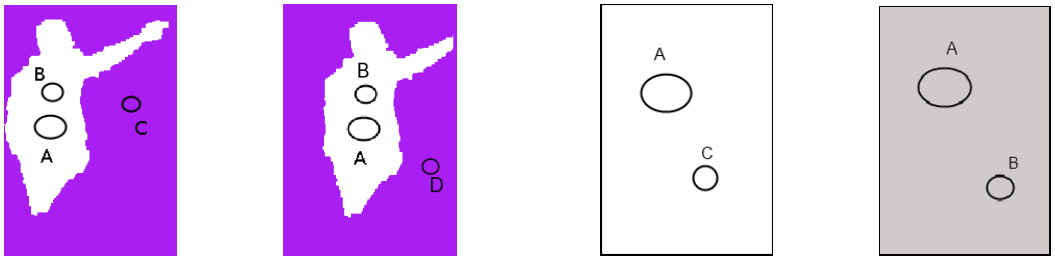
$E_{p(z|\bar{t})}[\log p(\bar{t}|\Phi)] = \prod_j \sum_i p(z_i) p(\bar{t}_j|z_i)$

M-step:

$\Phi_{new} = argmax_\Phi E_{p(z|\bar{t})}[\log p(\bar{t}|\Phi)]$

where Φ is a set of parameters to be estimated, c_1 is the nominalization factor to guarantee that the sum of $p(z|\bar{t})$ equals 1. The estimated $p(\bar{t}|z_+)$ identifies the location of WoI in the temporal motion coherence rank. We choose the visual words located within the standard deviation of the Gaussian distribution $p(\bar{t}|z_+)$ as WoI. After all, the frame level similarity can be measured based on the WoI.

In this section, it is demonstrated how to select the WoI using STC information. However, this selection arbitrarily set binary values to the visual words (to be WoI or not), which may not be precisely enough to distinguish descriptive power of visual words. In the next section, we will discuss how to softly weights the visual words with STC



(a) Frame 1 in the query.

(b) Frame 2 in the query.

(c) Key frame in Video 1

(d) Key frame in Video 2

Figure 4.2: Emphasizing the descriptive visual words to compensate the neglect of spatial-temporal internal structure. Query consists of two frames and is represented by visual words A, B, C and D, whilst the video 2 is likely to be more relevant than 1)

information.

4.2 Query Reformulation

As discussed in Section 4.1, spatial and temporal information can be utilized to address the interesting visual words in video content. However, the emphasizing of them for video retrieval is still arbitrary. This inspires us to further analyse and investigate on how to emphasis the descriptive visual content (i.e. WoI) and whether it may compensate the loss of spatial and temporal information in the retrieval function. The spatial and temporal information incurred by query is obviously a clue of relevance.

Figure 4.2 shows an illustrative example. Assume that the relevant videos should include the person appearing in the query. In Frame 1, visual words A and B, which represent the visual content of a person are messed up by another irrelevant visual word C. The BovW model, on one hand, discards the spatial relationship between A and B within Frame 1, and on the other hand, neglects the motions of A and B from Frame 1 to Frame 2. In the similarity measurement, the BovW model assumes that all visual words have an equal descriptive power. As a result, Video 1 and Video 2 are considered equally similar to the query. However, Video 2 is more likely to be relevant to the query, because the spatial relationship and motion coherence between A and B strongly implies that they belong to an identical object and should be more descriptive than C and D with respect to this query. Thus, the spatial-temporal correlation of visual words A and B is a strong clue of relevance.

Various approaches (described in detail in Chapter 2) have been proposed to incorporate the spatial-temporal constraints associated with visual words. Some recent image/video retrieval methods add the position, scale, main orientation and motion primitives of each visual word directly into the BovW representation, and then, for example, enhance the similarity measurement with Weak/Tight Geometric Constraint (WGC/TGC)

verification of matched visual words between two images/videos (Jégou et al. 2010), (Zhao et al. 2010). Nonetheless, the injected information results in a significant increase of computation cost in the similarity match.

To tackle the aforementioned limitations, instead of expanding or adding extra spatial-temporal information directly into the BovW representation, we propose to identify and emphasize the more descriptive visual words (for example, A and B in Figure 4.2) through exploiting the spatial-temporal correlation among different visual words in the query in an integrated manner. Emphasizing descriptive visual words would revise the query representation and exclude the irrelevant information, thus compensating the neglect of spatial-temporal information. Furthermore, the spatial-temporal information discovery from the query example does not result in extra storage cost for data representation nor increased complexity in similarity measurement.

Inspired by the improvement achieved based on our approach in the previous section, we propose to characterize the descriptive visual words based on spatial proximity and temporal motion coherence incurred by the query. In the consecutive frames of the query, an inherent object often has an explicit spatial structure and intensive spatial relationship. The motion of object layout across neighboring frames in the query often has a characteristic of coherence. This spatial-temporal relationship can be utilized to approximate the possibility that the visual words are associated with an identical object, and such visual words usually have more descriptive power for the query. We base our proposed method on two assumptions regarding the descriptive visual words: i) they co-occur closely in a frame; ii) they move coherently across sequential frames.

In this section, a STC-based similarity measure function is developed for adjusting the visual words weights to namely Query Correlation (QC) with respect to STC. Essentially, this leads to the key frames reformulation for the query video, effectively involving the descriptive visual words that may or may not originally appear in the key frame and excluding the noisy ones. The QC weights of the visual words are determined by both the STC matrices and their frequencies. Furthermore, it is important to note that the retrieval technology can be easily incorporated into standard inverted indexing architecture to achieve high computational efficiency.

4.2.1 Characterizing descriptive Visual Words

Let us start with a revisit to the formulation of the retrieval model based on classical BovW framework. A given query example, which is also a video, is represented as $v_q = \{\boldsymbol{f}_l\}$ where f_l is a frame. For efficiency, a number of key frames $\{\boldsymbol{f}_q\} \subset v_q$ are sampled for the video similarity measurement. Note, however, that we use all frames $\{\boldsymbol{f}_l\}$ in the query for spatial-temporal correlation detection and measurement. In addition, each video collection in the video collection is represented as a set of key frames $v_d = \{\boldsymbol{f}_d\}$, whereas each element of the i^{th} visual words w_i in $\boldsymbol{f}$ is its Term Frequency (TF).

The BovW model usually involves a very large vocabulary, and the representation vector $\boldsymbol{f}$ is sparse. Therefore, the inverted index architecture can be applied: for each visual word, a table is built to list all the frames where it appears and its occurence frequencies in these frames. The key frame similarity $sim(\boldsymbol{f}_d, \boldsymbol{f}_q)$ is measured by the cosine function, which can be approximated based on the inverted index structure using:

$$sim(\boldsymbol{f}_d, \boldsymbol{f}_q) \approx \frac{\sum_{i=1}^{K} score(w_i)}{l(\boldsymbol{f}_d) * l(\boldsymbol{f}_q)} \tag{4.7}$$

where $l(\boldsymbol{f})$ is the L^2-Norm of vector $\boldsymbol{f}$, and $score(w_i)$ is the scoring function for each matched visual word w_i across $\boldsymbol{f}_d$ and $\boldsymbol{f}_q$, given by multiplication of the corresponding TFs:

$$score(w_i) = \boldsymbol{f}_q(w_i) * \boldsymbol{f}_d(w_i) \tag{4.8}$$

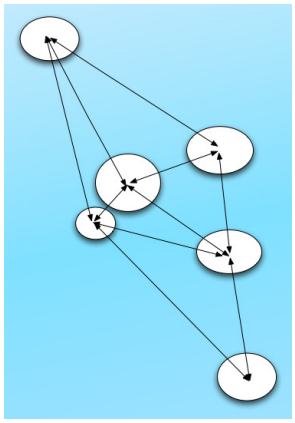
The scores are accumulated to compute a similarity score between two key frames. It has been assumed that the video data are well segmented shots or the videos are short and consisting of few shots. We then adopt the shot similarity measurement method proposed by Peng et al. (Peng & Ngo 2005a) where the highest similarity score among all possible pairs of key frames compared is used to measure the similarity between two video shots:

$$sim_{v_d, v_q} = \max_{\boldsymbol{f}_d \in v_d, \boldsymbol{f}_q \in v_q} sim(\boldsymbol{f}_d, \boldsymbol{f}_q) \tag{4.9}$$

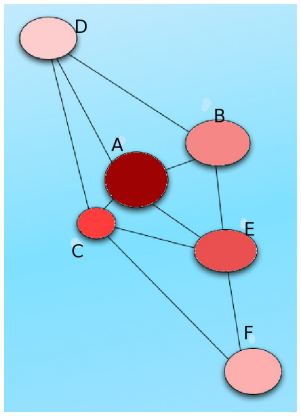
We aim to improve this similarity function by emphasizing the visual words in the query that fulfill the spatial proximity and temporal motion coherence constraints. Effectively, these visual words tend to be associated with an identical object in the query. Traditionally, this association was addressed by image segmentation, which is computationally expensive (Datta et al. 2008) and a high computationally cost in the online query is not acceptable for modern CBVR technology. It has been preliminarily shown that the visual words, which are more interesting when it comes to representing the visual content, can be extracted according to the two assumptions proposed in Section 4.1. Here, we develop a further characterizing scheme for the query video to improve the retrieval model.

Regarding the spatial correlation, it is measured by the proximity between visual words, e.g. the inverse of Euclidean Distance (Zhang et al. 2006, Liu & Chen 2009). An example is illustrated in Figure 4.3(a). Visual word A is located in a close proximity to B, C and E in a key frame of the query, and visual word A is assigned with a higher descriptive power as shown in Figure 4.3(b). In contrast, visual word D is located singularly, and as a result, it is assigned a lower descriptive power.

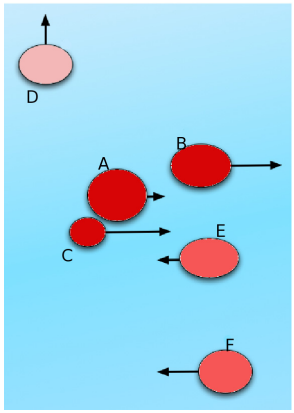
In addition, in the temporally consecutive frames, the reoccurring visual words are tracked according to the L^2 Norm. Each tracked visual word moves to a new position in the next frame, and is associated with a motion vector (Figure 4.3(c)). We propose



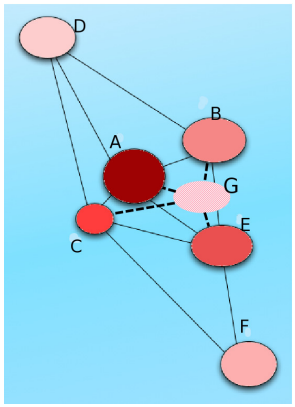
(a) Visual words in the key frame of query video



(b) Spatial correlation based emphasizing



(c) Motion coherence based emphasizing



(d) Emphasizing the descriptive visual word not appearing in the key frame

Figure 4.3: An example of visual words emphasizing approach. In (b), (c) and (d), the color intensity indicates the importance.

to measure the relative motion with respect to two visual words in order to indicate how the two visual words move coherently. In Figure 4.3(c), visual words A, B, and C move coherently to the right, and they are assigned with a higher descriptive power. The visual word D moves in a different direction, which has been defined as low motion coherence, and it would not be emphasized in the query.

Eventually, both the spatial and temporal relationship can be modelled by the STC technology in Chapter 3. Hence, we can state a hypothesis as follows:

When a visual word has a strong STC with other words appearing in the same frame of query, it has a higher descriptive power.

Accordingly, our scoring function based on STC for a visual word in a key frame of the query video is then formulated as follows:

$$score'(w_i) = \sum_{j=1}^{K} \boldsymbol{f}_q(w_i) * \boldsymbol{Corr}(i,j) * \boldsymbol{f}_d(w_j) \tag{4.10}$$

where $\boldsymbol{Corr}(i,j)$ measures the correlation between the i^{th} and j^{th} visual words. Here, we defined $\boldsymbol{Corr}(i,i) = 1$ and $\boldsymbol{Corr}(i,j) < 1$ as shown in the theoretical analysis in Chapter 3.

As shown in Equation 4.10, traditionally the similarity scoring of w_i is only determined by the matched instances of corresponding w_i in the video. In the STC based similarity measurement, it also depends on whether its correlated w_j appears in the data video. The score assigned $\boldsymbol{f}_q(w_i) * \boldsymbol{Corr}(i,j) * \boldsymbol{f}_d(w_j)$ has a coefficient $\boldsymbol{Corr}(i,j)$, which is determined by the STC discovered from the query. It can be replaced by $\boldsymbol{Corr}_s(i,j)$, $\boldsymbol{Corr}_t(i,j)$, or $\boldsymbol{Corr}_{st}(i,j)$ which represents different aspects of STC discovered respectively as introduced in Chapter 3.

It should be noted that, in the last section, $\boldsymbol{Corr}^f$ is utilized and the WoIs are selected based on the frame level correlation. When the query is a video example, then, the video level correlation can be leveraged to address the descriptive visual words. As shown in Figure 4.4, the $\boldsymbol{Corr}^q$ denotes the correlation matrix which is incurred by the query video. It means that this correlations between the visual words is learned from the query video, rather than the individual key frame, and we define it as **QueryCorrelation** (**QC**). Using this method, the hidden clue from the video query is discovered, to better predict the relevances.

Formally, Equation 4.10 is rewritten to incorporate the **QC** into the similarity scoring function as:

$$score'(w_i) = \sum_{j=1}^{K} \boldsymbol{f}_q(w_i) * \boldsymbol{Corr}^q(i,j) * \boldsymbol{f}_d(w_j) \tag{4.11}$$

Based on this scoring function, the video similarity measurement would be reformulated, and the formulation of QC based similarity measurement function will be discussed in the

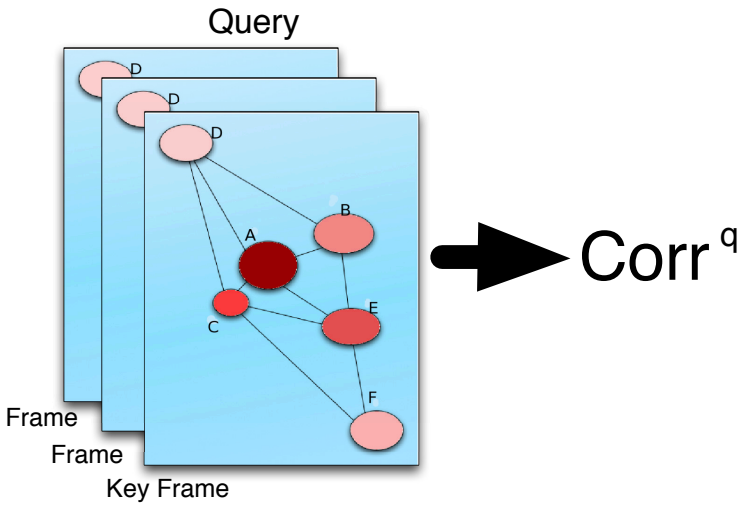


Figure 4.4: **QC**: Spatial and Temporal Correlation discovered from the Query

next subsection.

4.2.2 Key Frame Reformulation and Similarity Measurement

By incorporating a QC-based scoring function (Equation 4.11), the key frame similarity measure in Equation 4.8 becomes:

$$sim(\boldsymbol{f}_d, \boldsymbol{f}_q) \approx \frac{\sum_{i=1}^{K} score'(w_i)}{l(\boldsymbol{f}_d) * l(\boldsymbol{f}_q)} \tag{4.12}$$

However, direct computation of Equation 4.12 is difficult to implement for the inverted index architecture due to the STC computation for the query. The original scoring function of Equation 4.8 corresponds to the inner product between the query and data representation vectors, which can be easily applied to the inverted index system. To facilitate a similar computation, the numerator of Equation 4.12 is rewritten as follows:

$$\begin{aligned}\sum_{i=1}^{K} score'(w_i) &= \sum_{i=1}^{K}\sum_{j=1}^{K} \boldsymbol{f}_d(w_j) * \boldsymbol{f}_q(w_i) * \boldsymbol{Corr}^q(i,j) \\ &= \sum_{j=1}^{K} \boldsymbol{f}_d(w_j) * \sum_{i=1}^{K} \boldsymbol{Corr}^q(i,j) * \boldsymbol{f}_q(w_i) \\ &= \sum_{j=1}^{K} \boldsymbol{qc}(w_j) * \boldsymbol{f}_d(w_j)\end{aligned} \tag{4.13}$$

The weighting vector $\boldsymbol{qc} = \boldsymbol{Corr}^q \times \boldsymbol{f}_q$ and $qc \in \mathbf{R}^K$, where $qc(w_i)$ denotes the descriptive power (also called emphasizing weights) of the i^{th} visual word, and can also be directly

used as the emphasizing weights. $\boldsymbol{Corr}^q$ is the $K \times K$ STC matrix.

As shown in Equation 4.13, the QC is injected into the scoring function in order to emphasize the descriptive visual words in the keyframes, which is equivalent to key frame reformulation. It must be noted that the QC is measured using all frames in the query (Figure 4.4), and as a result, some visual words may be brought by the QC into the reformulated key frame, even though their original term frequency on the key frame is zero. For example, as shown in Figure 4.3(d), the visual word G is added with its corresponding weights, because of its strong QC with A, B, C and E in the whole query. In this way, the QC based approach would, to some extent, compensate the information loss caused by the key frame sampling.

It is interesting to compare Equations 4.10 and 4.13. Equation 4.13 demonstrates our assumption that the scoring function should not be completely determined by the matching of "independent" visual words, but also the visual words are correlated via spatial and temporal information. Equation 4.13 presents the approach to emphasize the descriptive elements in the visual content representation. After all, these two equations are mathematically equivalent, which shows that emphasizing the descriptive visual words in the query is equivalent to compensating the spatial-temporal correlation neglected by the BovW framework.

An example of QC weights computed for a random frame by Equation 4.13 is shown in Figure 4.5. As shown in the diagram, some weights are extreme large, which may greatly influence the relevance scores. We want to avoid the risk to over-estimate some relevance by strong bias on some individual visual words. As a result, Equation 4.14 has been utilized to further quantize the emphasizing weights of the descriptive visual words in the key frame:

$$\boldsymbol{qc}(w_i) = \begin{cases} 2 & \text{for} \quad \boldsymbol{qc}(w_i) > \sigma \\ 1 & \text{for} \quad \sigma/2 < \boldsymbol{qc}(w_i) < \sigma \\ 0 & \text{for} \quad else \end{cases} \tag{4.14}$$

As shown in Figure 4.6, the choice of σ determines the number of descriptive visual words. Figure 4.6 represents some examples of the $qc(w_i)$ computed for various queries. Given the same detector is used, the descriptive power distributions for different queries are not very different. Thus a static threshold σ can be empirically selected for all queries (σ=1 in our experiment). The effect of quantization scheme will be discussed in the Experiments section.

The keyframe representation of the query is finally reformulated as:

$$\boldsymbol{f}'_q = \boldsymbol{f}_q + k_{qc} * \boldsymbol{qc} \tag{4.15}$$

where k_{qc} is a parameter to determine the role of QC weight in the query representation. In this way, the Term Frequency is adjusted by QC weights, this adjust can be summarized

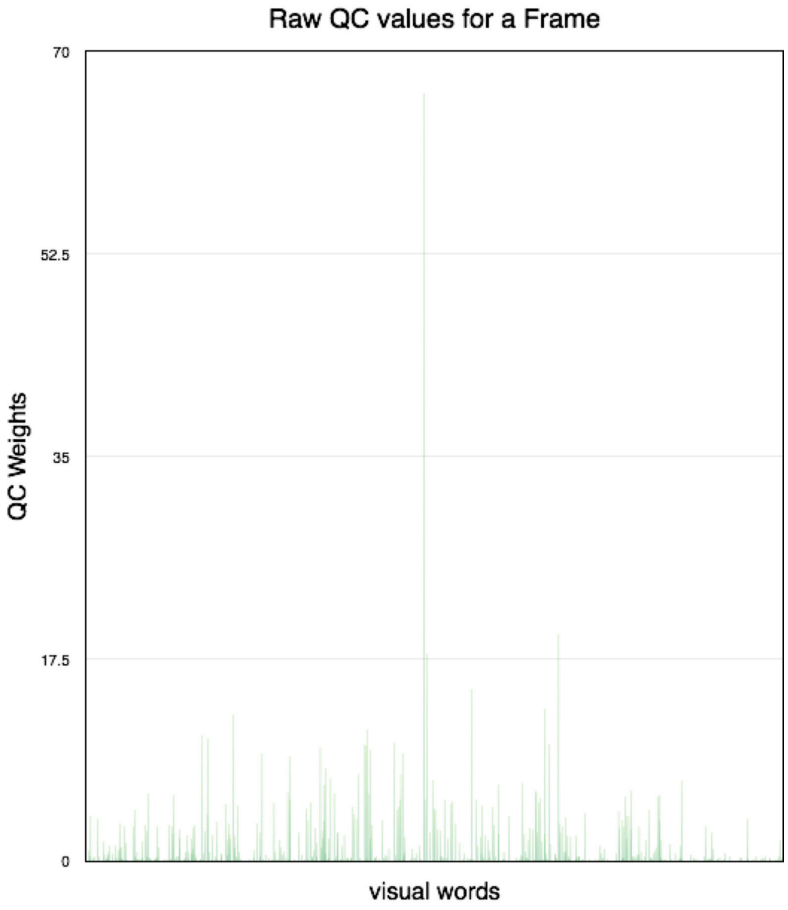


Figure 4.5: **QC** weights computed for an frame of a query

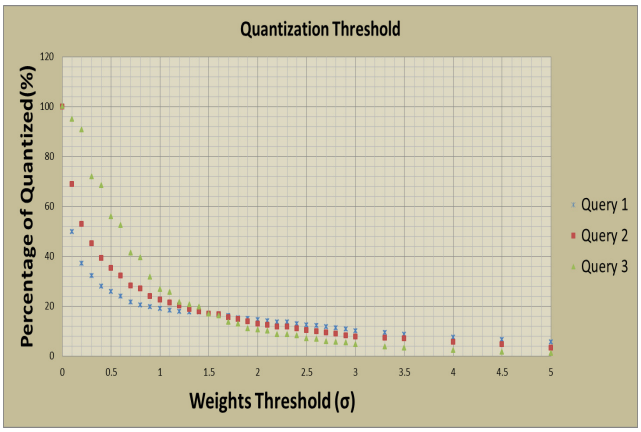


Figure 4.6: The quantization threshold σ decides the number of emphasized visual words.

as:

$$Weight = \textbf{TermFrequency} + \textbf{QueryCorrelation} \tag{4.16}$$

This equation shows that not only frequently appearing visual words is important for current query in our retrieval model, but also the strongly spatial-temporal correlated visual words are also important for it. As a result, QC is used as an additional weight to traditional term weights, for example, TF.

Based on Equations 4.12 and 4.15, the frame level similarity measurement function becomes:

$$sim'(f_d, f_q) \approx \frac{\sum_{i=1}^{K} \boldsymbol{f}'_q(w_i) * \boldsymbol{f}_d(w_i)}{l(\boldsymbol{f}_d) * l(\boldsymbol{f}_q)} \tag{4.17}$$

It is important to note that the key frame reformulation will not significantly increase the number of non-zero elements in the key frame representation. It would not only involve the descriptive visual words, but also exclude the noises. Furthermore, it avoids the extra computational and memory costs of the direct inclusion of the spatial-temporal information in video representation and indexing.

Having said that, in this thesis, we are more interested in how the STC-based approach can improve the retrieval effectiveness. In the following sections, we present an extensive empirical evaluation.

4.3 Inverse Documents STC

According to the assumptions made for the QC, STC incurred by the query have been assumed to reveal how visual words relevant to the topic. Considering the STC contained within video collection, an idea is that the weight of a visual word should be modified in the retrieval model whilst considering if they are correlated to other visual words in the current video collection.

We assume that the more likely a visual word w_i is to co-occur with other words, the less descriptive ability it has for the retrieval against the current video collection. For example, if there were two visual words representing "screen" and "keyboard" respectively. The "screen" co-appears more often with other visual information such as: "iPad", "tv-set", "GPS cable", and etc., and "keyboard" appears more singularly in the video collection. The visual word "screen" would be less descriptive as a result of the larger number of visual words it co-occurs with correlation. When a user searches with a visual query "Desktop", the "keyboard" is more descriptive for this topic than the "screen": a video only containing "keyboard" is more likely to be related to "Desktop" than the a video containing a single "screen" only.

