2015

Ensuring the quality of experience for mobile augmented visual search applications: Fast, low bitrate and high accuracy

Yi Cao
University of Wollongong, yc833@uowmail.edu.au

UNIVERSITY OF WOLLONGONG

School of Electrical, Computer and Telecommunications Engineering

Ensuring the Quality of Experience for Mobile Augmented Visual Search applications: Fast, Low Bitrate and High Accuracy

Yi Cao

"This thesis is presented as part of the requirements for the
Award of the Degree of
Doctor of Philosophy
From
University of Wollongong"

August 2015

ABSTRACT

The explosive growth of advanced smart devices, such as smart phones equipped with a high resolution camera, touch-screen, Wi-Fi and advanced multimedia functionalities, is dramatically changed the way people interact with multimedia content and provides new ways of interacting with traditional media publications such as newspapers and magazines. The Quality of Experience (QoE) of users of these new multimedia services has become a primary concern to ensure their success. The focus of this research is to investigate the user QoE for emerging Mobile Augmented Visual Search (MAVS) applications used to connect readers of printed media with corresponding online digital content. MAVS applications rely on automatically matching a captured visual scene to an image in a database to trigger the retrieval of augmented multimedia content to users. It is a complex issue to find an efficient solution to measure the QoE in such applications due to the diversity of users as well as variation of real world conditions, device limitations and bandwidth of the communication network. On the basis of the investigation of QoE related key influencing factors, the ultimate goal is to find optimal solutions to ensure the user QoE is maximized through this work.

A fast, low bitrate and high matching accuracy MAVS system is presented in this thesis. Focusing on two key influencing factors, namely matching accuracy and waiting time, the investigation starts from the matching accuracy of state-of-the-art local feature algorithms under realistic distortions. A number of local image feature algorithms are studied using various image compression schemes from the point view of matching accuracy using precision @ 1 and processing time. The trade-off between two general architectures (i.e. 1. sending compressed images and performing feature extraction and matching on a server; and 2. performing feature

extraction on the mobile device and sending these to a server for matching) for implementing MAVS applications is examined. The evaluation results suggest that the matching accuracy of sending compressed images at a very low bitrate is comparable to sending compact image features when using a high quality image coder, such as JPEG2000 and HDPhoto. Then, the joint effect of two common distortions, namely illumination changes and image blurring that occur when capturing images by the mobile camera, is investigated for print media when using state-of-the-art local feature algorithms from the aspect of ensuring matching accuracy. The results indicate that illumination changes have more influence on matching accuracy compared to image blurring and different cameras also influence the performance of local feature algorithms. Thus, flexible feature selection algorithms are required to improve the matching accuracy for MAVS applications within a heterogeneous camera phone environment.

On the basis of the investigation of the matching accuracy of various state-of-the-art local feature algorithms, two accurate and low bitrate MAVS systems are presented. One fast and accurate low bit rate system is proposed based on extracting Scale Invariant Feature Transform (SIFT) features from images reconstructed from the low spatial frequency components. The system applies a two-dimensional block-based Discrete Cosine Transform (DCT), as widely used in image coders on the mobile devices, and only encodes and transmits the resulting DC components at a low bitrate. This system achieves high matching accuracy of more than 97% precision @ 1 and is robust to a wide range of typical image distortions including scaling, rotation, additive noise, image blurring and illumination. Transmission data rates are comparable to existing compressed domain image features whilst significantly reducing system latency. For the adaptive purpose in the MAVS system,

another MAVS system based on feature selection to achieve low bitrate transmission while maintaining high matching accuracy is proposed. Novel feature selection methods are proposed, based on the entropy of the image content, entropy of extracted features and the DCT coefficients. The proposed methods are robust against complex real world capturing distortions and achieve better retrieval accuracy under low bit rate transmission than state-of-the-art peak based feature selection used within the MPEG-7 Compact Descriptor for Visual Search (CDVS).

For practical use of MAVS applications, the waiting time is studied as the primary perception, which significantly affects the user's QoE. A subjective experiment is conducted to study the impact factors in the MAVS applications, including the influence of linking different media types and using different progress bar indicators. The experimental results are compared to the traditional mouse-click-based-multimedia applications, which suggest that a logarithmic function of waiting time associated with QoE can be found. On the basis of the results from the subjective experiment, a QoE estimation approach based on waiting time and matching accuracy is studied by performing retrieval experiments on a realistic image dataset with real-world distortions caused by image capture. The ultimate goal is to achieve MOS greater than 4 ("4" stands for "good" in the MOS scale) in the proposed MAVS systems, which refers to achieving good QoE in this thesis. The predicted QoE using proposed QoE model proves that the proposed MAVS systems can provide good QoE to users under varying transmission conditions.