In Chapter 3, the co-occurring correlation has been quantized as a matrix $\boldsymbol{Corr}^d$ (shown in Figure 4.7), which accumulates the video level correlation through the entire

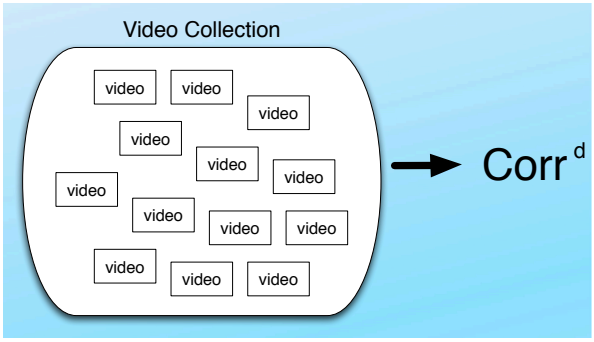


Figure 4.7: Co-occurrence correlation extracted from entire video collection.

collection. We can formulate a notion of "document correlation" for an individual visual word as follows:

$$\boldsymbol{dc}(i) = \sum_{j=0}^{K} \boldsymbol{Corr}^{d}(i,j) \tag{4.18}$$

where $\boldsymbol{dc}(i)$ denotes the document correlation for the i^{th} visual word. Theoretically, it should be normalized around 1. If the distribution of visual words in the visual content were uniform, all the $\boldsymbol{dc}(i)$ should be equally 1. For computation convenience, we scale the $\boldsymbol{dc}$ to:

$$\bar{\boldsymbol{dc}}(i) = K * \frac{\boldsymbol{dc}(i)}{\sum \boldsymbol{dc}(i)} \tag{4.19}$$

A high document correlation associated with single visual word has been assumed to harm its descriptive ability. We then define a concept Inverse Document Correlation (IDC) as follows:

$$\boldsymbol{idc(i)} = 1 + log(\frac{1}{\bar{\boldsymbol{dc}}(i)}) \tag{4.20}$$

where each $\boldsymbol{idc(i)}$ should be non-negative and inverse to the document correlation dc of w_i with respect to other visual words.

By incorporating the new term weights IDC, the keyframe representation of the query is then reformulated as:

$$\boldsymbol{f}'_q(i) = \boldsymbol{f}_q(i) + k_{idc} * tf(i) * \boldsymbol{idc}(i) \tag{4.21}$$

where $\boldsymbol{idc}$ is a K dimensional vector, in which each element is calculated by Equation 4.20, and k_{idc} is a parameter to scale the IDC weights in the representation. Combining QC and IDC results in:

$$\boldsymbol{f}_q^{qcidc} = \boldsymbol{f}_q + k_{qc} * \boldsymbol{qc} + k_{idc} * \boldsymbol{idc} \tag{4.22}$$

As shown in Euqation 4.22, IDC reduce the weights of visual words which are more likely

to co-occur with other visual words within the data sets. In another words, it relatively emphasis more on the visual words which appear solely and less, which are assumed to contain more useful information for retrieval. This IDC is similar to the idea behind the Inverse Document Frequency (IDF), which is another common term weighting scheme for textual retrieval. It implies that words appearing in more documents are less descriptive and should be assigned with a lower weights in the retrieval model. Similarly, IDC implies that the words appearing more likely with other words are less descriptive.

The weights for visual content representation can be reformulated as:

$$Weight = \textbf{TermFrequency} + \textbf{InverseDocumentCorrelation} \tag{4.23}$$

which can be combined with Equation 4.16 to characterize the representation reformulation function based on the STC discovered from both query and document collection:

$$Weight = \textbf{TF} + \textbf{QC} + \textbf{IDC} \tag{4.24}$$

Note that this representation reformulation is equivalent to the new similarity measurement scheme. The frame level similarity measurement is computed according to:

$$sim(f_d, f_q) \approx \frac{\sum_{i=1}^{K} \boldsymbol{f}_q^{qcidc}(w_i) * \boldsymbol{f}_d(w_i)}{l(\boldsymbol{f}_d) * l(\boldsymbol{f}_q)} \tag{4.25}$$

In summary, we build a retrieval model using term weights QC and IDC in addition to classical video representation, which is based on, for example, TF term weighting scheme.

4.4 Experiments

The goal of these experiments is to evaluate the effectiveness of the STC-based representation reformulation, so as to improve the BovW retrieval model. Accordingly two Query-by-Example video retrieval tasks are used: (1) QBE near-duplicate video search task; (2) general topics QBE video retrieval task. The QBE video retrieval for general topics is always a more challenging task. The topics cover the various user intentions, who may search for a specific object/scene or a category of shot. The visual similarity of the desired object may be relatively small in the relevant videos. For example, as shown in Figure 4.8, one of the topics of this task is searching for videos which contain women wearing long dresses. In the relevant videos, the persons may appear differently, and only a small amount of visual content of the relevant videos is visually matched to the query.

In our experiments, Points-of-Interest are detected by the Hessian detector, which works well to overcome the occlusion and cluttering (Mikolajczyk, Tuytelaars, Schmid, Zisserman, Matas, Schaffalitzky, Kadir & Gool 2005). The salient regions are described by the SIFT feature. Hierarchical K-means is used for visual vocabulary construction.

Figure 4.8: Typical frames of relevant videos of topic "Women in long dresses"

Mean Average Precision (MAP) is used as the main performance indicator, and we also show the Precision-Recall curves of different models.

The classical BovW model and a state of the art BovW enhancement approach based on Tight Geometric Constraint (TGC) (Zhao et al. 2010) are used as the baselines. The TGC method is implemented with a publicly available toolkit SOTU (Zhao 2009).

4.4.1 Experimental Set Up

Two commonly used video collections are selected for the experiments:

`(DATA1) CC_Web_Video` Near-duplicate video search is performed on video collection CC _Web_Video (Wu, Ngo, Hauptmann & Tan 2009). Most videos in this data-collection are short videos, most of which are 3-5 minutes long and not longer than 10 minutes. They are presented on the websites: Youtube, Yahoo and Google Video. From the original video collection, totally 336K key frames are extracted to represent the videos in the video collection. The videos in the ground truth are labeled by "*Exact duplicate*", "*Similar*", "*Major Changed*", "*Long version*" and "*Not Relevant*". The evaluation is performed for 24 topics, using 69 queries respectively selected from "*Exact*", "*Similar*" and "*Major Changed*" videos. 10 key frames are sampled from each query for retrieval. The average number of the relevant videos is 84.7 per topic.

`(DATA2) TRECVID2002` It is selected to perform the general topic video retrieval. The videos in this video collection contain diversified video sources: old film, news, documentary and advertisement, whilst the topics cover various information needs. The video collection consists of approximately 10K shots segmented from 133 videos based on the shot boundary ground truth provided by TRECVID2002. In total, 79K keyframes are

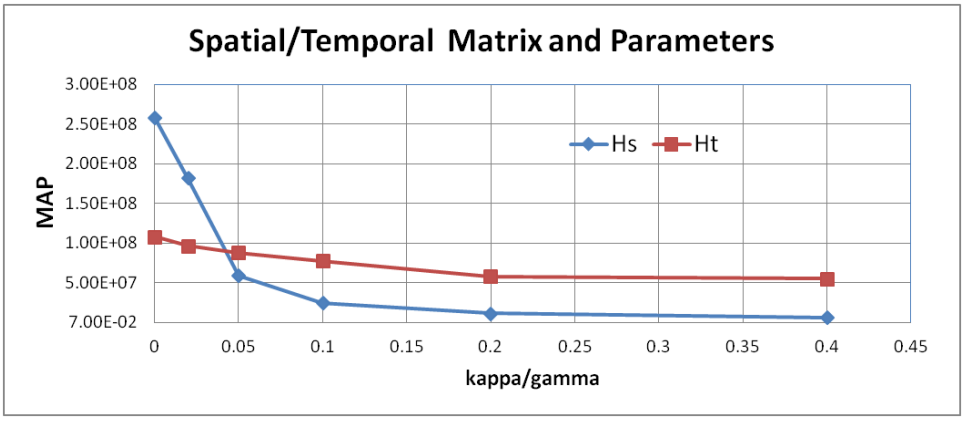


Figure 4.9: The influences of κ and γ on spatial and temporal matrix

sampled (1 keyframe per 2 second) to represents the shots. The TRECVID retrieval task consists of 22 topics, however, it does not provide queries in modality of video. To evaluate query-by-video-example retrieval model, the users of this video collection must compose the queries by themselves to perform the searching. To make our work more repeatable and comparable with other researchers, we do not use external visual resources to compose the queries for experiments in this thesis. We randomly select relevant videos from the ground truth as query examples for each topic, and then delete them from the ground truth. In total, 71 queries run are performed as QBE video search in the experiment. On average, each topic is associated with 23 relevant shots.

4.4.2 Parameters

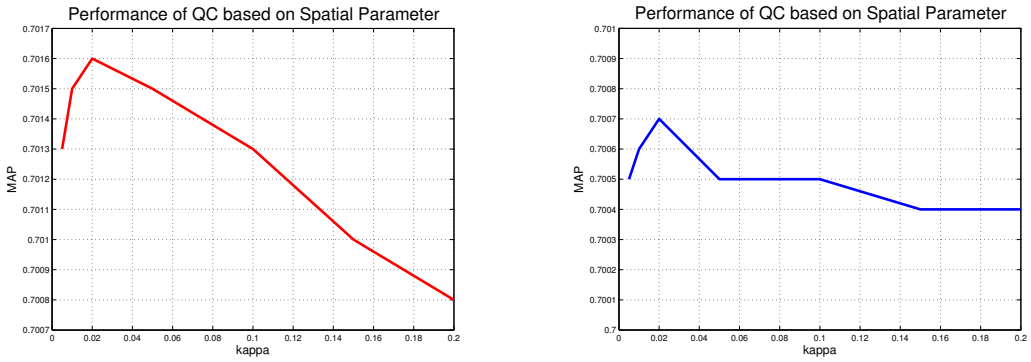
As presented in the previous sections, our STC based representation reformulation method includes several parameters: the spatial proximity parameter κ, the temporal coherence parameter γ, the QC parameter k_{qc}, and the IDC parameter k_{idc}.

As shown in Equations 3.29 and 3.49, the κ and γ determine the scale of STC discovery, which are defined in Chapter 3. We could demonstrate their influences on the STC matrix by a defined value H, which accumulates all entries of the spatial matrix or the temporal matrix as follows:

$$\begin{aligned} Hs(\kappa) &= sum(\boldsymbol{S}^d) \\ Ht(\gamma) &= sum(\boldsymbol{T}^d) \end{aligned} \tag{4.26}$$

Figure 4.9 demonstrates the changes of Hs and Ht computed from TRECVID2002 along with the κ and γ. It can be seen from Figure 4.9 that both Hs and Ht decrease with the increasing κ and γ, and it perfectly fulfills our theoretical analysis.

Furthermore, the Hs decreases more quickly where κ increases from 0.02 to 0.1. Ac-



(a) κ (b) γ

Figure 4.10: The influences of parameters on the performance of STC

cording to Equation 4.27, the Hs should change most quickly when more spatial distances are around $\frac{1}{\kappa}$, thus demonstrating that the the spatial distances between visual words are likely to be distributed in a range of 10-50 pixels.

$$\begin{aligned} & \frac{d^2 Hs}{\partial\kappa\partial d_{a,b}} &=& 0 \\ \Rightarrow\quad & \frac{\partial(-d_{a,b} * e^{-\kappa d_{a,b}})}{\partial d_{a,b}} &=& 0 \\ \Rightarrow\quad & e^{-\kappa d_{a,b}} * (1 - \kappa d_{a,b}) &=& 0 \\ \Rightarrow\quad & 1 - \kappa d_{a,b} &=& 0 \\ \Rightarrow\quad & \frac{1}{\kappa} &=& d_{a,b} \end{aligned} \tag{4.27}$$

Compared to the spatial proximity, the decease of Ht is a lot smoother. This is due to the fact that the distribution of the Norm of relative motion vectors between the visual words is more uniform. The Norm of relative motion vectors is relatively concentrated in a range $\gamma = 0.01 - 0.02$.

Furthermore, we preliminarily demonstrate the influence of κ and γ on the retrieval performance of STC-based technologies. The retrieval performance on TRECVID2002 are demonstrated in terms of MAP, and we utilize the QC based query reformulation as an example.

As shown in Figure 4.10(a), observation reveals that the performance of QC technology is not sensitive to the value of κ, and the value of MAP is always around 0.0701. The higher performance is achieved by $\kappa \approx 0.02$, which matches the spatial proximity between more visual words shown in Figure 4.9. In addition, γ also do have little effect on the performance of STC in Figure 4.10(b). A higher performance is achieved at $\gamma = 0.01 - 0.02$.

Generally, too large κ and γ ($\kappa > 0.05$ or $\gamma > 0.02$) always leads to a lower retrieval performance, which may be caused by the ignorance of some meaningful visual word correlation. Because, a larger parameter means that more visual word correlation will be neglected.

Simultaneously, the performance of temporal correlation based QC reformulation is more stable than spatial correlation, which is shown in Figure 4.10(b). It is also coincident with the relative smoothness of Ht demonstrated by Figure 4.9.

In summary, the performance of QC technology does not strongly rely on the value selection of parameters κ and γ. In the following experiments, we empirically choose $\kappa = 0.02$ and $\gamma = 0.02$, which have achieved a higher performance in these results.

Furthermore, k_{qc} determines the proportion of the QC weights in the formulated query representation for CBVR, and a higher k_{qc} will assign more QC weights to the representation. Naturally, it would also impact the retrieval performance of QC method. We still use QC method running on TRECVID2002 as an example, and the results are displayed in Figure 4.11.

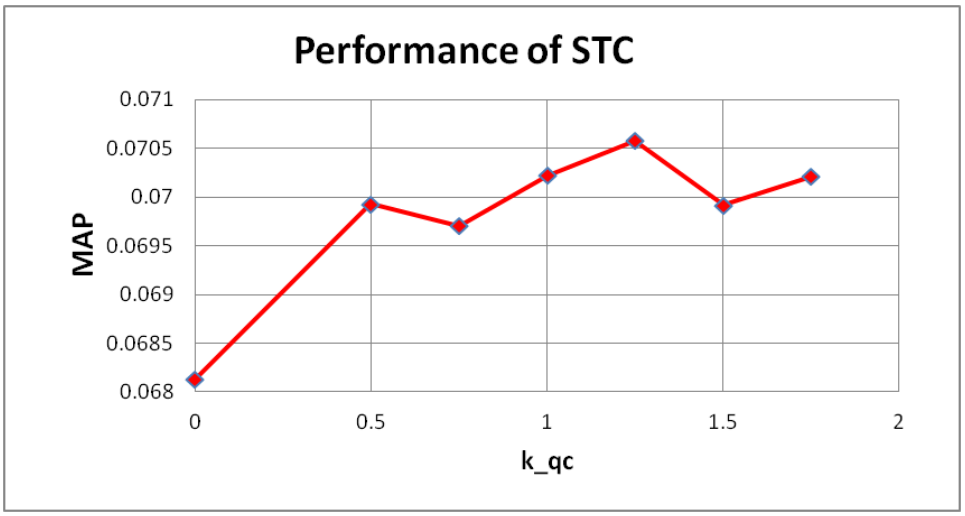


Figure 4.11: The parameter k_{qc} and the performance of STC

When $k_{qc} = 0$, the QC reformulated query is equivalent to the original query. Along with increasing k_{qc}, the retrieval performance also increases. However, it is not true that a larger QC would necessarily result in better results. When k_{qc} is larger than 1.5, the performance decreases. It means that over-emphasis the QC would lead to a risk of decreasing the retrieval performances. Generally, the better performance is achieved by $k_{qc} = 1$ or 1.25. In the following experiments, the value of parameter k_{qc} is empirically selected as 1 to simplify the computation.

4.4.3 QC Evaluation

In this section, we perform the experiments to evaluate the query representation modified method by QC weighting scheme. The performances of four variations of the approaches proposed in this section are reported, namely the QC-based retrieval models with and without weights quantization (Equation 4.14), denoted by **qc-st-BovW** and **qc-raw-st**

respectively; and retrieval model only based on either spatial or temporal correlation, denoted by **qc-s-BovW** and **qc-t-BovW** respectively.

In this experiment, we test the vocabulary of size 20k/5k for CC_WEB_VIDEO/TREC -VID2002 respectively. The goal is to evaluate whether or not the proposed QC technique can effectively compensate the neglect of spatial-temporal structure for the BovW model for either smaller or larger vocabulary.

Precision and Recall

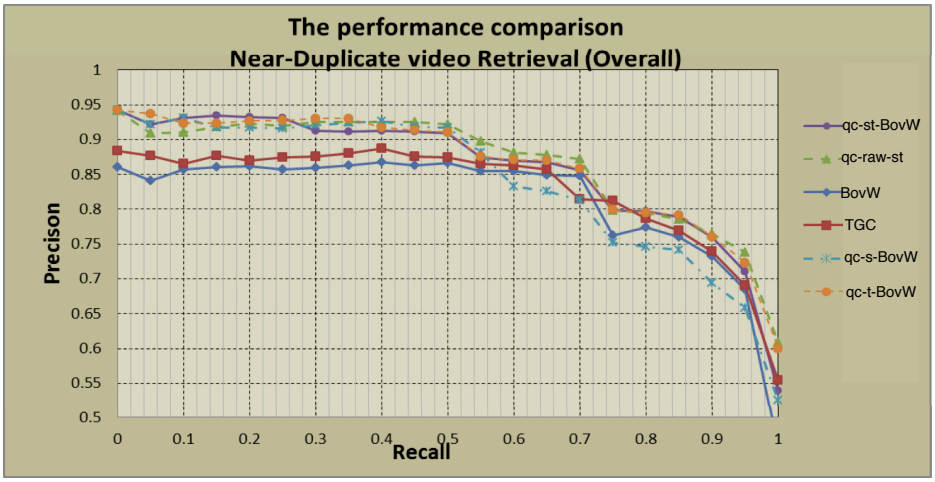


Figure 4.12: Overall performance in CC_WEB_VIDEO

The overall performance of QC based retrieval approaches in CC_WEB_VIDEO is displayed by the Precision-Recall curves in Figure 4.12. The qc-st-BovW simply outperforms both the classical BovW and TGC methods. The overall performance reveals that the QC approach effectively emphasises the descriptive visual words in the retrieval function, and more relevant videos are retrieved. Especially, the QC improves the quality of the top ranked results, which is important for real-world application and its potential cooperation with other technology, *e.g.*, pseudo feedback. The qc-raw-st performs similarly but slightly lower than the qc-s-BovW, qc-t-BovW and qc-st-BovW.

Noted that the qc-raw-st performs nearly the same as the qc-st-BovW, qc-s-BovW and qc-t-BovW in terms of Precision-Recall curves. However, as shown in Figure 4.12, the precision of top ranked results retrieved by qc-raw-st is not as good as qc-st-BovW and other two approaches. These results may be owing to the extreme weights assigned to some visual words which may actually be noise and involve irrelevant results. The weights quantization scheme has effectively reduced the risk of extreme weights for the descriptive visual words generated by the QC-based key frame reformulation and makes the qc-st-BovW perform more stably.

The performance difference between qc-s-BovW, qc-t-BovW and qc-st-BovW in terms of Precision-Recall curve is not obvious. It shows that the performances are very similar, and it may be because the spatial and temporal correlation has emphasized the same number of visual words as descriptive. As we know, both the spatial and temporal correlations are extensions of the co-occurrence correlation, and they have an overlapping part. The difference between two correlations will be shown in terms of MAP in the next section.

In CC_WEB_VIDEO, the queries are classified as "Exact", "Similar" and "Major Changed", based on the similarity between the query example and most relevances in the ground truth. The "Exact" and "Similar" queries always have a higher quality, and the quality of "Major Changed" is normally the lowest. As a result, the influence of query quality on the retrieval performance of QBE-CBVR system is obvious. The results of good queries ("Exact" and "Similar") are a lot better than the other type of queries, which can be observed in the experimental results.

Figures 4.13 and 4.14 present the retrieval performances of different approaches using "Exact" and "Similar" queries respectively. In terms of the precision-recall shown in Figure 4.13 and Figure 4.14, the classical BovW model performs adequately well on the "Exact" and "Similar" queries. Most positive results are ranked on top of the list, thus the precision is as high as nearly 100%. The other baseline TGC maintains a similar performance with BovW for both types of queries. The proposed qc-s-BovW, qc-t-BovW and qc-st-BovW perform comparable to the baselines, and qc-t-BovW performs better than others on top results for the "Similar" queries.

Generally, with "good" queries, which is not challenging for retrieval, all the approaches perform very well and closely to each other. However, it is not always an easy task to select a "good" query for video search. A robust CBVR system should be able to handle various queries of different qualities. We aim to test whether or not the QC improve the ability of the BovW framework to deal with hard queries.

For "Major Changed" queries, which are of low quality, as shown by the Precision-Recall curves in Figure 4.15, BovW performs a lot worse than on the "good" queries, and Precisions decrease very quickly around the point where Recall=0.55. It is also an evidence that these queries are harder tasks for the BovW framework. TGC performs a little better than the BovW, and it shows the effectiveness of additional TGC verification to enhance the BovW framework. Nevertheless, the qc-st-BovW outperforms the classical BovW model, and it also outperforms the TGC as shown in Figure 4.15. We can conclude that it substantially improves the performance of the BovW framework in terms of Precision-Recall for the hard queries.

If we summarize the above observations, it can be shown that the difference between the QC and the two baselines BovW and TGC, which is presented in the overall performance (Figure 4.12), is mainly due to the fact that retrieval performances of the "bad" queries have been improved .

In summary, the evaluation on CC_Web_Video shows that both TGC and qc-st-BovW effectively improve the performance of the BovW model in terms of Precision and Recall curve. The qc-st-BovW largely outperforms TGC on lower quality queries, which is more challenging for classical BovW model, and it has also performed comparably to TGC on high quality queries.

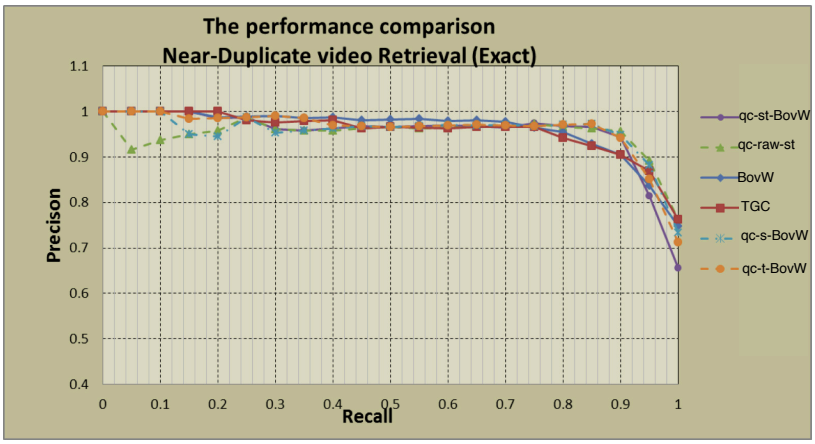


Figure 4.13: Performance of "Exact" queries in CC_WEB_VIDEO

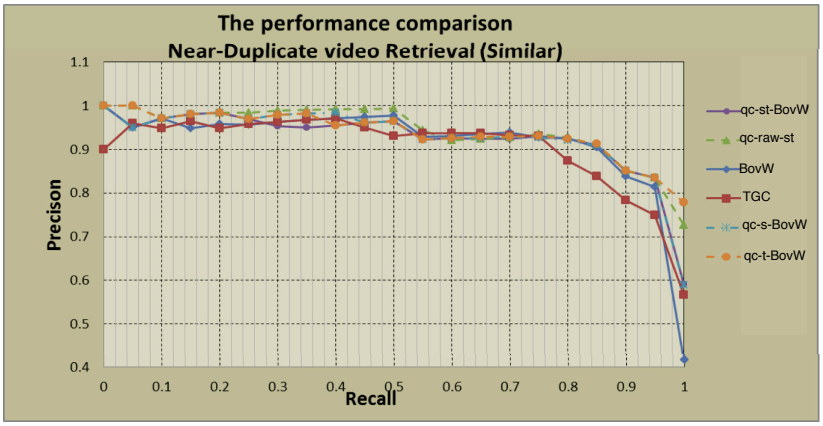


Figure 4.14: Performance of "Similar" queries in CC_WEB_VIDEO

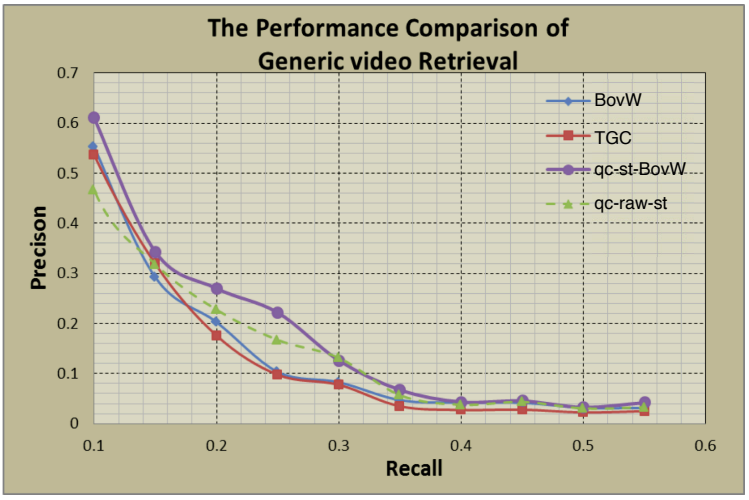


Figure 4.16: Precision-Recall curve for 7 typical queries for TRECVID2002

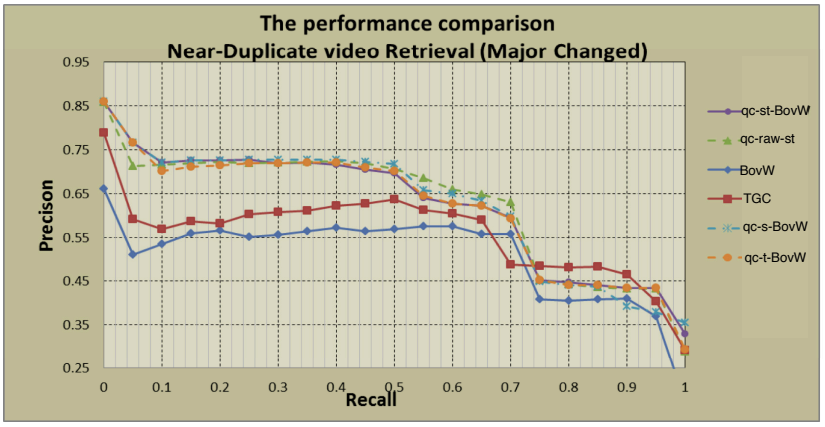


Figure 4.15: Performance of "Major Changed" queries in CC_WEB_VIDEO

TRECVID2002 is a more challenging task for the retrieval system, which can also be shown by the experimental results. As displayed by Figure 4.16, the precision-recall curve is lower than Figure 4.15, and the effectiveness of BovW for this task is much lower than for CC_WEB_VIDEO.