Thesis Certification

I, Yi Cao, declare that this thesis, submitted in partial fulfilment of the requirements for the award of Doctor of Philosophy, in the School of Electrical, Computer, and Telecommunications Engineering, University of Wollongong, is wholly my work unless otherwise referenced or acknowledged. The document has not been submitted for qualifications at any other academic institution.

Yi Cao

24 August 2015

ACKNOWLEDGEMENTS

I would like to sincerely thank my supervisors Dr. Christian Ritz and Dr. Raad Raad for their help, guidance, encouragement, support and inspiring suggestions that they have given to me for my research.

I also would like to thank the Smart Services CRC for the Phd scholarship and the research, student conferences that help me to expand my horizon on my research.

To my wife Yiwei and my parents for their love, support, patience and sacrifice during my whole Phd study, without which I could not be able to finish this degree.

A big thank to my Lab colleagues, Kevin, Samir, Shujau, Stephen, Xiguang, Jacob, Shahab, Yuxiao and Chao for their friendly suggestions and advice through my whole research, which opened my mind.

Many thanks to my family and friends for all the company and support, especially thank to the staff in Faculty of Engineering & Information Sciences for an active, and inspiring environment for research.

To my new born baby Shurui, my new motivation for life, I hope I make you proud.

TABLE OF CONTENT

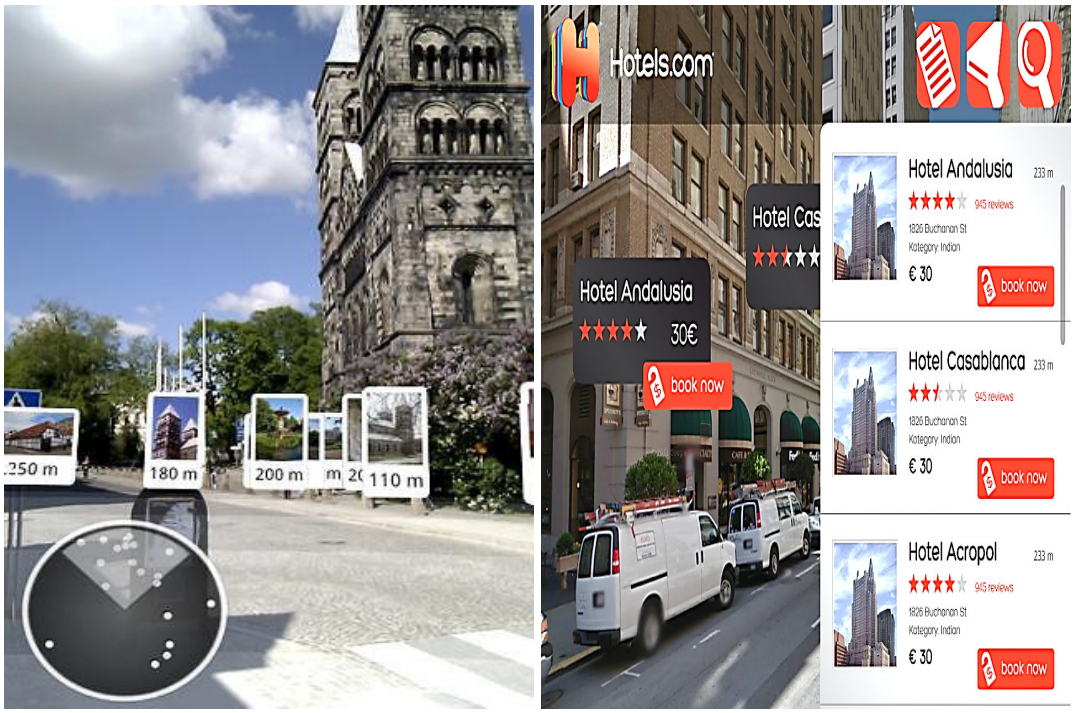
LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