Firstly, 7 typical topics (special individual, persons, objects, and land view) are selected to evaluate the performance of the QC method on the general topic video retrieval. The results in Figure 4.16 shows that qc-st-BovW outperforms the classical BovW model and TGC model. TGC and BovW perform comparable in this task. The relevances in the general topic retrieval experiments have fewer visual similarities, and the additional information verification in the similarity measurement would not be as effective as in the near duplicate visual content search.

Table 4.1: MAP performance of QC on CC_WEB_VIDEO

MAP	E	S	M	Overall
BovW	0.939	0.919	0.511	0.8120
TGC	0.940	0.913	0.553	0.8268(+0.010)
qc-st-BovW	**0.944**	0.927	0.619	**0.8503(+0.033)**
qc-raw-st	0.929	**0.930**	**0.625**	0.8477(+0.030)
qc-s-BovW	0.938	0.925	0.619	0.8360(+0.021)
qc-t-BovW	0.941	0.926	0.615	0.8486(+0.032)

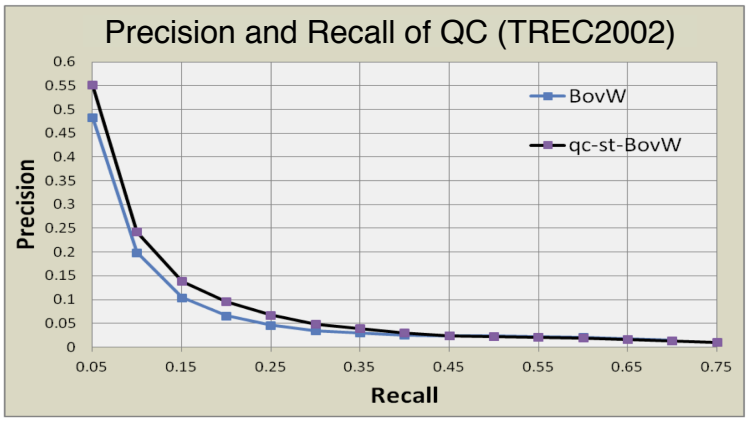


Figure 4.17: Performance of QC method Regarding Precision-Recall for all topics for TRECVID2002

The overall performance comparison of QC method and original BovW is illustrated in Figure 4.17. It shows that qc-st-BovW outperforms BovW model in terms of average Precision of all topics on different Recall points. We can conclude that the QC based approach could improve the performance of BovW framework for TRECVID2002 in terms of the Precision-Recall criteria.

Mean Average Precision

Mean Average Precision is a summarization criteria of precision and recall, which computes the average precision of all the retrieved positive results.

The performances of our approaches and baselines for CC_WEB_VIDEO in terms of MAP are presented in Table 4.1, where E, S, M denotes "Exact", "Similar" and "Major Changed" queries, respectively. On average, the qc-st-BovW outperforms the classical BovW by 4%, which is statistically significant (p-value=0.046). The qc-st-BovW performs much better than TGC on "bad" queries: it outperforms TGC by approximately 10% and BovW by around 20% on the "Major Changed" queries. It is clear that the improvement of the QC approaches is a result of its performances on "bad" queries.

It is also shown in Table 4.1 that qc-s-BovW performs similarly to qc-t-BovW, and

it means that both spatial and temporal correlation could be used to improve the BovW model in the near duplicate video search application. Furthermore, the STC, which combines both spatial and temporal information, generally outperforms any one of them for CC_WEB_VIDEO.

Furthermore, Table 4.2 presents the Average Precisions of different approaches on TRECVID2002. It is shown that qc-st-BovW outperforms the classical BovW model. The average improvement is 10%.

TGC fails to improve the classical BovW model for this general topic video retrieval task in terms of MAP. On average, qc-st-BovW outperforms both of the two baselines in terms of MAP.According to Table 4.2, the spatial correlation is more crucial than the temporal correlation, while qc-s-BovW outperforms qc-t-BovW on average. Differing with CC_WEB_VIDEO, the visual contents representing users' desire are often mixed up with a large amount of noise within the query videos of TRECVID2002. Thus the QC technique based on the spatial proximity assumption may contribute more than the motion coherence to the reduction of noise. Nevertheless, detailed discussion on the topic that to compare the spatial and temporal correlation will be demonstrated in Chapter 6.

The weights quantization scheme has improved the performance of QC-based approaches as shown in Table 4.2. qc-st-BovW performs more stably than qc-raw-st, and qc-raw-st performs worse than the BovW baseline on 3 topics although it achieves highest performance on the other two topics. Additional evidence is shown in Table 4.2, and the weights quantization scheme has promoted the precision on the top ranked retrieved results.

Table 4.2: MAP: QC performance on TRECVID2002

Approaches	qc-s-BovW	qc-t-BovW	qc-st-BovW	qc-raw-st
MAP (22 topics)	**0.0702**	0.0700	0.0701	0.0698
Average Precision on 10% Recall (22 topics)	**0.250**	0.250	0.250	0.246

In summary, the evaluation demonstrates that the QC technology can effectively improve the performance of the classical BovW framework in terms of MAP, and the qc-st-BovW performs more stably than the TGC-based approach on the two QBE CBVR tasks.

4.4.4 IDC Evaluation

IDC coefficients and Parameters

According to Equation 4.20, the number of generated IDC weights equals the scale K of visual vocabulary, and each visual word is associated with a fixed IDC weight for the current video collection. The statistic of the IDC weights is demonstrated by Table 4.3.

Both the average value of the IDC weights generated from TRECVID2002 and CC_WEB_VIDEO are around 1 after the normalization of Equations 4.19 and 4.20, which matches our theoretical analysis in the previous section. For both video collections, the deviation of IDC weights generated by temporal correlation is larger than the other two, thus meaning that the computed temporal correlations for the visual words differ more with each other. Compared to the other two, the temporal correlation extraction process has an extra step of visual word tracking, and it has reduced some co-occurrence of visual words. This step may contribute to the larger standard deviation existing in Table 4.3. The number of videos in the video collection CC_WEB_VIDEO is larger than TRECVID2002, and the standard deviation associated with the former video collection is also bigger.

Table 4.3: Statistics of IDC coefficients

	IDC Cooc	IDC SC	IDC TC
Average (TRECVID2002)	1.076	1.088	1.125
Standard Deviation	0.2154	0.2456	0.2715
Average (CC_WEB_VIDEO)	1.346	1.327	1.576
Standard Deviation	0.6726	0.6261	0.9163

The IDC weighting scheme is computed based on Equation 4.21, which has a parameter k_{idc}. It determines the role of IDC weights in the reformulated query example. We expect that different values of the k_{idc} would have effect on the performance of the IDC approach. The performance comparison is shown in Figure 4.18. Here, the IDC weighting with co-occurring correlation matrix is utilized for TRECVID2002 as an example, in which the preliminary evaluation is in terms of MAP.

Overall, the performance of IDC does change significantly with various values of the parameter k_{idc}: the variation of MAP is less than 0.001. As shown in Figure 4.18, the best performance is achieved by k_{idc} = 0.5-1. Note that the performance of IDC will decrease when k_{idc} is too large ($k_{idc} > 1.5$). In this case, too much IDC weights slightly harm the retrieval performance of the BovW model. For computation convenience, in the following experiments, the value of k_{idc} is empirically selected as 1. Nevertheless, it can be preliminarily concluded that the effectiveness of the IDC approach is not very sensitive to the parameter selection.

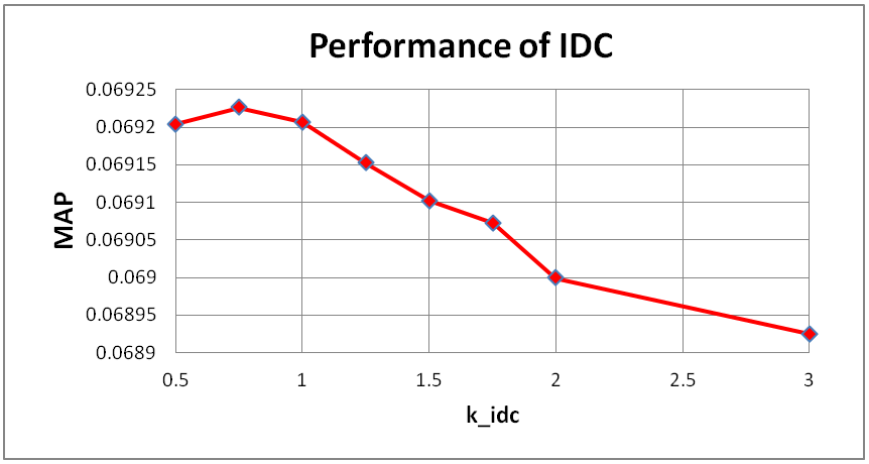


Figure 4.18: Performance of IDC and parameter k_{idc}

Precision-Recall Curve

Overall, the performance of the IDC approach on TRECVID2002 is displayed in Figure 4.19, in which the IDC weighting scheme computed based on co-occurrence, spatial proximity, and temporal correlation are denoted by "idc-c","idc-s", and "idc-t" respectively. It can be seen that the IDC approaches all outperformed the baseline original BovW model. The IDC weighting scheme slightly improves the descriptive ability of visual word for the video collection TRECVID2002. Especially, the retrieval performance improvement is more obvious for the top ranked results.

In Figure 4.19, the performance difference between "idc-c","idc-s", and "idc-t" is not very obvious. The reason may be that the IDC coefficients are computed based on the accumulation over entire video collection, and the spatial and temporal refinement do not significantly change the co-occurrence values of the visual words.

The IDC technology does not obviously outperform BovW for CC_WEB_VIDEO in terms of a Precision-Recall curve shown in the Figure 4.20. Only on the several Recall points less than 30%, the Precision of the IDC approach is higher than the BovW model, and for the Recall points 95% and 100%, the Precision of IDC approaches is slightly lower than the BovW model. This may reveal that representation reformulation based on the IDC weighting scheme has increased the risk of ignoring some relevant visual words.

In summary, according to the Precision-Recall criteria, the experimental results of two video collections shows that the IDC has improved the BovW model for the CBVR task. Some positive effects of the IDC weighting scheme on characterizing descriptive visual words could be observed in the experimental results.

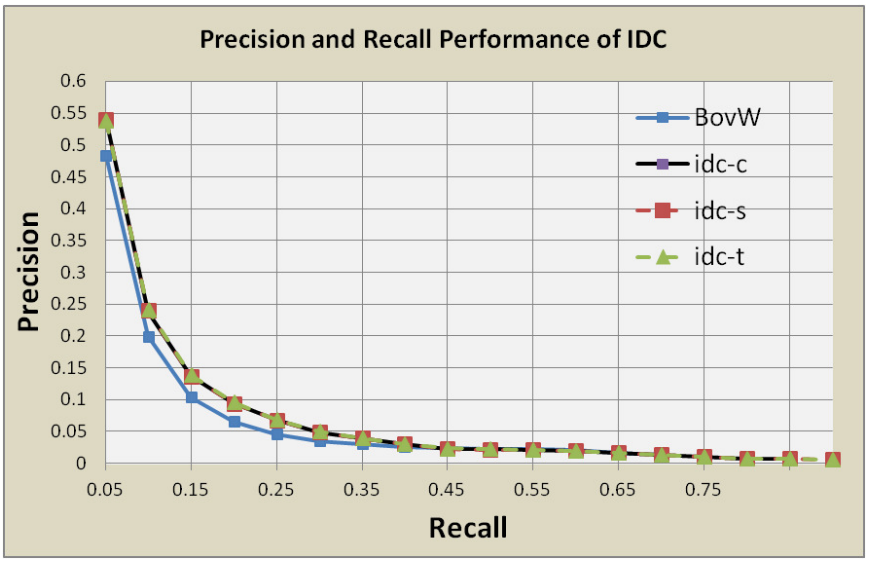


Figure 4.19: Performance of IDC method Regarding Precision-Recall for all queries for TRECVID2002

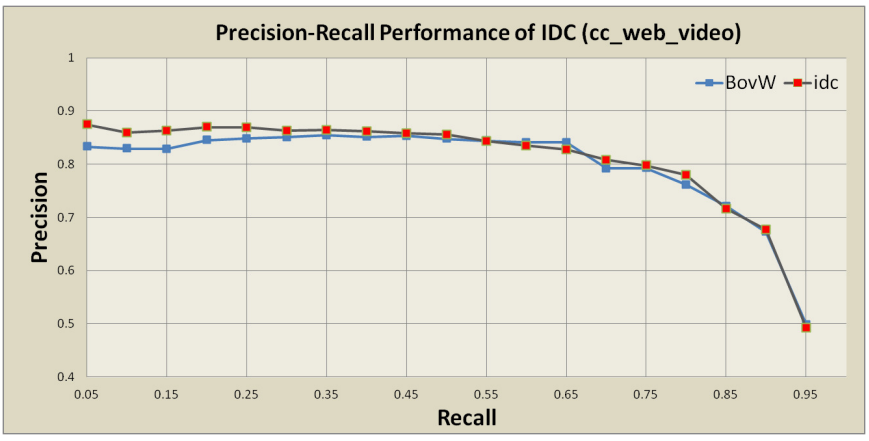


Figure 4.20: Performance of IDC method Regarding Precision-Recall for all queries for CC_WEB_VIDEO

Table 4.5: Average Precision for all topics (TRECVID2002) of IDC

topic	BovW	IDC-c	IDC-s	IDC-t
75	0.0879	0.0886	0.0887	**0.0890**
76	0.0631	0.0629	0.0640	**0.0672**
79	0.0407	0.0407	0.0409	**0.0410**
80	0.0493	0.0494	0.0495	**0.0496**
81	0.0679	0.0680	0.0680	**0.0681**
82	0.0526	0.0527	0.0528	**0.0529**
84	**0.0315**	0.0303	0.0312	0.0302
85	0.0510	0.0512	0.0511	**0.0512**
86	0.1659	0.1989	0.1967	**0.2018**
87	0.0319	0.0445	0.0445	**0.0454**
88	**0.0455**	0.0438	0.0448	0.0437
89	**0.0457**	0.0430	0.0441	0.0445
90	0.1242	0.1244	0.1242	**0.1245**
91	0.0487	0.0482	0.0481	**0.0489**
92	0.0351	0.0354	**0.0354**	0.0341
93	**0.0391**	0.0376	0.0375	0.0383
94	0.0324	0.0320	0.0316	**0.0331**
95	0.0616	0.0638	0.0638	**0.0645**
96	0.0773	0.0758	0.0753	**0.0765**
97	0.0596	0.0604	**0.0605**	0.0601
98	0.0251	0.0250	**0.0253**	0.0248
99	0.3779	0.3787	0.3786	**0.3788**

MAP

The performances of IDC approaches in terms of MAP are demonstrated in Table 4.4. It shows that the IDC-t approache outperforms BovW by 0.01 and 0.03 for the two video collections respectively. The IDC weighting scheme based on temporal correlation performs better than that based on the spatial correlation, and it also improve the performance of IDC-c. This is due to that the temporal correlation may be more stable in the IDC computation over entire video collection, while the visual words tracking process has reduced some meaningless or random co-occurrence between the visual words.

Table 4.4: MAP performance of IDC

	BovW	IDC-c	IDC-s	IDC-t
TRECVID2002	0.0681	0.0694	0.0692	**0.0698**
CC_WEB_VIDEO	0.812	0.814	0.813	**0.815**

Based on the experiments of general topics video retrieval, the average Precision for all the topics has been displayed in Table 4.5. In the experimental results of TRECVID2002, IDC approach statistically significantly (P-value=0.046) outperforms the baseline on 18

Table 4.6: 5 topics IDC outperforms mostly BovW(TRECVID2002)

topic	BovW	IDC-c	IDC-s	IDC-t	Difference
Eddie Rickenbacker	0.0879	0.0886	0.0887	**0.0890**	0.002
James Chandler	0.0631	0.0629	0.0640	**0.0672**	0.004
Overview of Cities	0.1659	0.1989	0.1967	**0.2018**	0.035
Oild Field	0.0319	0.0445	0.0445	**0.0454**	0.014
Nuclear Explosion	0.0616	0.0638	0.0638	**0.0645**	0.003
Average	0.082	0.092	0.092	**0.009**	12.2%

Table 4.7: 4 topics BovW outperforms IDC (TRECVID2002)

topic	BovW	IDC-c	IDC-s	IDC-t	Difference
Price Tower	**0.0315**	0.0303	0.0312	0.0302	-0.0015
Map of US	**0.0455**	0.0438	0.0448	0.0437	-0.002
Living Butterfly	**0.0457**	0.0430	0.0441	0.0445	-0.002
live Beef	**0.0391**	0.0376	0.0375	0.0383	-0.001
Average	**0.041**	0.039	0.038	0.039	-4.9%

out of 22 topics.

On average, IDC outperforms the **BovW** 5%. Further query-by-query analysis are shown in Table 4.6 and 4.7. It is shown that IDC outperforms BovW on top 5 topics by 12.2%, and the BovW performs better than IDC on 4 topics average by 4%.

We can find out that IDC would improve performance of classical BovW framework based on above experimental results. At first, the performance improvement is statistically significant. Secondly, IDC outperforms BovW on most topics. The average performance differences ,where IDC outperforms BovW, is 0.009 and it is 0.002 where BovW outperforms IDC. However, the improvement is not very obvious. Especially, if the improvement on topic "Overview of Cities" was excluded, the average improvement is 0.004.

Based on experimental results of near duplicate video search against video collection of CC_WEB_VIDEO. The Average Precision comparison is shown in Table 4.8 and the IDC approaches perform better on 17 out of 24 topics.

The statistic significance is not adequate enough to prove the effectiveness of IDC. At first, the average improvement of IDC method to classical methods is not very big. As shown in Table 4.4, the promotion of MAP is less than 3% for both data collection. Secondly, the improvement is sensitive to the parameters. As shown in Figure 4.18, larger k_{idc} will largely reduce the MAP of IDC. Finally, the query-by-query analysis shows that IDC perform not stable in between the queries. For example, although it improves the results of some queries, it performs a lot worse than as classical BovW framework for topic 17 and 19 of data collection CC_WEB_VIDEO. In summary, the p-value of AP is larger than 10%.

The 6 topics IDC performs best is shown in Table 4.9, and the average improvement

Table 4.8: Average Precisions of all topics (CC_WEB_VIDEO)

topic	BovW	IDC-c	IDC-s	IDC-t
1	0.9607	0.9724	0.9607	**0.9719**
2	0.9768	**0.9853**	0.9848	0.9846
3	0.9187	0.9483	0.9480	**0.9733**
4	0.9293	**0.9911**	0.9837	0.9827
5	0.9797	**0.9909**	0.9888	0.9858
6	0.6178	0.7529	0.7497	**0.7583**
7	0.7262	**0.8218**	0.8099	0.7929
8	**0.5657**	0.5533	0.5377	0.5114
9	**0.6622**	0.6099	0.6045	0.5941
10	0.9683	0.9749	**0.9752**	0.9724
11	0.7216	0.7154	**0.7638**	0.7128
12	**0.7939**	0.7524	0.7317	0.7531
13	0.9784	**0.9975**	0.9962	0.99562
14	0.7793	**0.8858**	0.8793	0.8816
15	0.6848	0.7130	0.7105	**0.7213**
16	**0.9868**	0.9859	0.9751	0.9722
17	**0.9422**	0.6567	0.6376	0.6588
18	0.4475	**0.4886**	0.4850	0.4359
19	**0.8252**	0.7019	0.7010	0.7251
20	0.8384	0.8924	**0.8831**	0.8334
21	0.7873	**0.8128**	0.7865	0.8116
22	0.5907	**0.6040**	0.5961	0.5856
23	0.5479	**0.6485**	0.6057	0.5693
24	0.9910	**0.9921**	0.9921	0.9858

Table 4.9: APs of 6 topics IDC outperforms mostly BovW (CC_WEB_VIDEO)

topic	BovW	IDC-c	IDC-s	IDC-t	Difference
"the urban ninja"	0.6178	0.7529	0.7497	**0.7583**	0.141
"kingdom hearts 2"	0.7262	**0.8218**	0.8099	0.7929	0.067
"little indian dancing kid"	0.7216	0.7154	**0.7638**	0.7128	0.052
"ronaldinho ping pong"	0.7793	**0.8858**	0.8793	0.8816	0.106
"2 pac-changes"	0.8384	0.8924	**0.8831**	0.8334	0.045
"shakira hips don't lie"	0.5479	**0.6485**	0.6057	0.5693	0.101
Average	0.705	0.786	0.782	0.758	11.5%

Table 4.10: APs of 5 topics BovW outperforms IDC (CC_WEB_VIDEO)

topic	BovW	IDC-c	IDC-s	IDC-t	Difference
"free hugs campaign"	**0.5657**	0.5533	0.5377	0.5114	0.054
"where the hell is matt"	**0.6622**	0.6099	0.6045	0.5941	0.068
"napoleon dynamite"	**0.7939**	0.7524	0.7317	0.7531	0.04
"i write sins not tragedies"	**0.9422**	0.6567	0.6376	0.6588	0.294
"sony bravia tv ad"	**0.8252**	0.7019	0.7010	0.7251	0.10
Average	0.758	0.655	0.649	0.651	13.3%

is 11.5%. The result of worst 5 topics is shown in Table 4.10, and on average, BovW outperform 13.3% than IDC. Obviously, we can not conclude the effectiveness of IDC based on the current experimental results on data collection CC_WEB_VIDEO.

The unstable retrieval performance can be caused by that IDC may over-weight some visual words and under-weight other visual words. As shown in Table 4.3, the IDC weights of TRECVID2002 has smaller variance than data collection CC_WEB_VIDEO. If we compare performances of the IDC methods for two data collection shown in Table 4.5 and Table 4.8, we can easily found that IDC performs more stable for TRECVID2002, and the p-values comparison (0.04 vs 0.11) also indicates the same scenario. The comparison shows that larger variance will involve in retrieval risk.

It is worthy to point out that the topic "i write sins not tragedies" looks like an extreme point, because only on this topic BovW outperforms IDC more than 0.1. If this topic was excluded, the average decreased performance is -7%, which is less than the average performance difference shown in Table 4.9. There is not such extreme points in the better results of IDC shown in Table 4.9. In summary, the effectiveness of IDC is not fully supported by the result, but the results have still revealed some possibility that performance of BovW may be improved through STC discovered from video collection, for example, idea similar to IDC.

The AP of best performing queries is 0.082 on TRECVID2002, which is nearly double as poor queries (AP = 0.041). But, APs of the best and worst queries for IDC are not different

in the other data collection CC_WEB_VIDEO. The best queries are normally persons and general scenario, and the poor queries are normally specific visual information. A possible explanation is that the IDC discovered from diversified information may be more useful for retrieval.