1.1 Overview

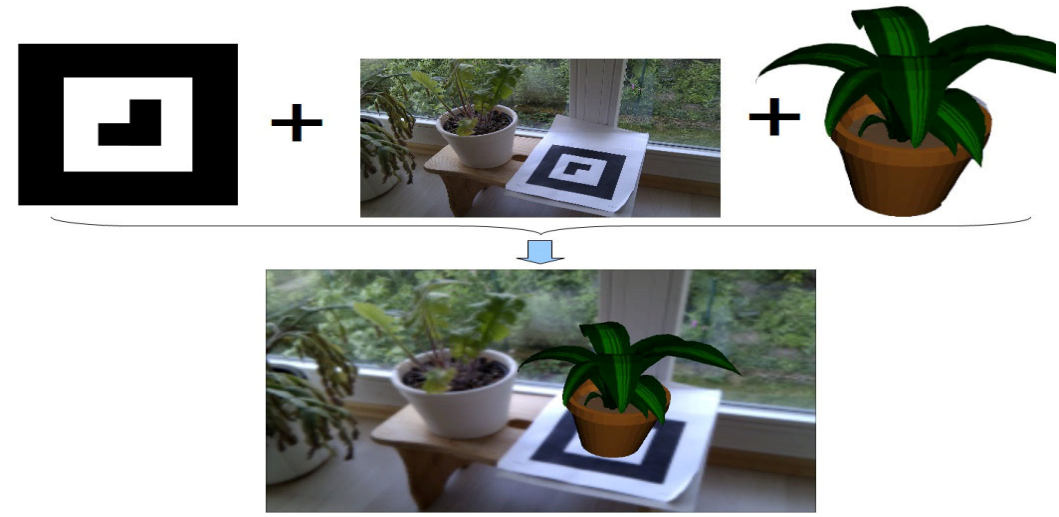
The concept of Augmented Reality (AR) has been around for decades. Since 1901, the first AR idea of a “character maker” was mentioned by L. Frank Baum in his novel The Master Key [1], where the augmented information was displayed through electronic spectacles. Ivan Sutherland developed the first AR head-mounted devices in 1968 [2] which linked the real world with the virtual world. In the 1990s, several AR prototype systems were developed, to name a few, the first fully functional immersive AR system named “Virtual Fixtures” developed by Louis Rosenberg [3] for enhancing operator performance; a head-mounted AR prototype to help maintenance for a laser printer developed by Steven Feiner et al. [4]; the first AR application which overlays map data onto video for 3D flight guidance developed by Michael Abernathy et al. [5]. These early AR systems were deployed in large size computers to fulfil the requirements of processing large amounts of visual data and real-time performance. Following the development of ARToolKit, which was created by Hirokazu Kato in 1999 [6], the first handheld the AR system was investigated by Dieter Schmalstieg et al. on an unmodified Personal Digital Assistant (PDA) with a commercial camera using Marker Tracking (MT) [7]. With the explosive development of modern multimedia technologies and services, traditional media services are being challenged by the new generation of augmented internet media. Since then, with the rapid expansion of more and more powerful smart mobile devices mounted with a variety of features, for example, high resolution camera, high definition touch screen, Wi-Fi and rich multimedia functionality such as video, voice and audio, mobile AR applications appear more and more.



(a) (b)

Figure 1.1 Examples of AR applications using GPS and LBS. (a): Popular photos near the current location are displayed in an AR-based view; (b): A list shows the augmented surrounding accommodation information on a camera view.

AR applications on the mobile devices are often described with reference to their two predominant modes [8]. One is a Location Based Service (LBS) that utilises a Global Position System (GPS) sensor, which makes use of user's geographic location and position to discern nearby objects [9] and then overlays the information on the camera's view to provide an augmented experience to users as shown in Figure 1.1. Figure 1.1-(a) shows an AR application named "Photos Around" which retrieves the most popular photos from Panoramio.com based on user's current location and display the images in augmented reality. The details and websites of the images can be accessed by tapping the images [10]. Figure 1.1-(b) shows an AR application for finding hotel accommodation. The price and the user's



(a)

(b)

Figure 1.2 Examples of AR applications using computer vision technologies. (a): An Android AR application named "AndAR" shows virtual object on a Marker, which is based on ARtoolKit. (b): Google goggles.

rating of the hotels around the user's current location are displayed by the augmented reality application [11]. These types of applications make use of the users' Point of Interest (POI) or tag made by users. The other kind of AR applications are based on computer vision, which use the image matching or object recognition technologies to process the image captured by a camera to retrieve relevant visual content or predefined virtual objects from a remote server or cloud as shown in Figure 1.2. Figure 1.2-(a) shows an AR application named "AndAR" running on Android

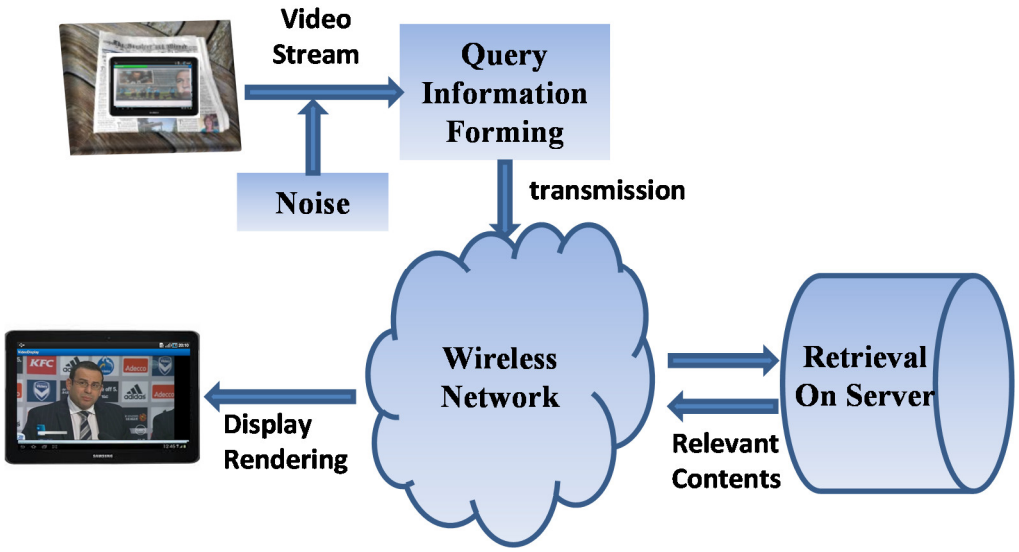


Figure 1.3 System diagram of a MAVS application.

platform which makes use of MT technology to display a virtual object on top of a marker [12]. Figure 1.2-(b) shows a well-known mobile AR application named "Google Goggles" developed by Google，which enables a user to use mobile phone camera to recognize product logos, landmarks, artworks and business cards to perform non-text search for relevant information [13].

The prosperous development of smart devices and modern technologies in signal processing, image processing and computer vision are dramatically changing the way people interact with multimedia content. In this thesis, an interactive mobile application named as Mobile Augmented Visual Search (MAVS) application is studied. Similar to the applications as described in [14]–[16], the targeted MAVS application can link user captured print media to a rich multimedia repository and bring augmented experiences to users by automatically matching captured scenes from a mobile camera to reference multimedia content in a database, such as video, a picture gallery and webpages as shown in Figure 1.3. The targeted MAVS application matches the captured image with a set of pre-defined images in the remote database. For each image in the database, there is only one corresponding

multimedia content that is linked. Once a match is found, the linked corresponding multimedia content automatically starts to playback. As there is only one correct match, it is desirable that the targeted MAVS application can achieve high matching accuracy that the first returned result (i.e. precision @ 1) is the correct correspondence to the captured image. There are several unique challenges for such interactive mobile applications. The captured scenes are needed to be processed in real time to generate meaningful query information whilst the computation capacity of mobile devices is often limited. The noise occurred during the capture varies with the real world conditions and increases the difficulty to extract accurate query information from captured scenes. The transmission bitrate of the query information should be as compact as possible because of spotty data transmission bandwidth. The retrieved content should correctly correspond to the captured scenes while more accurate algorithms normally require more processing time and transmission bandwidth, which influence the real time performance. Optimization of state-of-the-art technologies and development of new technologies are demanded to achieve good performance perceived by end users. Therefore, the emphasis of this research is to develop an innovation to maximize the user Quality of Experience (QoE) for MAVS applications which link print media, such as newspaper or magazine, with multimedia content, such as a website, picture gallery or video. It is a complex issue to ensure the QoE in such new emerging applications because such applications are the aggregator of state-of-the-art technologies in the ICT and there is no well-established QoE definition for MAVS applications. It naturally raises the question of how the quality perceived by users can be ensured. It is difficult to measure the QoE by using an online method as the users' perception requires subjective measurement which is usually performed offline. To answer this question, the QoE is studied