Considering the above evaluations based on both Precision-Recall curve and MAP criteria, the experimental results in this section have shown the effectiveness of IDC approaches. We can conclude that the improvement is not stable and the statistical significance is not large. The overall improvement of IDC approaches is not significant as the QC methods when we compare the performance displayed by the Table 4.2 and 4.4. However, we have at least opened a door in the direction that characterizing descriptive visual words via spatial-temporal correlation discovering from entire video collection could improve the BovW model for the CBVR technology development.

4.5 Summary

In this Chapter, we have proposed two novel term weighting schemes based on the different levels of spatial temporal correlation (STC) for the CBVR retrieval model, both of which reformulates the representation of the videos under the BovW model. The schemes are based on two defined concepts QC and IDC.

At first, the STC discovered from a query video is defined as Query Correlation (QC). It is used to charactorize the descriptive visual words in the similarity measurement, which is based on an assumption that the visual word with a higher QC is more descriptive. Furthermore, we developed a method to emphasize the weights of descriptive visual words and then reformulate the query representation. The similarity measurement function can be directly implemented based on this characterized QC weights for the CBVR technology.

Secondly, the STC discovered from a video collection is defined as Documents Correlation (DC). Inspired by the Inverse Document Frequency (IDF), the higher DC of individual visual word is assumed to be an indication that the corresponding visual word is less descriptive. Based on this hypothesis, a novel weighting scheme, namely Inverse Documents Correlation (IDC), has been proposed for the visual words to reformulate the representation. In this way, the spatial and temporal information discovered from the video collection is also utilized in the established retrieval function.

A series of experimental results on the near-duplicate web video search and general topic video retrieval tasks show that the QC weighting approaches substantially improve the classical BovW model without increasing storage cost for video representation. The QC weighting approach has also outperformed a state-of-the-art TGC-based approach on challenging tasks. We can conclude that the descriptive visual words can be characterized and emphasized based on the QC, and this strategy effectively compensates the neglection of the spatial-temporal information and the information loss during the key-frame

sampling.

The experimental results also show that the IDC could improve the performance of the BovW model for the CBVR tasks. The improvement for a video collection is not statistically significant enough. There is a need for further evidences to support the assumption to adjust the weights of visual words based on the computed IDC weights. Our evaluation has at least opened the door for future research in this direction.

Another future research direction is to investigate on the exploration of context knowledge, e.g. text information or users' feedback, with the spatial-temporal information to improve the descriptive ability of visual words. Furthermore, some optimization technologies could be explored to speed up the process of discovering descriptive visual words with STC.

In conclusion, it has been proven that the STC extracted from both the query video and the video collection can be incorporated into the retrieval function. The experimental results have shown its substantial effectiveness to improve the classical BovW framework on two types of CBVR tasks. Besides the representation and the similarity measurement function, another important step in the CBVR technology is the building of the visual vocabulary. How to utilize the STC to improve this process will be discussed in the next Chapter.

Chapter 5

Rebuilding the visual vocabulary based on STC

This chapter presents our approach, utilizing the discovered spatial and temporal information, to improve a key procedure of the Bag-of-visual-Words (BovW) model. We are aiming to rebuild a proper visual vocabulary, which is originally built by clustering a number of visual features (*e.g.* using K-means, GMM, etc). Two types of errors may occur in the building process. These will be referred to as the "UnderQuantize" and "OverQuantize" problems in this chapter. The former problem causes ambiguities and often leads to false visual content matches, and the latter generates synonyms and may result in losing true relevances. Unlike most state-of-the-art methods which concentrated on disambiguating the visual words, this chapter aims to address the "OverQuantize" problem by leveraging spatial and temporal context similarity between the visual words.

The context of a visual word is defined to its spatially/temporally co-occurring visual words, and then the STC model developed in Chapter 3 provides a computing framework to model this context. We assumed that the semantic of a visual word is not only determined by its position in the feature space, but it can also be detected or verified according to the characterized context. Based on this hypothesis, the visual words, which always appear in a similar context, are detected as synonyms caused by the quantization errors. In this Chapter, the detected synonyms are assumed to be the redundancy of the initial visual vocabulary.

These synonyms detected from the initial visual vocabulary are then merged to form a new vocabulary. We expecte that the new visual vocabulary would be more compact and descriptive. This scheme is evaluated on the TRECVID2002 and CC_WEB_VIDEO video collections for two typical Query-By-Example (QBE) video retrieval applications. Experimental results demonstrate substantial improvements in retrieval performance over the classical BovW model. We also show that our approach can be utilized in combination with a state-of-the-art disambiguation method to further improve the performance of the

CBVR.

5.1 Quantization Errors of Visual Vocabulary

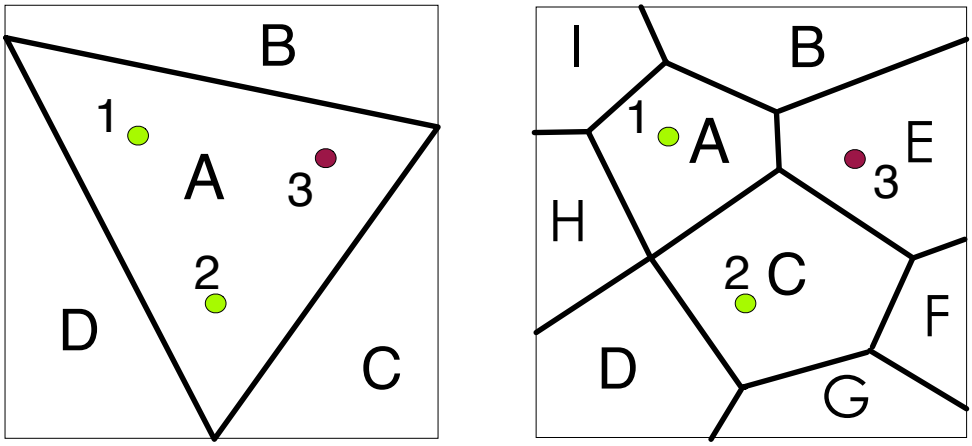
As was reviewed in Chapter 2, to efficiently index the local features, an unsupervised clustering method is always applied to a set of training features, and each feature is quantized to its nearest visual word. Although significant progress has been made, the visual words are not always as effective as textual words in information retrieval. One of the possible explanations is that the textual words always have relatively concrete semantics, but the meaning of a clustered visual word is not as stable. This is mainly because of two types of quantization errors which may hamper the retrieval.

The first error tends to occur when the quantization is too rough, and visual features with different meaning may be quantized into the same visual word. The ambiguous visual words will result in false visual content matches. In the present Chapter, we refer to this type of quantization error as "UnderQuantize". As can be seen in Figure 5.1, the feature space is shown as a square and the feature points with identical meaning are identified by the same color. When the size of the vocabulary K equals 4, then all features 1, 2, and 3 are quantized into the same visual word A in Figure 5.1(a). This "UnderQuantize" error is more likely to occur when a small vocabulary is used.

A larger sized vocabulary may help to disambiguate the visual words, and then the granularity of individual visual words becomes finer. As shown in Figure 5.1(b), the size of vocabulary increases from 4 to 9, and cells become smaller than those in Figure 5.1(a). This ensures distinction between features 1 and 3. However, increasing the vocabulary size to address the "UnderQuantize" problem could bring in the second type of quantization error, where visual features representing similar visual information will be quantized into different visual words. It is referred to as "OverQuantize" error, which would cause a "synonyms" problem. As shown in Figure 5.1(b), the features 1 and 2 are quantized into visual words A and C respectively. This "OverQuantize" error would cause a loss of the relevant visual information in the retrieval process, which would also hurt the performance of the BoVW model.

In previous research, intensive investigations were devoted to reducing the "UnderQuantize" error in order to achieve a more accurate visual content match. Recently, large vocabularies (Nister & Stewnius 2006), (Philbin et al. 2007) have often been used to address the "UnderQuantize" problem, and additional contextual information (Zhang et al. 2010) was utilized to further verify the matched visual words. However, there is a trade-off between the two types of errors, i.e., the smaller "UnderQuantize" error, and the bigger "OverQuantize" error.

The "OverQuantize" problem has attracted much attention in the past years. Philbin et al. (Chum et al. 2007) claimed that a major drawback of the BovW-based visual infor-



(a) K= 4. (b) K=9.

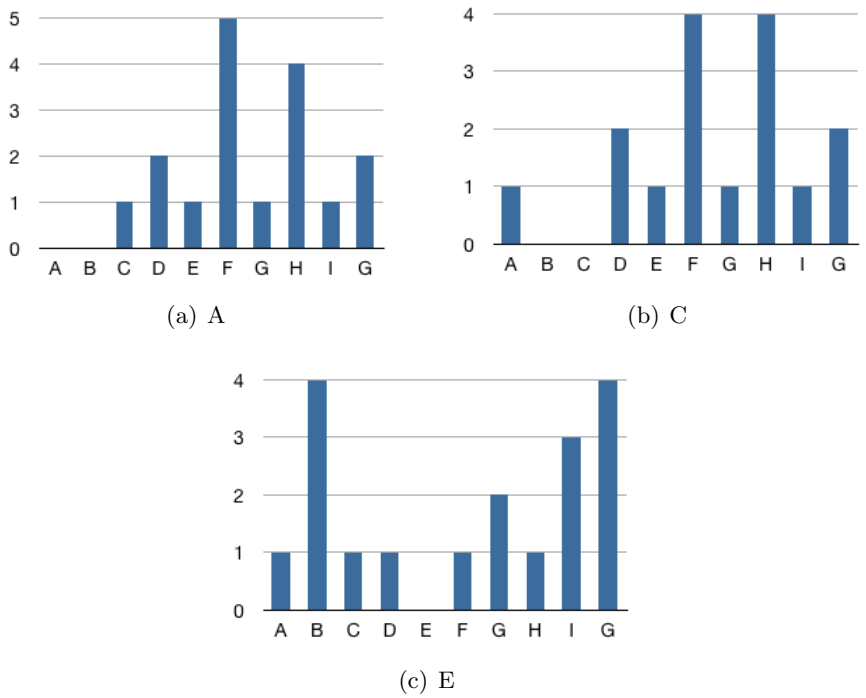
Figure 5.1: Two types of quantization error. K is the size of visual vocabulary.

mation retrieval is that it may miss some relevant information. Clearly, the "OverQuantize" error is one of the key factors contributing to the lost relevances. Jiang et al. (Jiang & Ngo 2009) have proposed to softly quantize a feature into multiple visual words to reduce the risk of missing relevant information. Another type of soft-matching equivalent schemes (Chum et al. 2011) was proposed to expand the query with the synonymous and related visual words to promote performance and indirectly tackle this problem.

Our work focuses on addressing the "OverQuantize" problem. The main idea is to integrate multiple synonymous visual words into an single visual word, and then rebuild a more compact and descriptive visual vocabulary. For example, if the visual words A and C in Figure 5.1(b) are merged, then features 1 and 2 can be correctly matched. Because the appearance of features has been used by the initial visual vocabulary to classify the meaning of visual words, extra context information should be used to detect the synonyms.

Building a contextual visual vocabulary has also attracted much attention. For example, the context aware visual word clustering methods proposed by (Yuan & Wu 2008) and (Wang, Yuan & Tan 2011) directly utilized the co-occurrence information of visual features in the regularized clustering. However, this regularized and iterative clustering algorithm can not be easily generalized to the unseen data. Our method detects and merges the synonyms in the initial visual vocabulary, and the generalization is as straightforward as the classical BovW framework.

We aim to utilize the context of visual words to address the problem of "OverQuantize". Here, we define the context of an individual visual word as its spatially or temporally co-occurring visual words. It is assumed that the visual words with similar meaning tend to occur in a similar context. For example, if the visual features representing "nose" are "OverQuantize"-ed into two different visual words by the BovW model, then, they both may still tend to appear in the same context, such as visual words representing "eyes",



(a) A

(b) C

(c) E

Figure 5.2: The histogram of co-occurring visual words of A, C, and E

"ears", and "mouth". A simple example can be illustrated by Figure 5.2. Both visual words A and C often co-occur with F and H, whilst at the same time E always co-occurs with B and I. Therefore, A and C are more likely to be synonyms and E is more likely to be different from them.

As was discussed in Chapter 4, the video content is always a mixture of large amounts of miscellaneous information, and the context of a visual word can be defined based on the STC model extracted from video collections. The most straightforward method is to count the number of the frames, in which it co-occurs with other visual words. Inspired by the effectiveness of STC to improve the CBVR retrieval model, this method is extended to define the spatial and temporal context of visual word.

A context vector (as can be seen in Figure 5.2) is then formulated for each visual word, w, in a video collection, and each entry of the vector represents the weight of a corresponding visual word that co-occurs with w. Thus, the context similarity between a pair of visual words is (inversely) measured by the distance between their context vectors.

Nevertheless, the original feature information (i.e., appearance) of visual words should not be completely ignored in determining the similar visual words. It can be used for the appearance similarity verification. For example, in Figure 5.1(b), A and G are far from each other in the feature space, and they may not be considered as synonyms, even though their contexts are similar.

In real world applications, the visual vocabulary could also be constructed with mul-

tiple features, either by concatenating multiple features into a long vector before the quantization (Hsu & Chang 2005) or combining the visual words after quantizing the different features separately (Zhang, Liu, Ouyang, Lu & Ma 2009). However, we only exploit vocabulary based on a single feature (SIFT) to demonstrate the effectiveness of our context similarity model and rebuilding approach, and it can be easily applied to the visual vocabulary based on other single or multiple features.

The framework of our method is shown in Figure 5.3. At first, an initial visual vocabulary is generated by clustering the training features. All features are then quantized as visual words according to the initial vocabulary. A context matrix representing the cross relationships between visual words is computed based on the spatial-temporal information. Furthermore, similar visual words are selected according to the context similarity computation. Finally, we merge these visual words after an appearance verification, and a more compact visual vocabulary is then formed. The details of each procedure will be presented in the following sections.

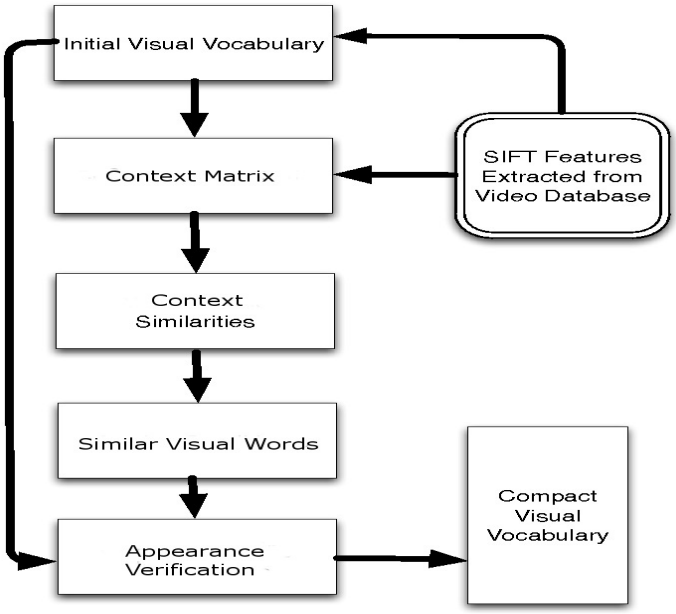


Figure 5.3: Proposed framework for the vocabulary rebuilding.

5.2 Spatial-Temporal Context of Visual Words

The initial visual vocabulary is generated by the approximate K-Means clustering algorithm (Philbin et al. 2007), which is able to build a large scale vocabulary. To focus our research on the "OverQuantize" problem, the initial visual vocabulary has been selected to be large enough, and the size K of the visual vocabulary could be as large as tens of thousands.

Recall that in a classical BovW framework, each video in a video collection D is represented as a sequence of key frames $v_d = \{\boldsymbol{f}_d\}$. In the query-by-example (QBE) video retrieval, the key frame similarity $sim(\boldsymbol{f}_d, \boldsymbol{f}_q)$ can be measured by a Cosine function, which is approximated by Equation 4.7. The video similarities are formulated based on this frame level similarity, which is applied to the key frames sampled from the query and video document.

As discussed in Section 5.1, the context of a visual word is defined as the co-occurring relationship with other visual words. It means that the features appearing in the neighborhood of a visual word should be recorded to represent its context as shown in Figure 5.4. When we discuss the context of a visual word, it is named as center visual word in this section.

It has been illustrated in Chapter 3 that the co-occurrence matrix quantitatively captured relationships between the visual words co-appearing within a frame. The defined context of a center visual word can be characterized by a corresponding row vector of this matrix.

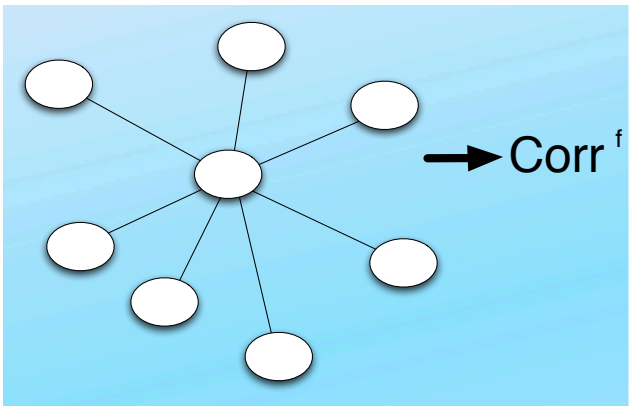


Figure 5.4: Context of a center visual word in a frame could be modeled by its neighboring visual words.

Furthermore, the STC of visual words extracted from a single frame may be not adequate for the context similarity measurement. Thus, we can utilize the STC extracted from the entire video collection. The context of a visual word is defined based on the correlation matrix modelled for the video collection:

$$ctxt_i^c(j) = \boldsymbol{Corr}^d(i, j) \tag{5.1}$$

where the $ctxt_i^c(j)$ represents the co-occurrence level correlation. A context vector is then formulated considering all other visual words, and each entry is calculated by Equation 5.1. In this way, the context vector is in the same form as a raw vector of the video collection level co-occurrence matrix:

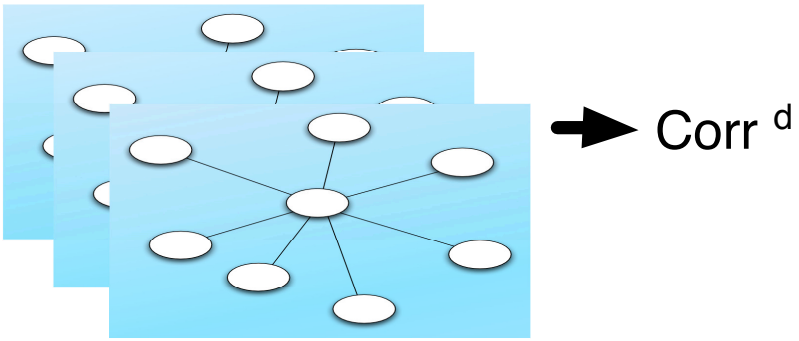


Figure 5.5: Context of a visual word in the video collection could be modeled as a correlation vector.

$$\boldsymbol{Ctxt}_i = \boldsymbol{Corr}^d(i) \tag{5.2}$$

where $\boldsymbol{Ctxt}$ is a K dimensional vector and $\boldsymbol{Corr}^d(i)$ is the i^{th} raw vector of the matrix $\boldsymbol{Corr}^d$.

Practically, a visual word always co-occurs with a variety of visual information contained in the frame, and the different visual words should not be equally related to its context. It is also reasonable to measure the relation based on the spatial proximity. Closer visual words should contribute more to the context of the central visual word. It has been demonstrated that correlation based on spatial proximity can be modelled using the STC matrix. Following this approach, the spatial proximity context of a visual word can be formulated as follows:

$$ctxt_i^s(j) = \boldsymbol{Corr}_s^d(i,j) \tag{5.3}$$

Similarly, a spatial context vector is constructed by aligning the elements computed by Equation 5.3 as follows:

$$\boldsymbol{Ctxt}_i^s = \boldsymbol{Corr}_s^d(i) \tag{5.4}$$

Note that the video is a sequence of temporally aligned frames, and a visual word also has the temporal context. The visual words, which moved more coherently with the central visual word, are assumed to contribute more to its temporal context. Similar to the spatial context, the temporal context is then also modelled by a temporal correlation matrix. The context computation is formulated as:

$$ctxt_i^t(j) = \boldsymbol{Corr}_s^d(i,j) \tag{5.5}$$

Given Equation 5.5. We can use it to derive the temporal context vector for a visual

word:

$$\boldsymbol{Ctxt}_i^t = \boldsymbol{Corr}_t^d(i) \tag{5.6}$$

where $\boldsymbol{Corr}_t^d(i)$ is the i^{th} raw vector of the temporal correlation matrix.

Furthermore, to make the modelled context vectors comparable between the different visual words, the vector $\boldsymbol{Ctxt}_i$ is normalized as:

$$\boldsymbol{Ctxt}_i = \frac{\boldsymbol{Ctxt}_i}{\parallel \boldsymbol{Ctxt}_i \parallel_1} \tag{5.7}$$

where $\parallel * \parallel_1$ denotes the 1^{st} order Norm of the vector. The co-occurrence context $\boldsymbol{Ctxt}^c$, spatial context $\boldsymbol{Ctxt}^s$, and temporal context $\boldsymbol{Ctxt}^t$, can all be normalized based on Equation 5.7. After the normalization, each element of the row vector $\boldsymbol{Ctxt}_i^c$ indicates the degree of probability that the i^{th} visual word co-occurring with other visual words.

In the next section, we will discuss our method in measuring the visual words context similarity to rebuild a visual vocabulary.

5.3 Context Similarity and Rebuilding the Visual Vocabulary

As discussed in Section 5.2, the synonymous visual words will be detected by examining the context similarity between visual words. The context of a visual word has been modelled as a K dimensional vector. The context similarity between visual words can be inversely measured by the distance between the correlation vectors. The distance d^c between two co-occurrence context vectors is formulated as:

$$d_D^c(i,j) \approx \sum_{k=1}^{K} |C_i^c(k) - C_j^c(k)| \tag{5.8}$$

where Manhattan Distance $|C_i^c - C_j^c|$ is used as an example, and other distance metrics can also be used for the context similarity measurement. Similarly, the distances d^s and d^t, which are based on the spatial correlation and temporal correlation respectively, can also be computed using Equation 5.8.

The larger the $d(i,j)$ is, the more similar the two visual words are in terms of the corresponding contexts. Thus, a set of pair-wise synonyms $W_s = \{(w_i, w_j)\}$ is selected. The selection procedure can be formulated as follows:

$$(w_i, w_j) \begin{cases} \in W_s & \text{if} \quad d_D(i,j) < \epsilon \\ \notin W_s & otherwise \end{cases} \tag{5.9}$$

where d_D can be replaced by either d_D^c, d_D^s or d_D^t, which would consider co-occurrence, spatial or temporal context respectively. The parameter ϵ is an empirically selected thresh-

old to determine the visual words which should be merged. In our experiment, ϵ is set to a range between 0.1 and 0.5, and the results will be demonstrated in Experiments section.

In addition, we utilize the appearance of the visual words for the similarity verification. The visual words close in the feature space should represent similar visual appearances. A set of nearest visual words $N_i = \{w_n\}$ to the visual word w_i is detected using the K-d tree algorithm. When the w_j does not belong to the set N_i, it is not allowed to be merged with w_i. This principle guarantees that the merged synonyms are not only similar with respect to the context, but also in their visual appearances. Based on this arrangement, the representation of the frames can be updated according to the rebuilt visual vocabulary. For example the term frequency of w_i can be updated as follows:

$$tf'(w_i) = \begin{cases} tf(w_i) + tf(w_j) & \text{if } (w_i, w_j) \in W_s \ \& \ w_j \in N_i \\ f(w_i) & otherwise \end{cases} \tag{5.10}$$

, and the other term specifies such as IDF can be updated accordingly.

If we denote the number of merged pair-wise visual words by K_m, and the initial visual vocabulary is pruned as a new visual vocabulary. Thus, the size of rebuilt visual vocabulary becomes $K - K_m$. The similarity between the query example and a data video is then measured based on the updated frame representation as:

$$sim_{v_d, v_q} = \max_{\boldsymbol{f}'_d \in v_d, \boldsymbol{f}'_q \in v_q} sim(\boldsymbol{f}'_d, \boldsymbol{f}'_q) \tag{5.11}$$

5.4 Experiments

The main objective of our experiments in the present chapter is to evaluate the effectiveness of our ocabulary rebuilding approach to detect the synonyms, which are caused by the "OverQuantize" problem. Hessian detector (Mikolajczyk et al. 2005) and the SIFT descriptor (Lowe 2004) are utilized for visual features extraction. The initial visual vocabulary is clustered by the K-means algorithm and visual features are quantized into the nearest cluster centroid and mapped to corresponding visual word.

The rebuilt visual vocabulary is evaluated by the performance of the Query-by-Example video retrieval on two different commonly used video collections. The classical BovW model is used as a baseline, which is denoted by **BovW**. The performances of three variations of visual vocabulary rebuilding approaches proposed in this Chapter are reported, namely the co-occurring, spatial, and temporal correlation based approaches (Equation 5.11), denoted by **c-corr** , **s-corr**, and **t-corr** respectively.