through two Key Impact Factors (KIFs) namely matching accuracy and waiting time in this thesis. It is intuitive that the MAVS applications should be as accurate as possible. If the wrong content has been returned to a user, it would degrade the user's satisfaction dramatically. Besides, the system response from the user's image capture to the multimedia content delivery should be as fast as possible to fulfil the real-time requirement of the MAVS applications. Too long a system delay is not acceptable in the MAVS applications. The ultimate goal is to propose an MAVS system which can achieve high matching accuracy meanwhile minimising the waiting time (e.g. processing delay and transmission delay) to result in good QoE (i.e. a MOS above 4 when users rate a MAVS service). Researches are conducted from various aspects, such as low complexity processing, low bitrate transmission, high matching accuracy under realistic distortion, and subjective evaluation and prediction of QoE, to develop efficient solutions to ensure the QoE is maximized in MAVS applications in this thesis.

1.2 Contribution

Due to the unique challenge of returning a sole correct result to users in the targeted MAVS applications, a novel evaluation using precision @ 1 is employed to study the performance of various state-of-art feature algorithms under complicated realistic distortions. On the basis of the evaluation, novel MAVS systems that achieve fast, low bitrate transmission and high accuracy are proposed. A new QoE estimation model for the targeted MAVS applications is proposed based on an extensive subjective experiment. The detailed contributions of this thesis are listed as follows, with reference to the list of publications arising from the research described in this thesis provided in section 1.3:

1. To achieve low bitrate transmission while keeping matching accuracy high, the trade-off for different MAVS application approaches (i.e. sending compressed images or sending image features) is discussed from the point view of QoE [C1], [C6].

2. A number of local image feature algorithms are studied using various image compression schemes from the point view of matching accuracy and processing time in the context of MAVS application. Based on the evaluation of using a new matching accuracy measurement named precision @ 1, which is specifically used for MAVS applications, results suggest that the matching accuracy of sending compressed images is comparable to sending compact image features for the targeted MAVS applications when using a high quality image coder, such as JPEG2000 and JPEGXR [C1].

3. The joint effect of the illumination changes and image blurring, which commonly occurs when capturing images using a mobile phone camera, is studied from image matching accuracy for print media when using state-of-the-art local feature algorithms. The evaluation is performed on a database of real camera images captured by two different camera models. Results suggest that the illumination changes have a more negative effect on matching accuracy compared to image blurring and this influence is camera-dependent [C2].

4. The influence of waiting time for the targeted MAVS applications is studied by conducting an extensive subjective experiment. The influence factors, including different user interaction methods (i.e. from traditional mouse click-based operation to camera-capture based operation), different multimedia types and different indicators (i.e. progress bar and spinning wheel), are studied from the users' satisfaction and acceptance in terms of waiting time. The result can be used as a

guidance to help the MAVS system designer to deploy and balance different technologies for maximizing the QoE perceived by users [C3].

5. Based on the investigation of the SIFT feature associated with the loss of spatial frequency information in the DCT domain, a new low bit rate, low complexity, low latency MAVS system with high accuracy is proposed. The novel system uses SIFT features extracted from low spatial frequency components represented by encoded block-based 2D DCT coefficients. The proposed system is proven to be robust against various image distortions, including additive white Gaussian, global illumination changes, out-of-focus blur, rotation and scaling, which commonly occur in the mobile visual search environment [C4].

6. Novel feature selection methods are proposed, based on the entropy of the image content, entropy of extracted features and the Discrete Cosine Transformation (DCT) coefficients. The methods proposed in the descriptor domain and DCT domain achieve better retrieval accuracy under low bit rate transmission than state-of-the-art peak based feature selection used within the MPEG-7 Compact Descriptor for Visual Search. The robustness of the proposed methods is evaluated under controlled single distortion and the retrieval performance is verified from image retrieval experiments and results for a realistic dataset with complex real world capturing distortion. The proposed method can improve the matching accuracy for various detectors and also indicate that the feature selection can not only achieve low bit rate transmission but also results in a higher matching accuracy than using all features when applied to distorted images [C5], [J1].

7. A QoE estimation for start-of-the-art feature selection in MPEG-7 CDVS is analysed based on waiting time and matching accuracy as judged by retrieval experiments on a realistic image dataset with real-world distortions caused by image

capture. The predicted QoE results suggest that feature selection can provide good QoE to users [C6].

1.3 Publications

1.3.1 Journal publications

[J1] Yi Cao; Ritz, C.; Raad, R., “Feature selection for low bit rate mobile augmented reality applications”. *Signal Processing: Image Communication*, Vol.36, pp.115-126, Aug, 2015

1.3.2 Conference publications

[C1] Yi Cao; Ritz, C.; Raad, R., "Image compression and retrieval for Mobile Visual Search," *Communications and Information Technologies (ISCIT), 2012 International Symposium on* , vol., no., pp.1027,1032, 2-5 Oct. 2012

[C2] Yi Cao; Ritz, C.; Raad, R., "The joint effect of image blur and illumination distortions for Mobile Visual Search of print media," *Communications and Information Technologies (ISCIT), 2013 13th International Symposium on* , vol., no., pp.507,512, 4-6 Sept. 2013

[C3] Yi Cao; Ritz, C.; Raad, R., "How much longer to go? The influence of waiting time and progress indicators on quality of experience for mobile visual search applied to print media," *Quality of Multimedia Experience (QoMEX), 2013 Fifth International Workshop on* , vol., no., pp.112,117, 3-5 July 2013

[C4] Yi Cao; Ritz, C.; Raad, R., "Fast and accurate low bit rate retrieval-by-capture applications," *Digital Signal Processing (DSP), 2014 19th International Conference on* , vol., no., pp.657,662, 20-23 Aug. 2014

[C5] Yi Cao; Ritz, C.; Raad, R., "Adaptive and robust feature selection for low bitrate mobile augmented reality applications," *Signal Processing and Communication Systems (ICSPCS), 2014 8th International Conference on* , vol., no., pp.1,7, 15-17 Dec. 2014

[C6] Yi Cao; Ritz, C.; Raad, R., "Quality of experience-based image feature selection for mobile augmented reality applications," *Signal Processing and*

Communication Systems (ICSPCS), 2014 8th International Conference on , vol., no., pp.1,6, 15-17 Dec. 2014

1.4 Thesis outline

This thesis is organised as follows:

A literature review is given in chapter 2 starting from fundamental knowledge of digital imaging on the mobile device camera in the targeted MAVS applications as well as the image compression technologies. Then, quality of experience, which includes definition of QoE, QoE modelling, objective and subjective QoE measurement, is reviewed from the point view of maximising the QoE perception in the targeted MAVS applications. Targeting two key influence factors that are waiting time and matching accuracy in the MAVS system, the studies in the field of waiting time as well as its quality perception are firstly reviewed in two multimedia services: a web service and a streaming content service. Secondly, the state-of-the-art techniques especially in feature detection and extraction, pairwise image matching, and efficient image retrieval are reviewed from the aspect of ensuring the matching accuracy. The feature selection is also reviewed from the point view of selecting the most important features to accelerate the image matching and retrieval speed meanwhile improving the matching accuracy. The rest of this chapter reviews an on-going MPEG-7 standard called MPEG-7 Compact Descriptor for Visual Search, which is closely related to the work in this thesis.

Chapter 3 starts from an extensive study of the performance of different feature detector and different descriptors under varying compression ratios using three different image compressors. Based on the study, the trade-off of different MAVS approaches is discussed both from the point view for matching accuracy and processing delay. Then, the joint effect of two common distortions, namely

illumination change and image blurring, is examined for print media (i.e. book covers, DVD covers, and museum paintings) with keypoints clustering and several state-of-the-art features.

Chapter 4 presents solutions for a fast, accurate and low bit rate MAVS application. Firstly, a solution for low bit rate transmission using low frequency DCT coefficients is proposed. The principle of the low frequency response of the SIFT feature is studied. Based on this theoretical exploration, a system using low frequency DCT coefficients is proposed. The performance of the matching accuracy under various realistic distortions is examined. The processing time and memory consumption are reduced. Secondly, low bit rate transmission by using feature selection is studied. Novel feature selection methods are proposed, based on the entropy of the image content in the keypoint domain, the entropy of the extracted features in the descriptor domain and the Discrete Cosine Transformation (DCT) coefficients in the compressed domain. The performance of proposed feature selection schemes is verified from image retrieval experiments and results for a realistic dataset with complex real world capturing distortion including varying lighting conditions, perspective distortion, foreground and background clutter.