Furthermore, we also evaluated the effectiveness of our approach to improve the visual word disambiguation methods. The effectiveness is also referred to as the compatibility between our approaches and disambiguation methods. We choose Tight Geometric Constraint (**TGC**) (Zhao et al. 2010) as an example to be in combination with the rebuilt

visual vocabularies.

5.4.1 Experimental Set-Up

Two commonly used video retrieval datasets are also selected for the experiments:

(1) `TRECVID2002` The dataset TRECVID2002 (Smeaton et al. 2006) is selected to perform the general topic video retrieval.

(2) `CC_Web_Video` Near-duplicate video search is performed on video collection CC_Web_Video (Wu, Ngo, Hauptmann & Tan 2009).

5.4.2 Experiment 1: General Topics QBE Video Retrieval

The relevance in the general topics QBE video retrieval is defined at the concept level. It is a challenging task, and the scale of initial vocabulary in this experiment is selected to be 5K, which is relatively small and suitable for this task.

Parameters

As shown in Figure 5.6, our approach merges a number of visual words based on the similarities of the co-occurrence, spatial, and temporal context modelled for corresponding pair-wise visual words. The number of merged visual words K_m increases proportionally to value of the threshold ϵ defined for context similarity measurement in Equation 5.10. This observation fulfills our theoretical analysis.

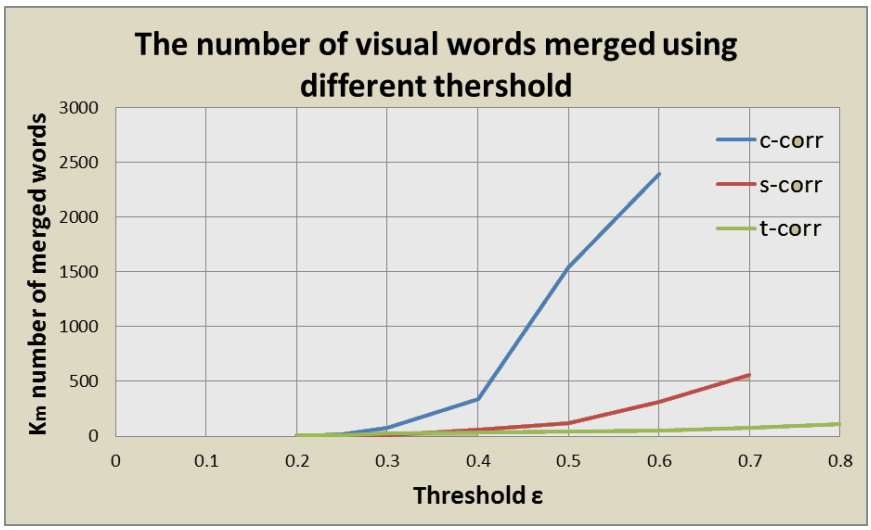


Figure 5.6: The number of merged visual words generated based on different threshold for TRECVID2002 video collection.

It is also clear that with increasing ϵ, the K_m of the **c-corr** increases much faster than that of the **s-corr** and the **t-corr**. This means that more visual words are similar with

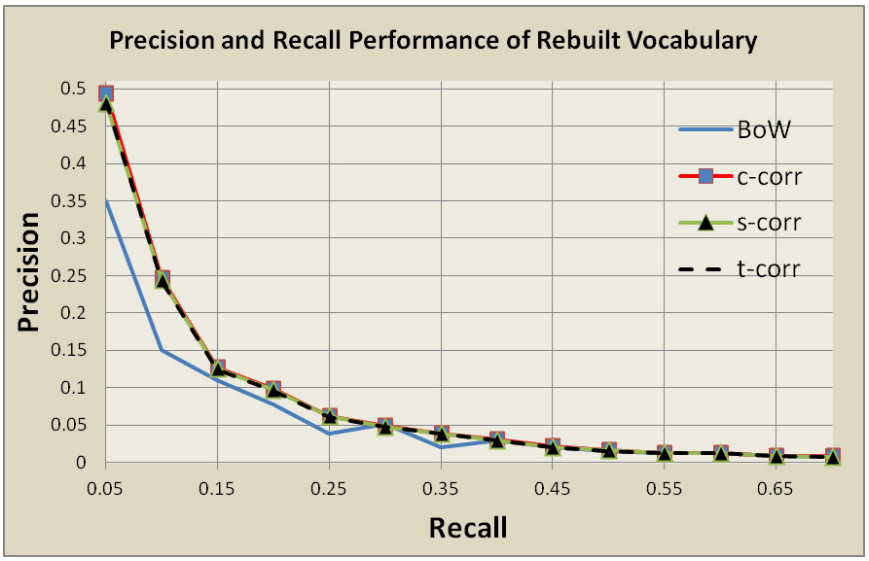


Figure 5.7: Overall Precision-Recall performance of rebuilt vocabulary for general topics QBE video retrieval (TRECVID2002).

respect to the co-occurring context, and few of them are similar in the spatial-temporal context. This result may be due to the fact that fewer visual words are related to the context of a center visual word, when the context is narrowed by spatial or temporal constraints.

Intuitively, different selected values of the parameter ϵ would lead to different performances of the rebuilt visual vocabularies. The performances difference would be demonstrated in the following experimental results in the form of MAP.

Improving the Retrieval Model

The overall performance in terms of Precision-Recall curve on TRECVID2002 is shown in Figure 5.7. It can be seen that the rebuilt visual vocabulary generally outperforms the original visual vocabulary. It is shown that at the Recall points ranging from 5% to 35%, the corresponding Precision of **c-corr** is higher than **BovW**, which means that more relevances are ranked higher. Thus, the **c-corr** seems to compensate the missing relevances by the classical BovW model in this experiment. Furthermore, while the differences in performance between **c-corr**, **t-corr**, and **s-corr** are not great, the three methods all outperforms the baseline **BovW**.

In order to capture influence of various ϵ (Figure 5.6) on the retrieval performance of the visual vocabulary rebuilding model, we compare the performances of visual vocabularies, which are rebuilt by merging different numbers of visual words. Figure 5.8 illustrates the performances of a variety of rebuilt vocabularies in terms of MAP criteria. The performances are arranged according to the number of merged visual words K_m.

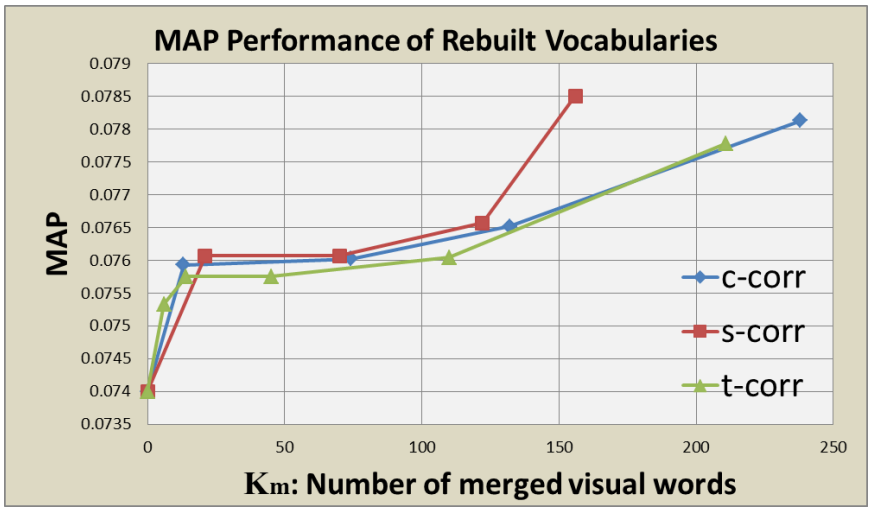


Figure 5.8: MAP performance of rebuilt vocabularies merging different numbers of visual words.

It can be seen from Figure 5.8 that the performances of the methods **c-corr**, **t-corr**, and **s-corr** generally arise along with the increasing of K_m. The best performance of the three methods is achieved by merging 150-250 out of 5K visual words. However, it does not necessarily imply that merging more visual words would always result in better performance. It shows a trade-off that too much merged visual words will lead to "UnderQuantization" error. For example, if we increase the K_m up to 1538, the MAP of **t-corr** will drop to 0.06891 from the best performance 0.0781. It can be concluded that, using the detection models, around 3%-5% visual words are "OverQuantize"-ed during the original quantization process of the BovW model for the TRECVID2002.

It is also clear from Figure 5.8 that the performance of **s-corr** is better than the **c-corr** and **t-corr**. This shows that the spatial context of visual words is more meaningful than the simple co-occurring correlation. Furthermore, **s-corr** outperforms **t-corr** with respect to the MAP criteria. For video collection TRECVID2002, we can preliminarily conclude that the discovered temporal relationship across visual words between consecutive keyframes may not be so effective as the spatial correlation. The detailed discussion about this topic will be demonstrated in Chapter 6.

The topic-to-topic average Precision comparison is demonstrated in Table 5.1. The **s-corr** outperforms **BovW** on more topics (16 out of 22) and on average, **s-corr** outperforms the initial vocabulary by 5.4%, which is statistical significant (P-value = 0.047). These results clearly demonstrate that the rebuilding process improves the initial vocabulary and partially solves the "OverQuantize" problem.

Table 5.1: Average Precision Comparison for TRECVID2002

AP	BovW	c-corr	s-corr	t-corr
75	0.0677	**0.0951**	0.0938	**0.0951**
76	0.0471	0.0633	0.0583	**0.0639**
79	0.0334	**0.0359**	0.0341	0.0356
80	0.0469	0.0471	**0.0483**	0.0472
81	**0.0718**	0.0681	0.0685	0.0681
82	**0.0384**	0.0316	0.0328	0.0319
83	0.0576	0.0634	**0.0634**	0.0607
85	0.195	0.194	**0.206**	0.1950
86	0.0444	0.04566	**0.04720**	0.04348
87	0.0397	0.05016	**0.0504**	**0.0504**
88	0.0544	**0.0564**	0.0545	0.0542
89	0.1231	**0.1250**	0.1245	0.1248
90	**0.0447**	0.0372	0.0356	0.0402
91	0.0783	0.0902	0.0902	**0.0904**
92	**0.0389**	0.0345	0.0378	0.0344
93	**0.0417**	0.0377	0.0383	0.0366
94	0.0315	0.0297	**0.0329**	0.0291
95	0.0681	0.0682	**0.0689**	0.0684
96	0.0626	0.0642	0.0641	**0.0646**
97	**0.0791**	0.0717	0.0639	0.0664
98	0.0255	0.0311	**0.0345**	0.0317
99	0.337	0.3788	0.3787	**0.3789**

Compatibility with TGC approach

Furthermore, to test whether or not the rebuilt vocabulary is compatible with the other enhancement approach for the BovW framework, we incorporated it into the **TGC** approach as an example. **TGC** is one of the common techniques used to disambiguate visual words with additional geometric information constraints. It has been proven to be effective to promote the accuracy of visual content similarity measurement.

As shown in Figure 5.9, **TGC** has an advantage that it could promote the average Precision of the top ranked results. However, the additional constraint of similarity measurement deteriorates the "OverQuantize" problem, and relevances are even more likely to be missed as shown by the relative lower Recall at tail of the curve. When we compare the performance shown in Table 5.2 and Table 5.1, the **TGC** does not performs so good as the **BovW** in terms of MAP. The "OverQuantize" problem does also exist for the TGC method.

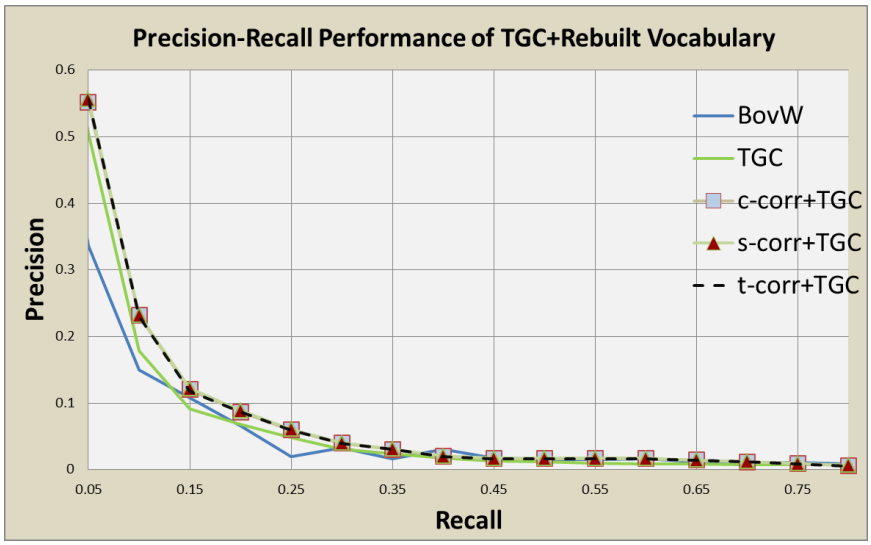


Figure 5.9: Precision-Recall performance of TGC approaches based on the rebuilt vocabularies for TRECVID2002

To evaluate the compatibility of the rebuilt vocabularies with the **TGC** method, we tested the **TGC** approach based on the videos and queries representation indexed according to the rebuilt visual vocabularies. The overall Precision-Recall curve is presented in Figure 5.9, and **c-corr**, **s-corr**, and **t-corr** outperforms the baselines **TGC** and **BovW**. The rebuilt vocabularies maintain high Precision of the **TGC** for the top ranked results, and also improved the performance on the middle Recall points.

As can be seen in Table 5.2 which presents an MAP performances comparison between **c-corr**, **s-corr**, and **t-corr**. **s-corr** performs better than the other two, however all three methods outperform the original **TGC** approach in terms of MAP. In comparison with

Table 5.2: MAP Comparison for TRECVID2002

	TGC	c-corr +TGC	s-corr +TGC	t-corr +TGC
MAP	0.0589	0.0716	**0.0720**	0.0710

the performances shown in Table 5.1, the **s-corr+TGC** performs comparable with the **BovW**. Moreover, the **s-corr+TGC** approach has a great advantage that it has higher Precision on top ranked retrieval results. The rebuilt vocabulary partially overcomes the drawback of the **TGC** approach. We conclude that the rebuilt vocabularies based on context similarity are compatible with the TGC approach.

As shown in Figure 5.10, when more visual words are merged, the performance of TGC is better. This observation is similar to the results shown in Figure 5.8. It shows that reducing redundancy in the initial visual vocabulary not only works for the classical BovW framework, but could also improve the performance of the TGC. It seems that fewer visual words are redundant for the TGC method. The corr+TGC methods perform stable in terms of MAP, when more than 100 visual words are merged. It may be because of that the TGC utilizes additional information to disambiguate the visual words in the similarity measurement.

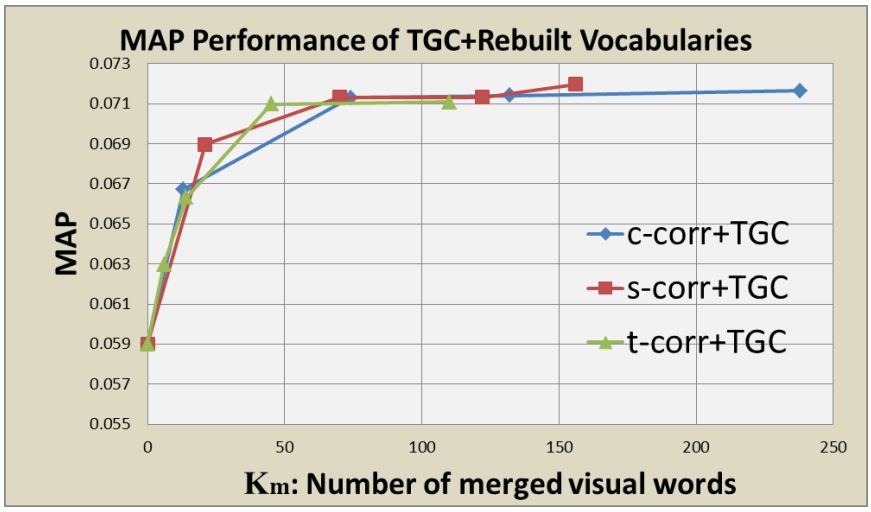


Figure 5.10: MAP performance of TGC approaches based on rebuilt vocabularies merging different number of visual words.

In summary, the evaluation on the QBE general topic video retrieval task shows that the rebuilt visual vocabulary based on the spatial-temporal context not only effectively improves the performance of the classical BovW model, but is also compatible with other BovW model enhancement methods such as the **TGC**.

5.4.3 Experiment 2: QBE Near-Duplicate Video Search

The near duplicate video search is another important application of QBE-CBVR technology. The search targets are always nearly identical to the queries. There is always a high degree visual similarity between the query example and the relevances, and it requires more accurate match of visual features. As a result, a larger size (20K) visual vocabulary is used to achieve higher visual content matching accuracy. Furthermore, another challenge is that the queries used are not always of high quality as introduced in Chapter 5. The robustness of the retrieval model with queries of various qualities is important for this application.

The overall Precision-Recall curve is shown in Figure 5.9, and it demonstrates that **c-corr** outperforms **BovW**. The curves associated with **c-corr**, **s-corr**, and **t-corr** are not very different with respect to the Precision-Recall criteria, and all rebuilt vocabularies outperform the initial visual vocabulary.

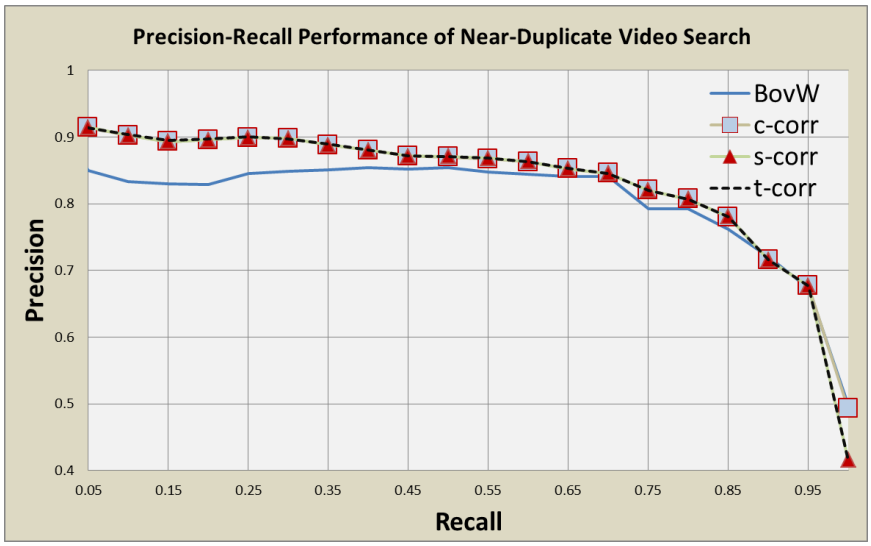


Figure 5.11: Precision-Recall performance of based on rebuilt vocabulary based

The performances comparison regarding MAP is displayed in Table 5.3. The first column shows the different values of parameter ϵ used in Equation 5.10. Firstly, it is shown that **t-corr** slightly outperforms the other two rebuilt vocabularies. This may be due to the fact that camera movements in the web videos are often slower than the professional videos, such as movies and advertisements, because the web videos are often recorded by simple instruments. The discovered temporal relationships may be more meaningful. Secondly, it has been pointed out in Section 5.4.1 that the larger the ϵ value is, the more visual words are merged. In this experiment, the best performance is achieved at $\epsilon = 0.2$. It is also shown that merging more visual words may not necessarily result in a better performance. For example, vocabularies based on $\epsilon = 0.5$ performed worse than

Table 5.3: MAP Comparison of rebuilt vocabularies generated by difference merging thresholds for CC_WEB_VIDEO

ϵ	c-corr	s-corr	t-corr
0	0.8010		
0.1	0.8243	0.8246	0.8346
0.2	**0.8295**	**0.8293**	**0.8364**
0.5	0.8181	0.8257	0.8360

Table 5.4: MAP Comparison of rebuilt vocabularies for different queries for CC_WEB_VIDEO

	BovW	c-corr	s-corr	t-corr
Exact	0.9395	0.9451	0.9451	**0.9454**
Similar	**0.9111**	0.9078	0.9082	0.9092
Major Changed	0.5498	0.6446	0.6446	**0.6448**

$\epsilon = 0.2$. Empirically, in this experiment, the number of merged redundant visual words should be around 40-70 (out of 20K visual words), which is less than the 150-200 visual words merged by visual vocabularies of the best performance in the general topic video retrieval experiment. In the near duplicate video search application, the relevances have high degrees visual similarity, and explicit visual content description always lead to more accurate similarity prediction and better retrieval performance. As a result, it is always necessary to build a larger visual vocabulary. Due to the same reason, fewer visual words should be considered as redundancy.

Table 5.4 presents the performances of different queries. It is shown that the queries grouped as "*Major Changed*" are harder than the other two. The rebuilt vocabularies outperforms initial vocabulary when "*Major Changed*" queries are running, whilst performs comparably with **BovW** for the other two. It is shown that the rebuilding model has improved the ability of the BovW model to deal with hard queries and maintained the good retrieval performance on the easier queries.

The topic-to-topic comparison regarding Average Precision is presented in Table 5.5, where ϵ is selected as 0.2 to demonstrate the best performances. Rebuilt vocabulary outperforms baseline **BovW** (P-value=0.0234, 0.0200, 0.01023 for **c-corr**, **s-corr**, and **t-corr** respectively). It can be concluded that the rebuilt visual vocabulary could effectively improve the performance of the initial clustered visual vocabulary. When the performances of the three rebuilt vocabularies are compared, the visual vocabulary rebuilt based on temporal correlation performs best. The extra tracking procedure of the temporal information extraction may have reduced some meaningless visual words, which just randomly appear in the context .

In summary, the evaluation on the QBE near-duplicate video search task shows that the rebuilt visual vocabulary based on spatial-temporal context can effectively improve

Table 5.5: MAP Comparison of rebuilt vocabularies for different queries of CC_WEB_VIDEO

topic	BovW	c-corr	s-corr	t-corr
1	0.9607	0.9606	0.9609	**0.9707**
2	0.9769	0.9822	0.9821	**0.9823**
3	0.9187	0.9706	0.9707	**0.9738**
4	**0.9293**	0.9911	0.9912	0.9912
5	0.9798	0.9887	0.9931	**0.9917**
6	0.6178	0.6933	0.6951	**0.6998**
7	0.7263	0.7618	0.7618	**0.7623**
8	**0.5657**	0.4740	0.4664	0.4744
9	0.6622	**0.9243**	0.8982	0.9205
10	0.9683	0.9707	0.9707	**0.9708**
11	0.7217	**0.7253**	0.7038	0.7086
12	0.7239	0.7024	0.7131	0.7345
13	0.9784	0.9952	0.9952	**0.9959**
14	0.7793	**0.8932**	0.8820	0.8869
15	0.6848	0.6929	0.6928	**0.6929**
16	0.9868	0.7083	0.9009	**0.9810**
17	0.9423	0.9837	0.9840	**0.9849**
18	0.4476	0.4854	0.4825	**0.4843**
19	0.8252	0.8510	0.8916	**0.9402**
20	0.8384	0.8929	0.8939	**0.8981**
21	0.7873	0.8417	0.8417	**0.8417**
22	0.5908	0.5921	**0.5922**	0.5921
23	0.5480	0.5627	0.5627	**0.5628**
24	0.9910	0.9903	0.9903	**0.9903**
MAP	0.8010	0.8295	0.8290	**0.8347**

the performance of initial visual vocabulary generated by the classical BovW model.

5.5 Summary

In this Chapter, we present our novel approach which rebuild the visual vocabularies to address the "OverQuantize" problem of the classical BovW model. The reasonable cause of this quantization error and the trade-off in between with "UnderQuantization" error have been theoretically discussed. We assumed that the synonyms can be detected based on the context similarity measurement.

The context of a visual word has been defined by the co-occurring feature instances with it, and its STC vector extracted from specific video collection has been directly utilized to quantitatively characterize its context. Furthermore, we compute the Euclidean Distance between the context vectors of pair-wise visual words to inversely measure the context similarity. The visual words, whose computed similarity are above a threshold, are detected as synonyms. Afterwards, the similar visual words are merged to rebuild a new visual vocabulary. The visual content representation and retrieval model of the BovW framework can be renewed based on the rebuilt visual vocabulary.

A series of experimental results on the general topic video retrieval and the near-duplicate video search tasks indicate that the rebuilt visual vocabulary generally promotes the retrieval performances. Experimental results also demonstrated that the rebuilt vocabulary is compatible with another state-of-the-art BovW enhancement approach TGC. It is clearly shown that rebuilding technology based on the spatial-temporal context effectively reduces the redundant visual words in the initial visual vocabulary, and contributes to solving the "OverQuantize" problem.