QoE estimation based on waiting time and matching accuracy is proposed in chapter 5. A subjective experiment to study the users' experience in term of waiting time in the context of MAVS application is conducted from the aspects of changed interaction between users and multimedia content, the QoE influence of different media types and the QoE influence of different indicators. Then, the satisfaction and acceptance of users for the waiting time in the context of an MAVS application is analysed. QoE estimations for employing start-of-the-art feature selection in MPEG-7 CDVS, the proposed feature selection method and low frequency DCT coefficients

in MAVS applications are then analysed based on the subjective experiment results and matching accuracy as judged by retrieval experiments on a realistic image dataset with real-world distortions caused by image capture.

Chapter 6 provides conclusions and suggestions of this thesis and further work.

2 AN OVERVIEW OF MAVS APPLICATIONS

2.1 Introduction

In this chapter, the literature and technologies related to the current development of MAVS applications are reviewed. The investigation starts from the digital image photography on the mobile phone camera. Considering the low bitrate transmission of visual information captured by a mobile phone, image compression technologies including JPEG, JPEG-XR and JPEG2000 are reviewed. Then, the definition and current development and standardization of QoE for multimedia applications in International Telecommunication Union Telecommunication Standardization Sector (ITU-T) [17] and European Network on Quality of Experience in Multimedia Systems and Services (QUALINET) [18] are reviewed. Following the QoE definition of targeted MAVS applications, the investigation of a primary key impact factor namely waiting time which influences the QoE is reviewed from the point view of the real time requirement for MAVS applications. The detailed technologies normally employed in the MAVS applications, including content-based feature detection and extraction, pair-wise image matching, feature selection and image retrieval are then discussed. The measurement of the performance for MAVS applications is reviewed in terms of the feature matching accuracy and retrieval accuracy. Finally, possible architectures for the targeted MAVS applications system are presented as well as the recent development of the MPEG-7 Compact Descriptor for Visual Search (CDVS) standard [19].

2.2 Digital imaging on mobile phone camera

The mobile phone camera equipped in the current smart device is a digital camera that encodes digital images. The technologies underneath a digital camera can be broadly categorised into a sensor and a program. There are two major types of sensors, Charge-Coupled Device (CCD) and Complementary metal–oxide–semiconductor (CMOS). The principle of a sensor is to capture the light and then convert the light to electrical signals. A program embedded in the digital camera firmware then translates the electrical signals into discrete signals, known as pixels. Each pixel presents the intensity of the light that hits at a given point on the sensor, the brighter the light is, the larger value that pixel has. Finally, all the pixels derived from the sensor form a digital image [20], [21]. The digital image is then compressed by an image encode for efficient storage and transmission.

During the image capture when using a mobile phone camera in a MAVS system, two major distortions normally occur, known as geometric distortion and photometric distortion. The geometric distortions investigated in this thesis are rotation, scaling and out-of-plane which are determined by the relative position between the camera and the object. These distortions result in the geometric transformation of the captured object. The photometric distortions, including illumination variation and image blurring, are caused by the poor environmental lighting condition and out-of-focus, respectively, which reduce the sharpness and contrast of the captured image. In practice, the geometric and photometric distortions are produced simultaneously and deteriorate the image quality [22], [23]. Therefore, it is difficult to match the captured image to a predefined image due to these distortions in a MAVS system.