In addition, the performances of rebuilt visual vocabulary, which merge different numbers of visual words, have also been discussed. We can conclude that the visual vocabularies for the different video collections always have a certain numbers of redundant visual words, and the number is empirically determined in this thesis.

Chapter 6

System and Additional Evaluation

A prototype retrieval system is implemented following the classical BovW framework. The present chapter introduces the architecture and components of this software platform. We implemented the presented representation reformulation and visual vocabulary rebuilding methods, and integrated them into the system.

With the developed system, we conducted a series of additional experiments to evaluate the effectiveness of grouped approaches, which combined the QC, the IDC, or the rebuilt vocabulary methods.

Based on the demonstrated experimental results, we summarized and discussed the performance differences between the approaches leveraging spatial information and temporal information respectively.

6.1 Function Modules of the Retrieval System

The retrieval system is implemented based on a C++ library OpenCV, which is publicly available. The SIFT and SURF features have been utilized for the video representation. The visual vocabulary is clustered by an open software **cluto**, which is able to handle large scale and high-dimension features clustering. The videos are inversely indexed using MySQL database.

The prototype system is built as a set of function modules, each of which is designed to complete a specific process. Its architecture is shown in Figure 6.1. The red line and arrow indicate the flow direction of data originally extracted from the query example. The blue line and arrow present data generated from the video collection. For example,the computations of function modules Rebuilt Visual Vocabulary and IDC Generation are based on the video collection, and QC computation uses STC data generated from the query video.

In this system, the function modules Feature Extraction, Video Indexing, and Query Representation follow the classical BovW framework. We briefly introduce the other

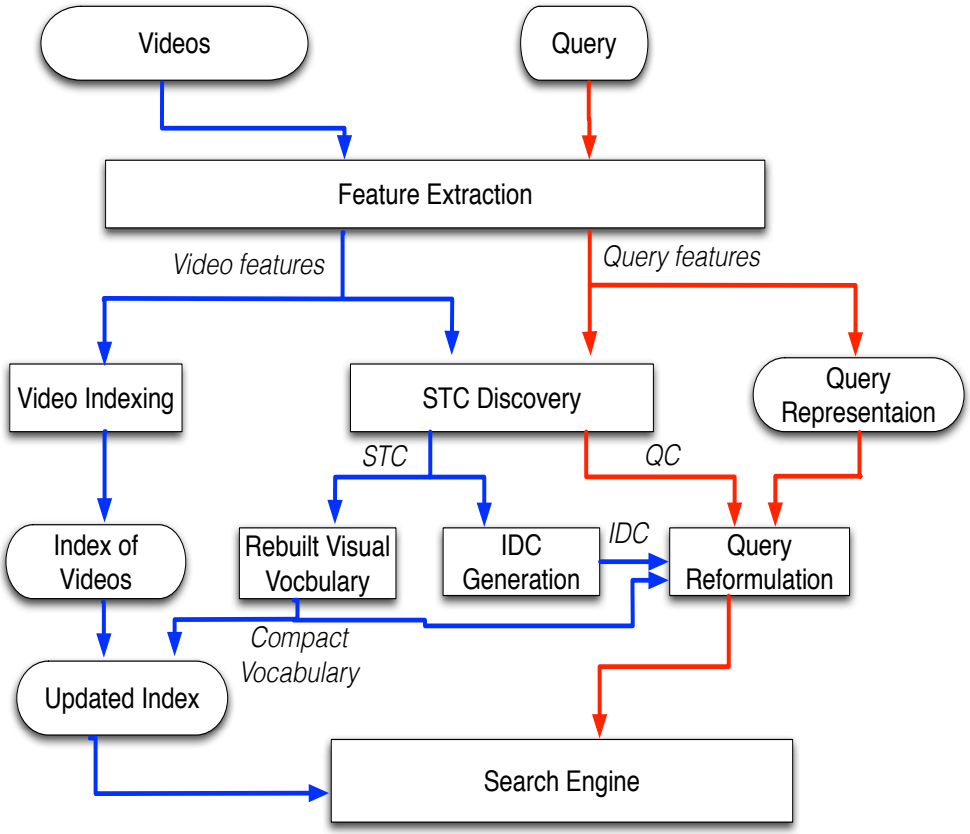


Figure 6.1: The Architecture of the prototype CBVR System based on the STC

function modules, which are associated with methods presented in this thesis, as follows:

STC discovery Module: It could be applied to a video or a video collection. Output of the module is a $K \times K$ matrix, which is written down into .txt files. The module could be utilized to discover either QC or DC based on co-occurring, spatial, or temporal correlation. It should be pointed out that the input must be the raw visual words data, which maintains the spatial and temporal information. Its input and output are summarized in Table 6.1.

Table 6.1: STC Discovery

Input	Output
raw visual words	$K \times K$ STC matrix

IDC generation Module: it generates a IDC weight for each visual word with inputed STC matrix. IDC coefficients should be generated for specific video collection. The output of this module is a K dimension vector. Each entry of it is associated with a specific visual word. Its formulation is illustrated in Table 6.2

Table 6.2: IDC generation

DC	idc value
$K \times KMatrix$	K Vector

Visual Vocabulary Rebuilding Module: It has synonyms detection and index updating functions. The detection uses the STC matrix discovered from a video collection. The index updating function renews the index with rebuilt vocabulary. The structure of this module is shown in Table 6.3.

Table 6.3: Visual Word Rebuilding

Input	Output
STC matrix, Original Index	Updated Index

Query Reformulation Module: it reformulates the query with QC or IDC weights. It could also use the two STC based term weighting schemes simultaneously. The function format of the Query Reformulation module is shown in Table 6.4:

Table 6.4: Query Reformulation

Input			Output
QC matrix	IDC coefficients	Query	Reformed Query

Finally, the Search Engine module measures the similarities between the query and videos, and it ranks the videos with the similarity scores.

In the next section, the system is applied to the two different CBVR applications respectively, and a series of additional experimental results are demonstrated to evaluate

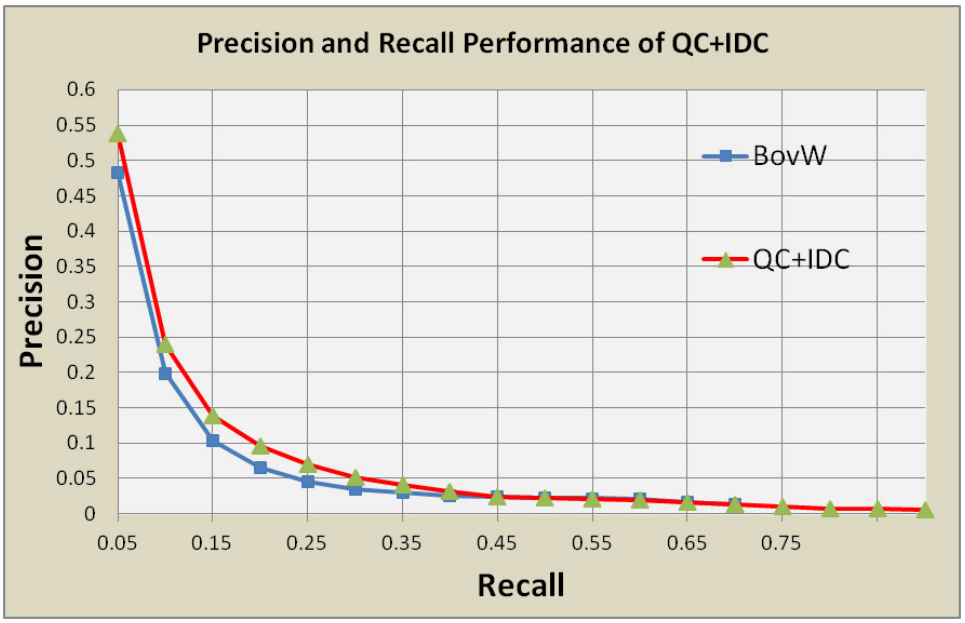


Figure 6.2: The performance of QC+IDC regarding the Precision-Recall Curve (general topic video retrieval)

the effectiveness of our approaches.

6.2 General Topic Video Retrieval

The video collection TRECVID2002 selected to perform the general topic video retrieval has been introduced in previous chapters. The retrieval task consists of 22 topics. To search for relevant videos for each topic, a video in the ground truth is selected as query example and the similar videos are retrieved by the CBVR system.

In Chapter 5, both QC and IDC term weights have been shown to be able to improve retrieval performance of the classical BovW model. In this section, we are interested in a particular question: whether or not the two approaches could co-operate with each other?

The combination of the QC and IDC approaches is to switch on both the QC and IDC function modules, and it is denoted as **QC+IDC**. The overall performance of the **QC+IDC** approach in terms of precision-recall curve is demonstrated in Figure 6.2. The **QC+IDC** approach generally outperforms the original BovW framework. Because, pure Qc or IDC is already able to improve the retrieval performance of the BovW model, we need further evidence to support the effectiveness of the **QC+IDC** approach.

According to Equation 4.22, the **QC+IDC** approach has two parameters k_{qc} and k_{idc}, which control the weights of QC and IDC in the representation, *e.g.* if k_{qc} equals zero, the **QC+IDC** approach will be equivalent to pure IDC. If we gradually increase the value of k_{qc} from 0, more QC will be added into the query representation.This procedure is defined

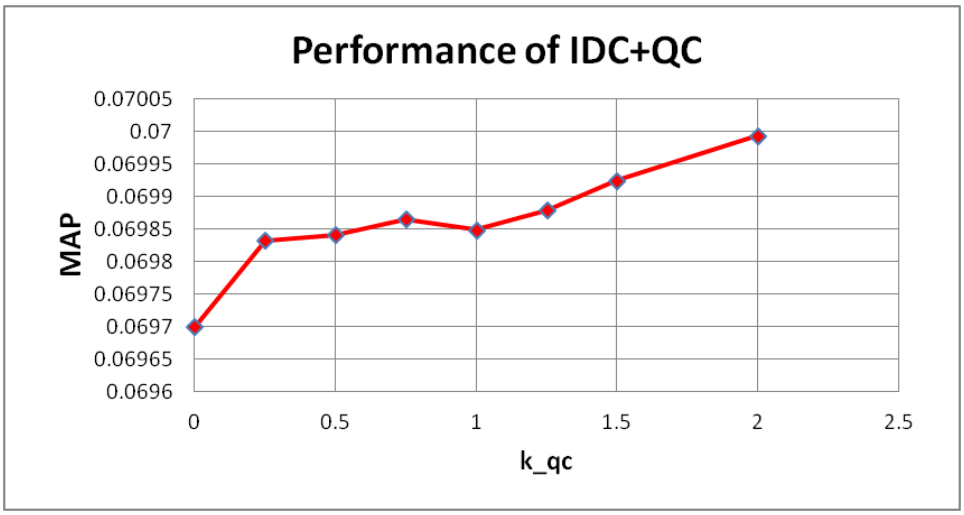


Figure 6.3: The performance of IDC+QC regarding the MAP (general topic video retrieval)

as the **IDC+QC**. To be distinguished from the aforementioned method, the method which uses a consistent k_{qc} and gradually increasing k_{idc} is denoted by **QC+IDC**.

The experimental results of **IDC+QC** are demonstrated in terms of MAP in Figure 6.3. Here, the QC is computed based on the spatial correlation and IDC is computed based on the temporal correlation as an example, because they perform more stably in previous experiments. It can be seen from Figure 6.3 that the performance increases almost linearly with the increasing k_{qc}. It shows that the QC weights works stably to improve the retrieval performance. This result matches the stable performance of the pure QC observed in previous experiments, and it provides additional evidence for the conclusion that the QC method could effectively improve the retrieval performance for the general topic video retrieval application.

The performance of **QC+IDC** is illustrated in Figure 6.4. The results are not stable: when $k_{idc} = 0.25$ or 2, the **QC+IDC** outperforms the original QC in terms of MAP, and on the other points the **QC+IDC** does not performs so good as the pure QC. It shows that although the added IDC could improve the performance of QC approach, but it is sensitive to the value of the parameter and is not stable. This is similar to the performance of the pure IDC method, which is also not stable.

Again, it can be confirmed that the IDC formulated by this thesis could improve the performance of the general topic video retrieval, although the effectiveness is not stable.

Considering all results shown in Figures 6.2, 6.3, and 6.4, the approaches QC and IDC can co-operate with each other. The two methods could be combined to improve the classical BovW model for this application.

Moreover, we also want to discuss a combination of the query reformulation and the rebuilt visual vocabulary methods, which are both based on STC. The QC and rebuilt visual

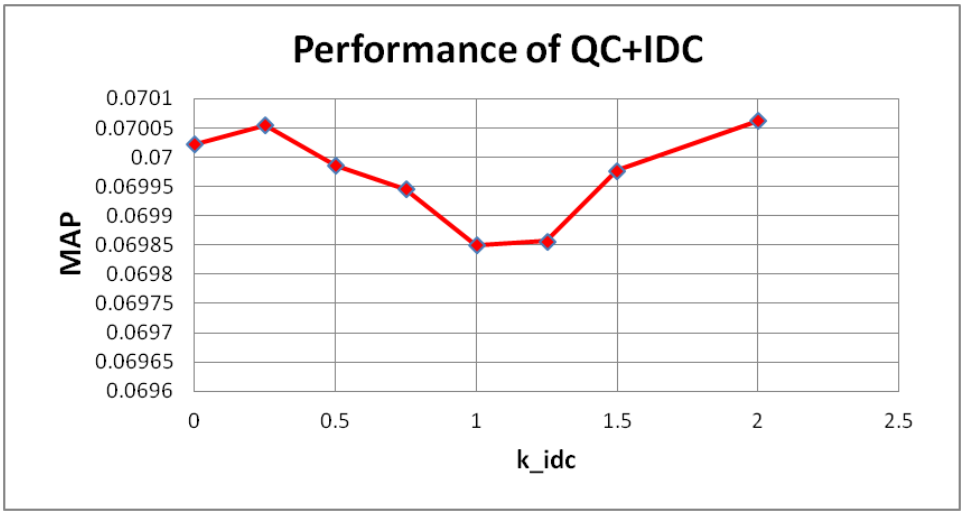


Figure 6.4: The performance of QC+IDC regarding the MAP (general topic video retrieval)

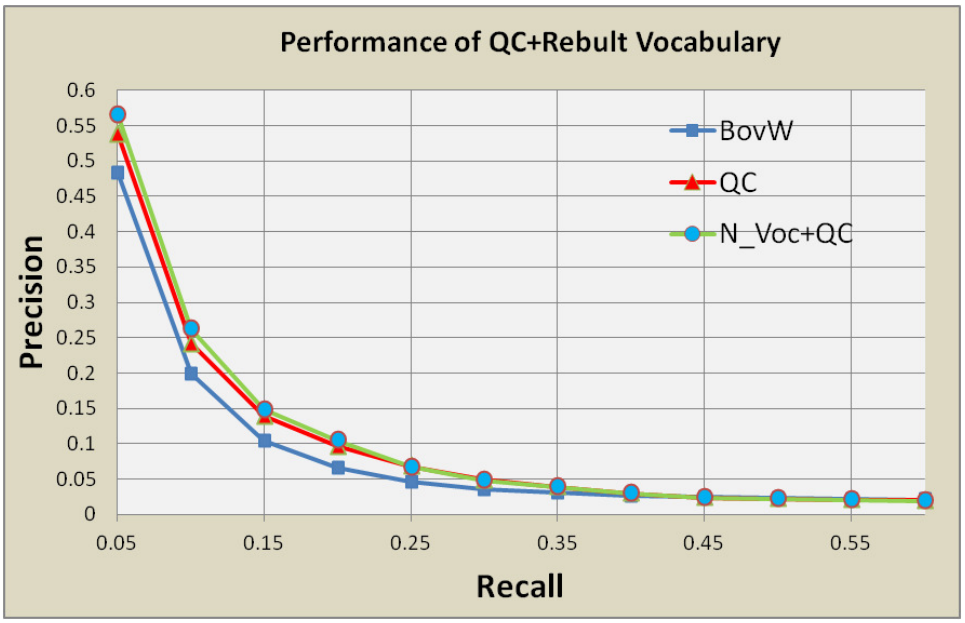


Figure 6.5: The performance of QC+N_VOC (general topic video retrieval)

vocabulary function modules are switched on simultaneously to complete this evaluation. It should be pointed out that the query is reformulated based on a co-occurring correlation as an example, and the visual vocabulary rebuilding model considers the temporal correlation. The combination method of the rebuilt visual vocabulary and QC approaches is denoted as the **QC+N_VOC** method.

The overall performance of the **QC+N_VOC** method in terms of precision-recall curve is illustrated in Figure 6.5, which compares the **QC+N_VOC** with the baseline classical

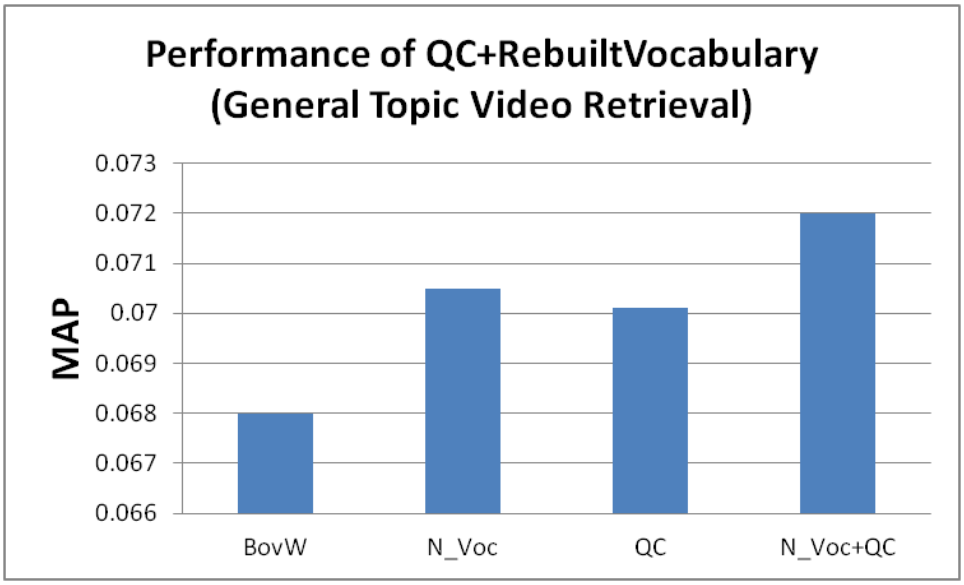


Figure 6.6: The MAP performance of QC+N_VOC (general topic video retrieval)

BovW Model and pure QC method (denoted by QC). As can be seen in Figure 6.5, the **QC+N_VOC** generally outperforms the pure QC method and classical BovW model.

Furthermore, the performance of QC+N_VOC is compared with pure QC and N_VOC methods in terms of MAP. The general comparison is shown in Figure 6.6. As shown, the combination of QC + N_VOC outperforms both pure methods regarding the MAP criteria. This result shows that it is effective to integrate the query reformulation and visual vocabulary rebuilding approaches.

It can be concluded: i) the methods QC and IDC via STC discovery could be integrated with each other in the retrieval system for the general topic video retrieval; ii) the query reformulation and visual vocabulary rebuilding approaches could be effectively integrated to improve the performance of the classical BovW framework.

6.3 Near-Duplicate Video Detection

We utilize a public available video collection CC_WEB_VIDEO, which has been introduced in previous chapters, for the near duplicate video detection application. For each topic, a series of videos are defined by the CC_WEB_VIDEO as seed videos, and ground truth files listed their near duplicate videos. All these videos are set as relevances for the the near-duplicate video detection task.

Similar to the last section, the combination of QC and IDC query reformulation approaches is firstly performed. The terms **QC+IDC** and **IDC+QC** are named in the same manner, which increases k_{idc} or k_{qc} with consistent k_{qc} or k_{idc} respectively.

Performance of the **QC+IDC** regarding MAP is illustrated in Figure 6.7. It is pre-

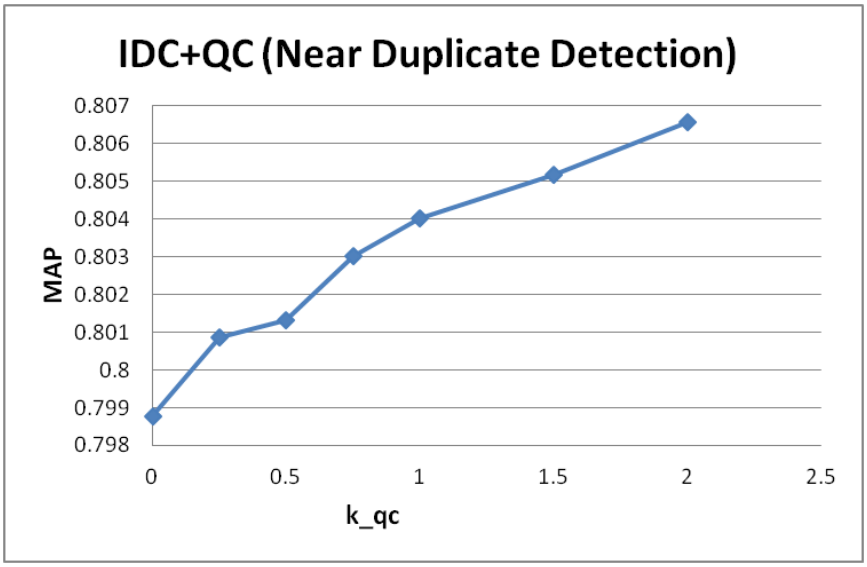


Figure 6.7: The MAP performance of IDC+QC (Near duplicate video detection)

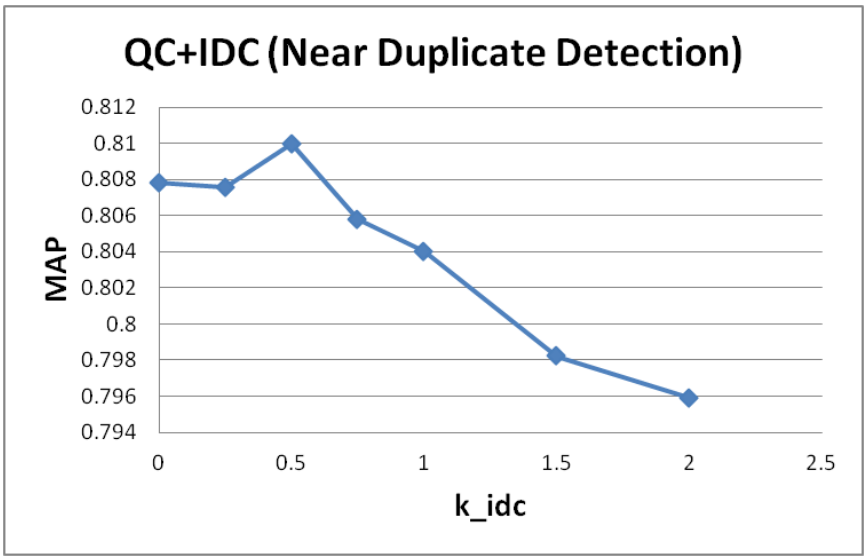


Figure 6.8: The MAP performance of QC+IDC (Near duplicate video detection)

sented that the performance of the **QC+IDC** is proportional to the value of the k_{qc}. This result shows that the QC is not in conflict with IDC and more QC will promote the retrieval performance. This observation supports our conclusion made in Chapter 4 that the QC is effective for near duplicate videos detection.

Moreover, performance of the **IDC+QC** regarding MAP is demonstrated in Figure 6.8. The added IDC weights do not exhibit stable performance in this application. Although it could improve the performance of the QC method on an individual point ($k_{idc} = 0.5$), it harms the retrieval performance on other points.

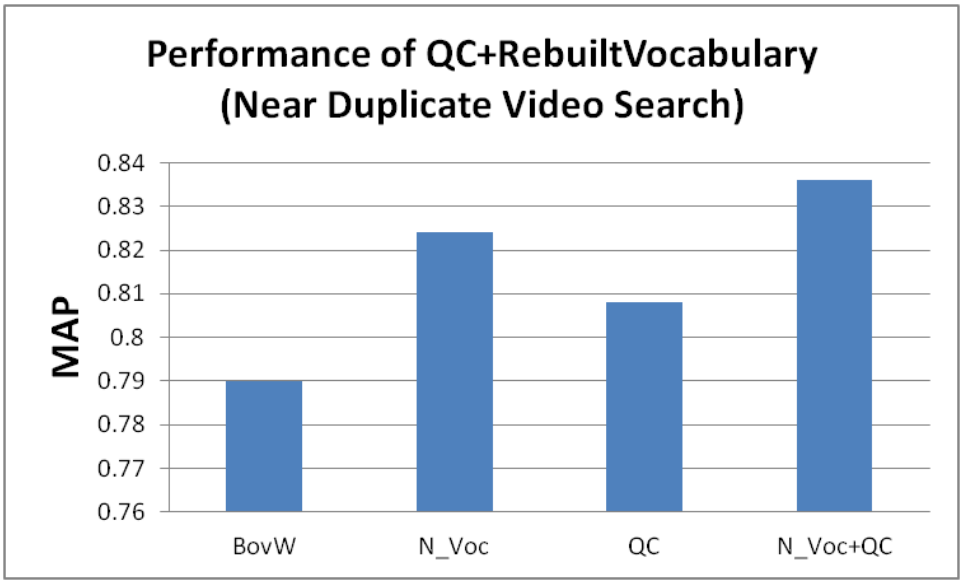


Figure 6.9: The MAP performance of QC+N_VOC (Near duplicate video detection)

The results may imply that the added IDC weights have increased the risk of the retrieval system. A reason may be that the IDC has enlarged the variance of the visual words weights. It has been shown in Table 4.3 that the IDC coefficients have a large standard deviation (larger than 0.9), which would involve in larger variance of the visual words weights in the reformulated queries. The large variance means that too much prior importances may have been assigned to some visual words, and tends to ignore certain others. As a result, the retrieval performance becomes more risky.

The standard deviation of the IDC coefficients generated for CC_WEB_VIDEO is a lot larger than TRECVID2002 as shown in Table 4.3. The performance of IDC in near-duplicate video detection is more unstable than the former application as shown in Figures 6.4 and 6.7.

The above performances of the IDC can be summarized, and it could improve the classical BovW for the near duplicate video retrieval when the appropriate parameter is selected. However, we can say that it is unstable and sensitive to the value of the parameter.

The visual vocabulary rebuilding has been shown to be effective in improving the retrieval model for the near-duplicate video detection task in Chapter 5. We combine the query reformulation method and the rebuilt visual vocabulary, whilst the method is denoted by **QC+N_VOC**. As an example, the QC is computed based on the **co-occurring correlation** matrix, and the new vocabulary is built based on the co-occurring correlation matrix extracted from CC_WEB_VIDEO data collection.

The performance of **QC+N_VOC** is illustrated by the Figure 6.9. As demonstrated by the histogram, the QC+N_VOC outperforms both the N_VOC and the QC methods,

as well as the classical BovW model. We can conclude that integrating query reformulation and visual vocabulary rebuilding would reinforce the two methods for this query-by-example near duplicate video detection.

6.4 Spatial vs Temporal Information

According to Chapter 3, the spatial and the temporal correlation are both extensions of the co-occurring correlation. As a result, the two correlations always contribute similarly to the retrieval model, which has been shown in the experimental results demonstrated in previous chapters. However, the two correlations still have two main differences: i) the temporal correlation generation involved in a visual word tracking process, which reduces some correlation; ii) the spatial correlation is more directly related with the visual object than the temporal correlation. For example, visual words associated with still background cannot be completely distinguished with pure temporal information. There must be performance differences between approaches based on the two types of difference information. This section will summarize the experimental results of the approaches based on the two different correlations.

The QC can be quantized according to the spatial and temporal correlation in the query video respectively, and the experimental results regarding MAP are summarized in Table 6.5. The experimental results show that the spatial correlation outperforms the temporal correlation for TRECVID2002 but the temporal correlation performs better for CC_WEV_VIDEO.

Table 6.5: The MAP of QCs

Data collection	spatial correlation	temporal correlation
CC_WEB_VIDEO	0.8090	**0.8186**
TRECVID2002	**0.0702**	0.0700

Table 6.6 demonstrates the comparison of IDC weights generated based on the spatial correlation and temporal correlation respectively. The experimental results of the two data collections show that temporal correlation always outperforms spatial correlation for the IDC approaches.

Table 6.6: The MAP of IDCs

Data collection	spatial correlation	temporal correlation
CC_WEB_VIDEO	0.803	**0.804**
TRECVID2002	0.0692	**0.0698**

The comparison of visual vocabularies rebuilt based on spatial and temporal correlation respectively is shown in Table 6.7. The spatial correlation performs better for TRECVID2002 and temporal correlation performs better for the CC_WEB_VIDEO.

Table 6.7: The MAP of Rebuilt Visual Voabulary

Data collection	spatial correlation	temporal correlation
CC_WEB_VIDEO	0.8293	**0.8364**
TRECVID2002	**0.0708**	0.0706

As shown in Tables 6.5, 6.6, and 6.7, for CC_WEB_VIDEO, temporal correlation outperforms the spatial correlation, and for TRECVID2002, the spatial correlation always outperforms the temporal correlation, with the exception of the IDC approach. The data collection CC_WEB_VIDEO contains the un-segmented web videos, in which key frames are more different. The temporal correlation may have reduced noisy correlations between the visual words. However, the data collection TRECVID2002 is composed of well segmented video shots by the given shot boundary ground truth. The correlation based on the spatial information may be more straightforward and meaningful. The performance differences between the spatial and the temporal correlations are never very large, because both of them are extended from the co-occurring correlation.

The spatial and temporal correlations can be combined to form the spatial-temporal correlation matrix as demonstrated in the Chapter 3. The combination utilizes the spatial and temporal information simultaneously as shown in Equation 3.60, which has a parameter k_{st} to control the weights of temporal correlation within the ST matrix. For example, if $k_{st} = 0$, the ST correlation becomes pure spatial correlation.

Several experiments are conducted to evaluate the effectiveness of combining the two different correlations, and the QC with more stable performance in the previous CBVR experiments is utilized as an example.

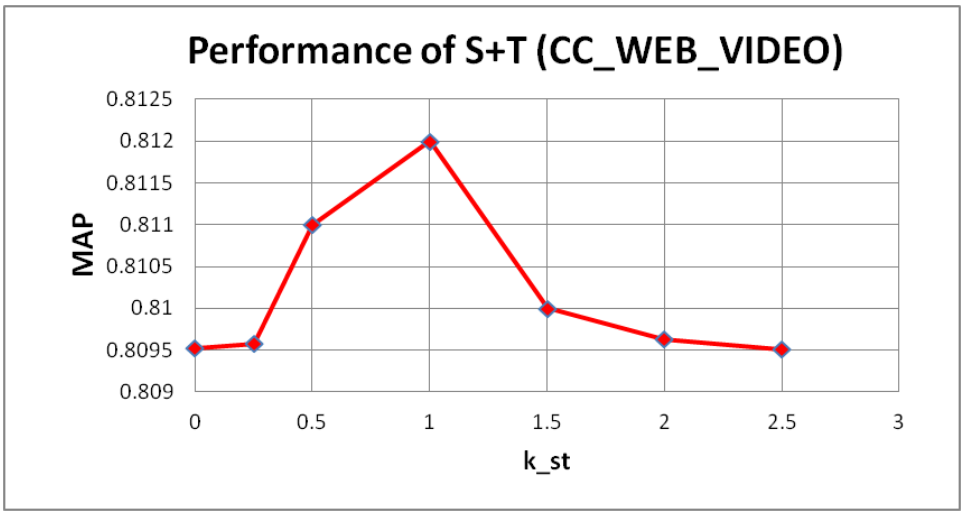


Figure 6.10: The MAP performance of spatial+temporal correlation (near duplicate video detection)

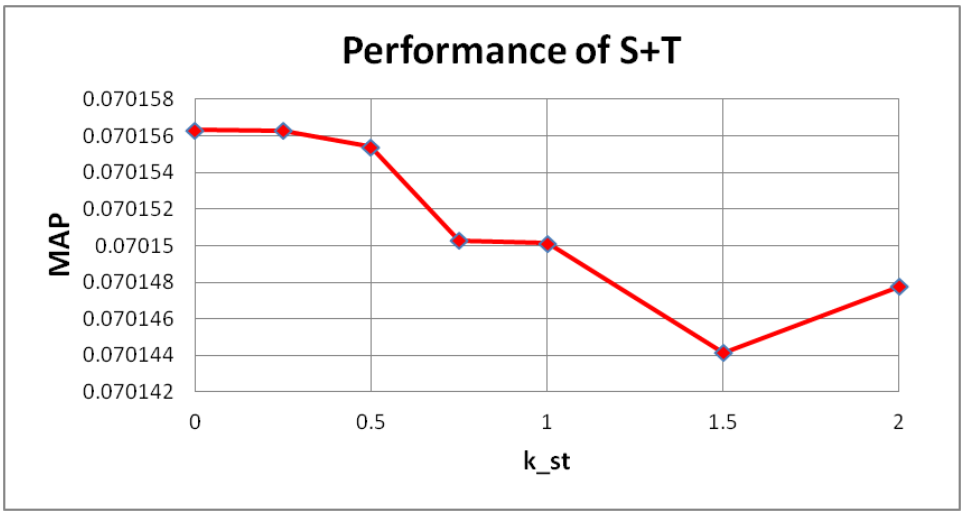


Figure 6.11: The MAP performance of spatial+temporal correlation (general topic video retrieval)

The performances of the QC based on the spatial-temporal correlation (denoted as S+T) computed with different k_{st} is shown in Figures 6.10 and 6.11 in terms of MAP respectively. Larger k_{st} means that more temporal correlation is contained in the S+T. For near duplicate video detection, the S+T outperforms the pure spatial correlation on some points. It has been shown in previous experimental results that temporal correlation performs better in this task. The added temporal correlation makes the S+T outperform the pure spatial correlation on these points.

The demonstrated performance of the S+T is not as good as pure spatial correlation for general topic video retrieval. This may be because of that the temporal correlation is not as meaningful as the spatial correlation in this application.

Direct integration of the spatial and temporal correlations does not obviously outperform the pure temporal or spatial information in these two applications, as shown in Table 6.5. The effectiveness of the direct integration has not been fully supported by the experimental results.

6.5 Summary

In the present chapter, the architecture of the experiment system has been introduced.

A series of additional experiments are conducted for the two common tasks, which are the near duplicate video detection and general topic video retrieval. The two term weighting schemes: QC and IDC are firstly combined together. They are shown to be effective to co-operate with each other to promote the retrieval performances.

Furthermore, the query reformulation method is integrated into the rebuilt visual vocabulary. A series of experimental results revealed that the combination method is

effective to improve the CBVR technology.

The performance comparison between the spatial and the temporal correlation is also demonstrated in the present chapter. It is illustrated that the spatial correlation works better for unsegmented web videos, and the temporal correlation performs better for well segmented TRECVID2002 videos.

A linear method to directly integrate the spatial and temporal correlation is proposed in Chapter 3. However, the experimental results in the present chapter show that it is not always effective to improve the pure spatial or the pure temporal correlation. A better method to integrate the two correlations should be investigated in the future.

Chapter 7

Conclusions and Future Work

In this thesis, we aimed to tackle the limitations of the BovW framework by considering the relationship across the visual words, which is characterized by spatial and temporal correlation discovered from visual information. A novel framework, theory and methods have been developed to improve descriptive ability of visual content representation and retrieval model. This chapter concludes our main contributions and points out a future direction.

7.1 Contributions

A novel spatial and temporal information discovery and quantization framework has been proposed. It is assumed that spatially or temporally co-occurring instances are clues of the correlation between the corresponding visual words.

The retrieval function is modified to incorporate the spatial and temporal correlation. The correlations discovered from the query example and video collections would characterize novel term weights of the visual words, and this model emphasizes the descriptive visual words for different retrieval topics. The model reformulates the query representation with the defined term weighting schemes and establishes a new similarity measurement function.

The STC could be leveraged to define the context of a visual word in terms of its co-occurring visual words. With the context similarities, an approach is developed to detect synonymous visual words, and the detected synonyms are merged to rebuild a more compact and effective visual vocabulary.

The following sections outline in details the main thesis contributions.

7.1.1 The Spatial and Temporal Information Discovery Framework

Our framework is aiming to quantitatively discover spatial and temporal correlation between co-occurring visual words, which is an approximation of semantic correlation be-

tween the visual words. We must approximate this correlation for our retrieval technology development, because this correlation can not be discovered directly from the visual information with existing technology.

If we consider the co-occurring visual words as a joint term, an instance of the joint term is composed of instances of the corresponding visual words. Thus, the number of the joint terms are K^N, where N equals the order of the co-occurrence. The occurrence of a joint term can be mathematically modelled by counting its frequency. The occurrence of K^2 joint terms (in this thesis, we focus on 2-words correlation) is formulated as a $K \times K$ matrix. Each entry of the matrix represents the frequency of a joint term, which is equivalent to the co-occurrence of the corresponding visual words. To make the co-occurrence comparable between different visual words, we normalize the co-occurrence with the term frequency of corresponding visual words, and define the normalized co-occurrences as **co-occurring correlation**. It is the first STC concept established for the discovery framework. The computed co-occurring correlation is also in the form of $K \times K$ matrix.

Furthermore, the co-occurring correlation could be refined according to additional geometric information, because not all co-occurring visual words are actually related to each other. We developed a method to refine the occurrence of a joint term according to several spatial and temporal constraints. The constraints are established to approximate the probability that co-occurring visual words perceptually belongs to identical visual entity. In this case, the visual entity should geometrically cover all visual words, which are truly "correlated". In this way, we extend the co-occurrence to get a finer approximation of the semantic correlation.

To achieve this objective, we set up the constraints using spatial proximity and temporal motion coherence between the visual words. Firstly, when the visual words tend to appear spatially closely to each other, they are more likely to be correlated. Secondly, if the visual words always move coherently between continuous frames of a video, they are more likely to be correlated.

The spatial proximity is quantized according to physical distance, and relative motion is tracked to model the motion coherence between the visual words. The co-occurrence computation model is adapted with a Gaussian Function to incorporate the modelled spatial proximity or the motion coherence. Similar to the co-occurrence, the spatially or temporally co-occurrences are also normalized by the term frequency of corresponding visual words, and then are defined as two concepts **spatial correlation** and **temporal correlation**. The constructed spatial or temporal correlation are both represented in the form of $K \times K$ matrices.

The three concept **co-occurring correlation**, **spatial correlation** and **temporal correlation** are generally known as Spatial-Temporal Correlation, which is denoted by STC in this thesis. The spatial correlation and temporal correlation are both extensions

of the co-occurring correlation.

We also attempt to combine the spatial correlation and the temporal correlation. The spatial proximity and the temporal motion coherence are hard to be directly fused, because they used different units of measurement. Then, we formulate a spatial-temporal correlation matrix via linearly combing the two correlation matrices, because both of them are estimated as a degree of probability of the true correlation.

Based on a series of practical experimental results, we concluded that the STC discovered by the developed framework could effectively promote retrieval performance of the BovW model. It is also shown that the spatial and the temporal correlation perform differently in different CBVR applications. It can not be concluded that any correlation is better, because no correlation consistently outperforms the other one.

7.1.2 The Video Retrieval Model via STC Discovery

We developed two novel visual word weighting schemes to modify the representation of visual information based on the discovered STC. The schemes utilized the STC discovered from the query video and the video collection respectively.

It is shown in preliminary exploration that a number of selected descriptive visual words would better described the visual information for the retrieval model. These descriptive visual words are called Words-of-Interest (WoI) in this thesis, and the WoI selection algorithms are established based on the proposed spatial-temporal correlation.

Motivated by this exploration, we proposed an assumption that the visual words with higher STC incurred by query video are of higher descriptive ability. Based on this hypothesis, we emphasized the discriminative visual words in the similarity measurement function. According to the theoretical analysis, this modified similarity measurement model is actually equivalent to reformulating key frames of the query with a new term weighting scheme. Here, the weight is computed by integrating the STC matrix discovered from the query and the term frequency (TF) vector. We refer to this term weight as Query Correlation (QC) weight. With this QC weighting scheme, the modified similarity measurement function can be easily implemented for the inverted index videos structure.

The QC can be seen as an analog of a common concept: TF in traditional textual IR. Inspired by another important concept Inverse Document Frequency (IDF), the STC discovered from entire video collection are defined as Documents Correlation (DC). We set up a hypothesis that higher DC with more visual words would harm descriptive ability of a visual word. Based on this assumption, we defined another new term weighting scheme: Inverse Documents Correlation (IDC), in which the visual words are inversely weighted according to the discovered DC in similarity measurement model.

Both the QC and IDC weighting schemes reformulate the video representation for the similarity measurement and relevance prediction. In this thesis, the two methods are sometimes generally referred to as the query reformulation method. The method can be

implemented without extra storage cost for the video index, because it does not directly integrate additional geometric information into videos representations.

A series of experimental results demonstrate that the QC method substantially improves the classical BovW model. It has also outperformed a state-of-the-art TGC-based approach on challenging tasks. The results have verified our hypothesis that the discriminative visual words characterized and emphasized according to QC would effectively compensate the neglected spatial-temporal information by the classical BovW model. The combination of the QC and rebuilt vocabulary (MAP = 7.2%) outperforms state-of-the-art achievement on TrecVID2002 shown in (Donald & Smeaton 2005) (MAP = 6.9%), and it should be noticed that the QC only uses visual information but the literature (Donald & Smeaton 2005) fuse multiple information including textual information.

Experimental results also show that the IDC approach could improve the performance of the BovW model for CBVR tasks. The improvement on an CBVR application is not statistically significant. The failure of IDC may be a results of enlarged weighting difference. IDC may involve in risk of overemphasizing some visual words and then lose relevances, because it is found that the difference of visual words weights have increased by IDC. Some isolated visual words may have been assigned with too large weights with our linear IDC weighting function. However, the demonstrated improvement has at least indicated a possibility to develop the CBVR technology with IDC, and more investigations could be completed for this direction in the future.

In this thesis, the compatibility between the two query reformulation technologies has also been discussed. It is shown by the experiments that the QC and IDC technologies could co-operate with each other to improve the retrieval performances.

Overall, we can conclude that the developed video retrieval model via discovering both spatial and temporal information from the query and the video collection is effective. Discovering and utilizing STC in the retrieval model may involve in additional computation complexity. However, considering the demonstrated performance improvement, the computational cost is acceptable. Especially, the online computation of QC technology only considers the spatial-temporal information within the query without adding extra storage expense to the data videos. The IDC technology does not involve extra computational complexity in on-line searching, and the IDC coefficients could be pre-computed offline.

7.1.3 Visual Vocabulary Rebuilding Method based on the STC

A novel approach is developed for rebuilding the visual vocabularies to address quantization errors. Many factors may cause these errors in one of the most important procedures of the BovW framework: visual vocabulary generation. Indeed, these errors would definitely lead to relevances prediction mistakes of the retrieval system. We aimed to solve a defined “OverQuantize” problem, which is a special type of quantization error. It means that various visual words always represent identical visual information, which are called

synonymous visual words. We developed a technology to detect the synonyms based on context similarity measurement between the visual words.

The context of a visual word in terms of its co-occurring visual words is utilized in this method. The STC discovery framework provides a convenient tool to characterize this context, and the STC vector extracted from a video collection is defined as the context of the corresponding visual word. Following this, the spatial-temporal context of individual visual word is defined accordingly with the spatial correlation and temporal correlation respectively.

With the quantitative visual words context, the similarity of the context between pairwise visual words can be inversely measured by Euclidean distance between the two STC vectors. Any visual words, whose computed similarity is below a threshold, are detected as synonyms. The synonyms are merged to rebuild the initial visual vocabulary to be more compact. Then, the visual content representations are renewed based on the rebuilt visual vocabulary.

A series of experimental results on two common CBVR applications have shown that the rebuilt visual vocabulary effectively promotes the performance of the initial vocabulary. It is also demonstrated that the rebuilt vocabulary is compatible with TGC, which is an state-of-the-art disambiguation approach under the BovW framework.

Based on the practical experimental results, we concluded that the proposed rebuilding technology based on the STC effectively reduces redundancies in the initial visual vocabulary and contributes to solving the "OverQuantize" problem. It can also be concluded that some synonymous visual words can be detected according to their physical context without additional semantical information.

In addition, the retrieval performance increases gradually, when the rebuilt visual vocabulary increases the threshold and merges more detected synonymous visual words. But if too many visual words are merged, the retrieval performance will decrease.

Furthermore, this thesis combines the query reformulation method and this visual vocabulary rebuilding technology. The combination method has outperformed each single technology in the experiments. We concluded that the two technologies co-operate effectively with each other to enhance the BovW framework.

In summary, this thesis has presented a series of novel contributions: to discover spatial and temporal information in the videos whilst to retrieve the videos considering this information. It can also be concluded that the spatial and temporal information discovery is effective for CBVR technology development.

7.2 Limitation

Although our method has made progress in developing CBVR technology, it has some limitations at current stage.

Current STC discovery technology relies on fixed parameters as shown in Chapter 3. These parameters should be adaptively selected according to individual visual words, because the true correlation measurement actually varies from case to case due to change of *e.g.* scale, direction, and background.

The query reformulation technology in Chapter 4 also has a limitation, namely that directly combining the spatial and the temporal correlation has not succeeded in improving the single correlation. The performance of IDC technology is not very stable.

Synonymous visual words detection technology in Chapter 5 utilizes the single geometrical context and a hard-decision threshold to detect the perceptive similarity, which may not be sufficient to find out all synonyms in the visual vocabulary.

7.3 Future Work

In the future, we will endeavor to overcome the limitations of this work highlighted in Section 7.2 and make the work more general. First of all, additional constraints such as appearance could be investigated to refine the parameters selection for STC discovery technology. Other advanced spatial and temporal information discovery methods, for example, foreground and background extraction, could also be utilized to characterize the accurate spatial-temporal correlation.

One possible way in which to refine the query reformulation technology is to utilize users' feedback. The current STC based retrieval model has utilized the correlation from query and video collection as shown in Chapters 4 and 5. Positive feedbacks from users could provide a knowledge space, which is less noisy than the video collection but more diverse than the query example. The discovered spatial and temporal information differs from QC and IDC, and it may be used to establish a new weighting scheme for the retrieval function, which will result in a new interactive video search model.

The synonymous visual words detection technology could incorporate additional semantic information into the context similarity measurement. To achieve this objectives, we will systematically investigate the semantic correlation between the visual words. Currently, the semantic correlation is approximated by the spatial temporal constraints. The big gap between the visual correlation and the semantic correlation may be one of major obstacles. We believe that the incorporating semantic modelling technology into the spatial temporal correlation will effectively bridge this gap.

The spatial-temporal correlation technology can be developed with additional semantic knowledge. For example, more videos are displayed in webs containing multi-modal information, including textual information. Our work could represent a good starting point, and we will extend the correlation formulation to multi-modal correlation, for example text word-visual word correlation. We will investigate on whether or not the multi-modal correlation would benefit the multi-modal information access and retrieval.

The additional semantic knowledge can be used as a guide to build more complex spatial-temporal correlation model. It has been proposed to build a hierarchical relationship (Jiang & Ngo 2009) between the visual words with the semantic information. It set up a vertical relations rather than horizontal correlation in our work. It is interesting here to answer one particular question: how can we build a uniform model combining the subordination and flat correlations to develop the retrieval technology?

Other than video retrieval, we believe that it is possible to use the correlation model for other video processing and understanding tasks. In Chapter 4, a new similarity measurement scheme is built based on the STC. The core idea is to softly measure the similarity between a visual word and its correlated words. Minor expansions would enable this model to compute the videos differences for supervised learning. The effectiveness of the correlation model should be evaluated in more experiments, for example, videos classification or categorization.

Our ultimate aim is to integrate all above research works into a uniform correlation model. We believe that no visual feature should be considered independently in visual information perception, especially the visual information retrieval and access. Not only intra- or inter- videos but also the inter- modals correlations maintained in the multimedia information should be modelled. The discovery would definitely be a new direction of the video access and retrieval technology development.

Bibliography

Adali, S., Candan, K. S., C, K. S., shing Chen, S., Erol, K. & Subrahmanian, V. S. (1996). Advanced video information system: Data structures and query processing, *Multimedia Systems* **4**: 172–186.

Baeza-Yates, R. A. & Ribeiro-Neto, B. (1999). *Modern Information Retrieval*, Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

Bakker, E. M. & Lew, M. S. (2002). Semantic video retrieval using audio analysis, *Proceedings of the International Conference on Image and Video Retrieval*, CIVR '02, Springer-Verlag, London, UK, UK, pp. 271–277.
URL: *http://dl.acm.org/citation.cfm?id=648261.753216*

Ballan, L., Bertini, M., Bimbo, A. & Serra, G. (2009). Video event classification using bag of words and string kernels, *in* P. Foggia, C. Sansone & M. Vento (eds), *Image Analysis and Processing ICIAP 2009*, Vol. 5716 of *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, pp. 170–178.

Ballard, D. (1981). Generalizing the hough transform to detect arbitrary shapes, *Pattern Recognition* **13**(2): 111 – 122.
URL: *http://www.sciencedirect.com/science/article/pii/0031320381900091*

Bay, H., Ess, A., Tuytelaars, T. & Van Gool, L. (2008). Speeded-up robust features (surf), *Comput. Vis. Image Underst.* **110**: 346–359.
URL: *http://portal.acm.org/citation.cfm?id=1370312.1370556*

Bescos, J., Cisneros, G., Martinez, J. M., Menendez, J. M. & Cabrera, J. (2005). A unified model for techniques on video-shot transition detection, *Trans. Multi.* **7**(2): 293–307.
URL: *http://dx.doi.org/10.1109/TMM.2004.840598*

Bolovinou, A., Pratikakis, I. & Perantonis, S. (2013). Bag of spatio-visual words for context inference in scene classification, *Pattern Recogn.* **46**(3): 1039–1053.
URL: *http://dx.doi.org/10.1016/j.patcog.2012.07.024*

Boreczky, J. S. & Rowe, L. A. (1996). Comparison of video shot boundary detection techniques, *Journal of Electronic Imaging* **5**(2): 122–128.

Bosch, A., Zisserman, A. & Muñoz, X. (2006). Scene classification via pLSA, *In Proc. ECCV*, pp. 517–530.
URL: *http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.65.3509*

Bouchard, G. (2005). Hierarchical part-based visual object categorization, *In Proc. CVPR*, pp. 710–715.

Brand, M., Oliver, N. & Pentland, A. (1997). Coupled hidden markov models for complex action recognition, *Proceedings of the 1997 Conference on Computer Vision and Pattern Recognition (CVPR '97)*, CVPR '97, IEEE Computer Society, Washington, DC, USA, pp. 994–.
URL: *http://portal.acm.org/citation.cfm?id=794189.794420*

Bretzner, L. & Lindeberg, T. (1998). Feature tracking with automatic selection of spatial scales, *Comput. Vis. Image Underst.* **71**: 385–392.
URL: *http://portal.acm.org/citation.cfm?id=306695.306726*

ByJason Farquhar, Sandor Szedmak, H. M. & Shawe-Taylor, J. (2005). Improving "bag-of-keypoints" image categorisation, *Technique Report*, University of Southampton.

Cao, L. & Li, F.-F. (2007). Spatially coherent latent topic model for concurrent segmentation and classification of objects and scenes, *Computer Vision, IEEE International Conference on* **0**: 1–8.

Cao, L., Tian, Y., Liu, Z., Yao, B., Zhang, Z. & Huang, T. S. (2010). Action detection using multiple spatial-temporal interest point features, *ICME*, pp. 340–345.

Carson, C., Belongie, S., Greenspan, H. & Malik, J. (2002a). Blobworld: image segmentation using expectation-maximization and its application to image querying, *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **24**(8): 1026 – 1038.

Carson, C., Belongie, S., Greenspan, H. & Malik, J. (2002b). Blobworld: Image segmentation using expectation-maximization and its application to image querying, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **24**: 1026–1038.

Carson, C., Thomas, M., Belongie, S., Hellerstein, J. M. & Malik, J. (1999). Blobworld: a system for region-based image indexing and retrieval (long version), *Technical Report UCB/CSD-99-1041*, EECS Department, University of California, Berkeley.
URL: *http://www.eecs.berkeley.edu/Pubs/TechRpts/1999/5567.html*

Chen, T.-W., Chen, Y.-L. & Chien, S.-Y. (2008). Fast image segmentation based on k-means clustering with histograms in hsv color space, *2008 IEEE 10th Workshop on Multimedia Signal Processing* pp. 322–325.
URL: *http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4665097*

Chum, O., Mikulik, A., Perdoch, M. & Matas, J. (2011). Total recall ii: Query expansion revisited, *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pp. 889 –896.

Chum, O., Philbin, J., Sivic, J., Isard, M. & Zisserman, A. (2007). Total recall: Automatic query expansion with a generative feature model for object retrieval, *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pp. 1 –8.

Chum, O., Philbin, J. & Zisserman, A. (2008). Near duplicate image detection: min-hash and tf-idf weighting, *British Machine Vision Conference*.

Cooper, M. (2004). Video segmentation combining similarity analysis and classification, *Proceedings of the 12th annual ACM international conference on Multimedia*, MULTIMEDIA '04, ACM, New York, NY, USA, pp. 252–255.
URL: *http://doi.acm.org/10.1145/1027527.1027584*

Datta, R., Joshi, D., Li, J. & Wang, J. Z. (2008). Image retrieval: Ideas, influences, and trends of the new age, *ACM Computing Surveys* **40**(2): 1–60.

Decleir, C., Hacid, M.-S. & Kouloumdjian, J. (1999). A database approach for modeling and querying video data, *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING* **12**: 729–750.

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm, *JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B* **39**(1): 1–38.

Donald, K. & Smeaton, A. (2005). A comparison of score, rank and probability-based fusion methods for video shot retrieval, *in* W.-K. Leow, M. Lew, T.-S. Chua, W.-Y. Ma, L. Chaisorn & E. Bakker (eds), *Image and Video Retrieval*, Vol. 3568 of *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, pp. 61–70.

Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J. & Zisserman, A. (n.d.). The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results, http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html.

Fergus, R., Perona, P. & Zisserman, A. (2003). Object class recognition by unsupervised scale-invariant learning, *In CVPR*, pp. 264–271.

Fergus, R., Perona, P. & Zisserman, A. (2005). A sparse object category model for efficient learning and exhaustive recognition, *In CVPR*, pp. 380–387.

Fernando, B., Fromont, E., Muselet, D. & Sebban, M. (2012). Supervised learning of gaussian mixture models for visual vocabulary generation, *Pattern Recogn.* **45**(2): 897–

907.
URL: *http://dx.doi.org/10.1016/j.patcog.2011.07.021*

Flickner, M., Sawhney, H., Niblack, W., Ashley, J., Huang, Q., Dom, B., Gorkani, M., Hafner, J., Lee, D., Petkovic, D., Steele, D. & Yanker, P. (1995). Query by image and video content: The qbic system, *Computer* **28**(9): 23–32.
URL: *http://dx.doi.org/10.1109/2.410146*

Galleguillos, C., Rabinovich, A. & Belongie, S. (2008). Object categorization using co-occurrence, location and appearance, *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pp. 1–8.

Hauptmann, A. G. (2002). Multi-modal information retrieval from broadcast video using ocr and speech recognition, *in JCDL 02: Proceedings of the 2nd ACM/IEEE-CS joint conference on Digital libraries*, ACM Press, pp. 160–161.

Hauptmann, A. G., Christel, M. G. & Yan, R. (2008). Video retrieval based on semantic concepts, *Proceedings of The IEEE* **96**: 602–622.

Hopfgartner, F. (2007). Evaluating the implicit feedback models for adaptive video retrieval, *ACMMIR* **7**: 323–331.

Hsu, W. H. & Chang, S.-F. (2005). Visual cue cluster construction via information bottleneck principle and kernel density estimation, *Proceedings of the 4th international conference on Image and Video Retrieval*, CIVR'05, Springer-Verlag, Berlin, Heidelberg, pp. 82–91.

Hu, S. (2005). Efficient video retrieval by locality sensitive hashing, *Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP '05). IEEE International Conference on*, Vol. 2, pp. 449–452.

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

Huang, J., Kumar, S. R., Mitra, M., Zhu, W.-J. & Zabih, R. (1997). Image indexing using color correlograms, *Proceedings of the 1997 Conference on Computer Vision and Pattern Recognition (CVPR '97)*, CVPR '97, IEEE Computer Society, Washington, DC, USA, pp. 762–.
URL: *http://dl.acm.org/citation.cfm?id=794189.794514*

Illingworth, J. & Kittler, J. (1988). A survey of the hough transform, *Computer Vision, Graphics, and Image Processing* **44**(1): 87 – 116.
URL: *http://www.sciencedirect.com/science/article/pii/S0734189X88800331*

Jain, A. K. & Vailaya, A. (1996). Image retrieval using color and shape, *Pattern Recognition* **29**(8): 1233 – 1244.
URL: *http://www.sciencedirect.com/science/article/pii/0031320395001603*

Jégou, H., Douze, M. & Schmid, C. (2010). Improving bag-of-features for large scale image search, *Int. J. Comput. Vision* **87**: 316–336.
URL: *http://dx.doi.org/10.1007/s11263-009-0285-2*

Jiang, Y.-G. & Ngo, C.-W. (2009). Visual word proximity and linguistics for semantic video indexing and near-duplicate retrieval, *Comput. Vis. Image Underst.* **113**: 405–414.
URL: *http://portal.acm.org/citation.cfm?id=1502816.1503023*

Kaliciak, L., Song, D., Wiratunga, N. & Pan, J. (2012). Improving content based image retrieval by identifying least and most correlated visual words, *8th Asia Information Retrieval Societies Conference (AIRS2012)*.
URL: *http://oro.open.ac.uk/34650/*

Ke, Y. & Sukthankar, R. (2004). Pca-sift: A more distinctive representation for local image descriptors, pp. 506–513.

Ke, Y., Sukthankar, R. & Hebert, M. (2005). Efficient visual event detection using volumetric features, *Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1 - Volume 01*, ICCV '05, IEEE Computer Society, Washington, DC, USA, pp. 166–173.
URL: *http://dx.doi.org/10.1109/ICCV.2005.85*

Ko, B. & Byun, H. (2002). Integrated region-based image retrieval using region's spatial relationships, *Proceedings of the 16 th International Conference on Pattern Recognition (ICPR'02) Volume 1 - Volume 1*, ICPR '02, IEEE Computer Society, Washington, DC, USA, pp. 10196–.
URL: *http://portal.acm.org/citation.cfm?id=839290.842694*

Kovashka, A. & Grauman, K. (2010). Learning a hierarchy of discriminative space-time neighborhood features for human action recognition, *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on* **0**: 2046–2053.

Laptev, I. & Lindeberg, T. (2003). Space-time interest points, *IN ICCV*, pp. 432–439.

Lazebnik, S., Schmid, C. & Ponce, J. (2006). Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories, *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2*, CVPR '06, IEEE Computer Society, Washington, DC, USA, pp. 2169–2178.
URL: *http://dx.doi.org/10.1109/CVPR.2006.68*

Lienhart, R. (2001). Reliable transition detection in videos: A survey and practitioners guide, *International Journal of Image and Graphics* **1**: 469–486.

Lienhart, R. W. (1998). Comparison of automatic shot boundary detection algorithms, pp. 290–301.

Lindeberg, T. (1996). Edge detection and ridge detection with automatic scale selection, *Proceedings of the 1996 Conference on Computer Vision and Pattern Recognition (CVPR '96)*, CVPR '96, IEEE Computer Society, Washington, DC, USA, pp. 465–.
URL: *http://dl.acm.org/citation.cfm?id=794190.794608*

Liu, D. & Chen, T. (2009). Video retrieval based on object discovery, *Comput. Vis. Image Underst.* **113**: 397–404.
URL: *http://portal.acm.org/citation.cfm?id=1502816.1503022*

Lloyd, S. P. (1982). Least squares quantization in pcm, *IEEE Transactions on Information Theory* **28**: 129–137.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints, *Int. J. Comput. Vision* **60**: 91–110.
URL: *http://portal.acm.org/citation.cfm?id=993451.996342*

Mallows, C. L. (1972). A note on asymptotic joint normality.

Matas, J., Chum, O., Urban, M. & Pajdla, T. (2004). Robust wide-baseline stereo from maximally stable extremal regions, *Image and Vision Computing* **22**(10): 761 – 767. ¡ce:title¿British Machine Vision Computing 2002¡/ce:title¿.
URL: *http://www.sciencedirect.com/science/article/pii/S0262885604000435*

Mikolajczyk, K. & Schmid, C. (2004). Scale & affine invariant interest point detectors, *Int. J. Comput. Vision* **60**: 63–86.
URL: *http://portal.acm.org/citation.cfm?id=990376.990402*

Mikolajczyk, K. & Schmid, C. (2005). A performance evaluation of local descriptors, *IEEE Transactions on Pattern Analysis & Machine Intelligence* **27**(10): 1615–1630.
URL: *http://lear.inrialpes.fr/pubs/2005/MS05*

Mikolajczyk, K., Tuytelaars, T., Schmid, C., Zisserman, A., Matas, J., Schaffalitzky, F., Kadir, T. & Gool, L. V. (2005). A comparison of affine region detectors, *Int. J. Comput. Vision* **65**(1-2): 43–72.
URL: *http://dx.doi.org/10.1007/s11263-005-3848-x*

Niebles, J. C., Wang, H. & Fei-fei, L. (2006). Unsupervised learning of human action categories using spatial-temporal words, *In Proc. BMVC*.

Nister, D. & Stewnius, H. (2006). Scalable recognition with a vocabulary tree, *IN CVPR*, pp. 2161–2168.

Oomoto, E. & Tanaka, K. (1993). Ovid: Design and implementation of a video-object database system, *IEEE Transactions on Knowledge and Data Engineering* **5**: 629–643.

Othman, H. & Aboulnasr, T. (2003). A separable low complexity 2d hmm with application to face recognition, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **25**: 1229–1238.

Palm, C. (2004). Color texture classification by integrative co-occurrence matrices, *Pattern Recognition* **37**(5): 965 – 976.
URL: *http://www.sciencedirect.com/science/article/pii/S0031320303003686*

Park, D. K., Jeon, Y. S. & Won, C. S. (2000). Efficient use of local edge histogram descriptor, *Proceedings of the 2000 ACM workshops on Multimedia*, MULTIMEDIA '00, ACM, New York, NY, USA, pp. 51–54.
URL: *http://doi.acm.org/10.1145/357744.357758*

Peng, Y. & Ngo, C.-W. (2005a). Emd-based video clip retrieval by many-to-many matching, *CIVR*, pp. 71–81.

Peng, Y. & Ngo, C.-W. (2005b). Emd-based video clip retrieval by many-to-many matching., *CIVR'05*, pp. 71–81.

Pentland, A., Picard, R. W. & Sclaroff, S. (1995). Photobook: Content-based manipulation of image databases.

Perronnin, F., Dance, C., Csurka, G. & Bressan, M. (2006). Adapted vocabularies for generic visual categorization, *In ECCV*, pp. 464–475.

Philbin, J., Chum, O., Isard, M., Sivic, J. & Zisserman, A. (2007). Object retrieval with large vocabularies and fast spatial matching, *Computer Vision and Pattern Recognition, 2007. CVPR 2007. IEEE Conference on.*

Ponceleon, D., Srinivasan, S., Amir, A., Petkovic, D. & Diklic, D. (1998). Key to effective video retrieval: effective cataloging and browsing, *Proceedings of the sixth ACM international conference on Multimedia*, MULTIMEDIA '98, ACM, New York, NY, USA, pp. 99–107.
URL: *http://doi.acm.org/10.1145/290747.290760*

Qi, Y., Hauptmann, A. & Liu, T. (2003). Supervised classification for video shot segmentation, *Proceedings of the 2003 International Conference on Multimedia and Expo - Volume 1*, ICME '03, IEEE Computer Society, Washington, DC, USA, pp. 689–692.
URL: *http://dl.acm.org/citation.cfm?id=1153922.1154326*

Qu, W., Bashir, F. I., Graupe, D., Khokhar, A. & Schonfeld, D. (2005). A motion trajectory based video retrieval system using parallel adaptive self organizing, *Maps, International Joint Conference on Neural Networks, July 31 - Aug*, p. 2005.

Ren, H., Lin, S., Zhang, D., Tang, S. & Gao, K. (2009a). Visual words based spatiotemporal sequence matching in video copy detection, *Multimedia and Expo, 2009. ICME 2009. IEEE International Conference on*, pp. 1382 –1385.

Ren, H., Lin, S., Zhang, D., Tang, S. & Gao, K. (2009b). Visual words based spatiotemporal sequence matching in video copy detection, *Proceedings of the 2009 IEEE international conference on Multimedia and Expo*, ICME'09, IEEE Press, Piscataway, NJ, USA, pp. 1382–1385.
URL: *http://portal.acm.org/citation.cfm?id=1698924.1699265*

Rubner, Y., Tomasi, C. & Guibas, L. J. (1998). A metric for distributions with applications to image databases, pp. 59–66.

Savarese, S., Winn, J. & Criminisi, A. (2006). Discriminative object class models of appearance and shape by correlatons, *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2*, CVPR '06, IEEE Computer Society, Washington, DC, USA, pp. 2033–2040.
URL: *http://dx.doi.org/10.1109/CVPR.2006.102*

Shang, L., Yang, L., Wang, F., Chan, K.-P. & Hua, X.-S. (2010). Real-time large scale near-duplicate web video retrieval, *Proceedings of the international conference on Multimedia*, MM '10, ACM, New York, NY, USA, pp. 531–540.
URL: *http://doi.acm.org/10.1145/1873951.1874021*

Shi, J. & Malik, J. (2000). Normalized cuts and image segmentation, *Technical report*.

Sivic, J. & Zisserman, A. (2006). Video Google: Efficient visual search of videos, *in* J. Ponce, M. Hebert, C. Schmid & A. Zisserman (eds), *Toward Category-Level Object Recognition*, Vol. 4170 of *LNCS*, Springer, pp. 127–144.

Skov, M., Larsen, B. & Ingwersen, P. (2008). Inter and intra-document contexts applied in polyrepresentation for best match ir, *Inf. Process. Manage.* **44**(5): 1673–1683.
URL: *http://dx.doi.org/10.1016/j.ipm.2008.05.006*

Smeaton, A. (2006). Techniques used and open challenges to the analysis, indexing and retrieval of digital video, *Information Systems* **2006**.

Smeaton, A. F., Over, P. & Kraaij, W. (2006). Evaluation campaigns and trecvid, *MIR '06: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*, ACM Press, New York, NY, USA, pp. 321–330.

Smeaton, A. F., Over, P. & Kraaij, W. (2009). High-Level Feature Detection from Video in TRECVid: a 5-Year Retrospective of Achievements, *in* A. Divakaran (ed.), *Multimedia Content Analysis, Theory and Applications*, Springer Verlag, Berlin, pp. 151–174.

Smeaton, A. F., Wilkins, P., Worring, M., de Rooij, O., Chua, T.-S. & Luan, H. (2008). Content-based video retrieval: Three example systems from trecvid, *Int. J. Imaging Syst. Technol.* **18**(2-3): 195–201.
URL: *http://dx.doi.org/10.1002/ima.v18:2/3*

Smeulders, A. W. M., Worring, M., Santini, S., Gupta, A. & Jain, R. (2000). Content-based image retrieval at the end of the early years, *IEEE Trans. Pattern Anal. Mach. Intell.* **22**(12): 1349–1380.
URL: *http://dx.doi.org/10.1109/34.895972*

Su, Y. & Jurie, F. (2011). Visual word disambiguation by semantic contexts, *Computer Vision (ICCV), 2011 IEEE International Conference on*, pp. 311–318.

Tu, Z. & Zhu, S.-C. (2002). Image segmentation by data-driven markov chain monte carlo, *IEEE Trans. Pattern Anal. Mach. Intell.* **24**: 657–673.
URL: *http://portal.acm.org/citation.cfm?id=513073.513079*

Vasconcelos, N. & Lippman, A. (2000). Statistical models of video structure for content analysis and characterization, *Trans. Img. Proc.* **9**(1): 3–19.
URL: *http://dx.doi.org/10.1109/83.817595*

Wang, F., Jiang, Y.-G. & Ngo, C.-W. (2008). Video event detection using motion relativity and visual relatedness, *Proceeding of the 16th ACM international conference on Multimedia*, MM '08, ACM, New York, NY, USA, pp. 239–248.
URL: *http://doi.acm.org/10.1145/1459359.1459392*

Wang, H., Yuan, J. & Tan, Y.-P. (2011). Combining feature context and spatial context for image pattern discovery, *Proceedings of the 2011 IEEE 11th International Conference on Data Mining*, ICDM '11, IEEE Computer Society, Washington, DC, USA, pp. 764–773.
URL: *http://dx.doi.org/10.1109/ICDM.2011.38*

Wang, J. Z., Li, J. & Wiederhold, G. (2001). Simplicity: Semantics-sensitive integrated matching for picture libraries, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **23**: 947–963.

Weng, L., Li, Z., Cai, R., Zhang, Y., Zhou, Y., Yang, L. T. & Zhang, L. (2011). Query by document via a decomposition-based two-level retrieval approach, *Proceedings of the 34th international ACM SIGIR conference on Research and development in*

Information, SIGIR '11, ACM, New York, NY, USA, pp. 505–514.
URL: *http://doi.acm.org/10.1145/2009916.2009985*

Winn, J., Criminisi, A. & Minka, T. (2005). Object categorization by learned universal visual dictionary, *Proceedings of the Tenth IEEE International Conference on Computer Vision - Volume 2*, ICCV '05, IEEE Computer Society, Washington, DC, USA, pp. 1800–1807.
URL: *http://dx.doi.org/10.1109/ICCV.2005.171*

Wong, C.-F. & Pun, C.-M. (2008). Content-based image retrieval based on rectangular segmentation, *Proceedings of the 7th WSEAS International Conference on Signal Processing*, World Scientific and Engineering Academy and Society (WSEAS), Stevens Point, Wisconsin, USA, pp. 75–80.
URL: *http://portal.acm.org/citation.cfm?id=1404086.1404100*

Wu, L., Hu, Y., Li, M., Yu, N. & Hua, X.-S. (2009). Scale-invariant visual language modeling for object categorization, *Trans. Multi.* **11**(2): 286–294.
URL: *http://dx.doi.org/10.1109/TMM.2008.2009692*

Wu, X., Ngo, C.-W., Hauptmann, A. G. & Tan, H.-K. (2009). Real-time near-duplicate elimination for web video search with content and context, *Trans. Multi.* **11**: 196–207.
URL: *http://dl.acm.org/citation.cfm?id=1652976.1652978*

Youtube (2013). https://www.youtube.com/yt/press/en-gb/statistics.html.

Yuan, J., Wang, H., Xiao, L., Zheng, W., Li, J., Lin, F. & Zhang, B. (2007). A formal study of shot boundary detection, *Circuits and Systems for Video Technology, IEEE Transactions on* **17**(2): 168–186.

Yuan, J. & Wu, Y. (2008). Context-aware clustering, *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pp. 1–8.

Zhang, C., Liu, J., Ouyang, Y., Lu, H. & Ma, S. (2009). Concept-specific visual vocabulary construction for object categorization, *Proceedings of the 10th Pacific Rim Conference on Multimedia: Advances in Multimedia Information Processing*, PCM '09, Springer-Verlag, Berlin, Heidelberg, pp. 936–942.

Zhang, H. & Petkovic, D. (1996). Content-based representation and retrieval of visual media: A state-of-the-art review, *Multimedia Tools and Applications* **3**: 179–202.

Zhang, J., Marszalek, M., Lazebnik, S. & Schmid, C. (2006). Local features and kernels for classification of texture and object categories: A comprehensive study, *Computer Vision and Pattern Recognition Workshop, 2006. CVPRW '06. Conference on*, p. 13.

Zhang, S., Huang, Q., Hua, G., Jiang, S., Gao, W. & Tian, Q. (2010). Building contextual visual vocabulary for large-scale image applications, *Proceedings of the international conference on Multimedia*, MM '10, ACM, New York, NY, USA, pp. 501–510.
URL: *http://doi.acm.org/10.1145/1873951.1874018*

Zhang, S., Tian, Q., Hua, G., Huang, Q. & Li, S. (2009). Descriptive visual words and visual phrases for image applications, *Proceedings of the 17th ACM international conference on Multimedia*, MM '09, ACM, New York, NY, USA, pp. 75–84.
URL: *http://doi.acm.org/10.1145/1631272.1631285*

Zhang, Y., Jia, Z. & Chen, T. (2011). Image retrieval with geometry-preserving visual phrases, *CVPR*, pp. 809–816.

Zhao, W.-L. (2009). http://www.cs.cityu.edu.hk/ wzhao2/ sotu.htm.

Zhao, W.-L., Wu, X. & Ngo, C.-W. (2010). On the Annotation of Web Videos by Efficient Near-Duplicate Search, *IEEE Transactions on Multimedia* **12**(5): 448–461.
URL: *http://dx.doi.org/10.1109/TMM.2010.2050651*

Zheng, Q.-F. & Gao, W. (2008). Constructing visual phrases for effective and efficient object-based image retrieval, *ACM Trans. Multimedia Comput. Commun. Appl.* **5**: 7:1–7:19.
URL: *http://doi.acm.org/10.1145/1404880.1404887*

Zheng, W., Yuan, J., Wang, H., Lin, F. & Zhang, B. (2005). A novel shot boundary detection framework, pp. 596018–596018–11.
URL: *+ http://dx.doi.org/10.1117/12.631547*

Zhou, G., Wang, Z., Wang, J. & Feng, D. (2010). Spatial context for visual vocabulary construction, *Image Analysis and Signal Processing (IASP), 2010 International Conference on*, pp. 176–181.

Zhou, W., Tian, Q., Yang, L. & Li, H. (2010). Latent visual context analysis for image re-ranking, *Proceedings of the ACM International Conference on Image and Video Retrieval*, CIVR '10, ACM, New York, NY, USA, pp. 205–212.
URL: *http://doi.acm.org/10.1145/1816041.1816073*