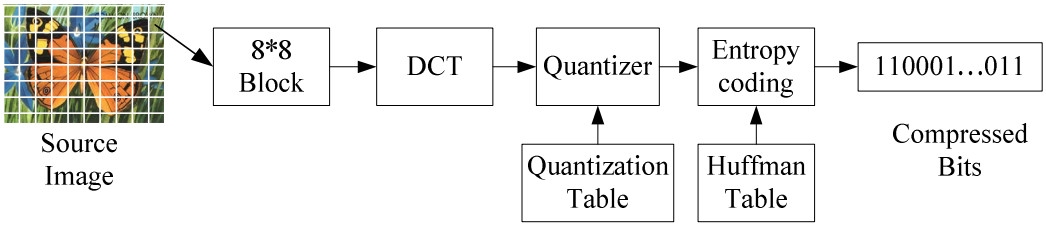


Figure 2.1 JPEG compression diagram

2.3 Image compression

The majority of MAVS applications normally operate with Client/Server architecture. The captured visual information is interactively processed between client side and server side. However, due to the increasing number of pixels on the mobile devices, the raw digital visual information from image acquisition contains a large amount of data [24]–[27]. An efficient form to remove the coding redundancy, inter-pixel redundancy and psych visual redundancy known as image compression is widely deployed on mobile devices for data transmission and storage [28]. In this session, three popular image compression technologies, including Joint Photographic Experts Group (JPEG), JPEG eXtended Range (JPEGXR) and JPEG 2000, are reviewed. Moreover, the image is used for matching not for viewing in an MAVS system. The distortion caused by image compression degrades the visual quality of images and the pixel values of an image are then changed. The features derived from the statistical information of pixel values are changed too. Therefore, the image compression affects the performance of the features for image retrieval [29]–[31]. Therefore, the influence of highly compressed images by different compressors at a low bit rate is studied from the aspect of image matching accuracy when employing state-of-art features in a MAVS system in this thesis.

The JPEG compression is widely used in the digital image system, which is based on the Discrete Cosine Transform (DCT). The diagram of the JPEG

compression is shown in Figure 2.1. The source image is subdivided into blocks of 8*8 pixel size. The 2D-DCT transform is performed at each block to transfer image block signal from spatial domain to frequency domain using (2.1):

$$T(u,v) = \sum_{x=0}^{M-1}\sum_{y=0}^{N-1} f(x,y)\alpha(u)\alpha(v)cos\frac{(2x+1)u\pi}{2M}cos\frac{(2y+1)v\pi}{2N} \tag{2.1}$$

$$\alpha(u) = \begin{cases}\sqrt{1/M} & u=0\\ \sqrt{2/M} & u=1,\cdots M-1\end{cases}, \alpha(v) = \begin{cases}\sqrt{1/N} & v=0\\ \sqrt{2/N} & v=1,\cdots N-1\end{cases}$$

where $T(u,v)$ is the DCT coefficients (8*8 array), *u,v* is the index of DCT coefficients, *M*=*N*=8, *f(x, y)* is the pixel value of the block. The result for each block is an 8*8 coefficient array in which first top-left coefficient is known as the DC (zero-frequency) component and the other coefficients are known as AC (high frequency) component. The higher AC components represent higher vertical and horizontal spatial frequencies. The DCT coefficients can be used as an efficient feature for image retrieval [32], [33]. The detailed JPEG information can be found in [34]–[36]. The compression ratio of JPEG is controlled by a quality factor *Q* being used in the quantization procedure ranging from 1 to 100. Normally, the compression is invisible to human eyes by setting *Q* above 75 while extreme compression can be done by setting *Q* below 10 to achieve a low bit rate, at the expense of image quality (i.e. distinctly block artefacts [37]).

The JPEGXR was developed by Microsoft, originally known as HD Photo, for compression of continuous tone photographic content [38] and became ITU-T Recommendation T.832 on 2012 [39]. The compression scheme of JPEGXR is similar to JPEG including fixed block subdivision, spatial to frequency space transform, frequency coefficients quantization and entropy coding, but with some

improvements over JPEG to achieve better compression with equivalent visual quality [40]. The detailed technology used in JPEGXR can be found in [38]–[40].

JPEG 2000 is a wavelet-based compression standard for still images, which provides a better compression performance and flexibility such as scalability and editability compared to JPEG. The JPEG2000 reduces the blocking artefacts that occur in JPEG at high compression ratios by means of aforementioned technologies to achieve approximately 20% compression gain over JPEG [41]. The detailed technologies specifications can be found in [42]–[44].

2.4 QoE definition, model and measurement

To ensure the QoE for new emerging applications and services, it is essential to understand what QoE is and how to model and evaluate the QoE. In this section, the current development of QoE is reviewed at a high level including the research activities in ITU-T [45] and QUALINET [46]. Following the definition of QoE, the challenges to ensure the QoE for emerging mobile multimedia services and applications in the mobile devices are introduced and discussed.

2.4.1 QoE definition

The QoE has become a crucial element for deploying a successful multimedia service or novel application, which has gained more and more attention both in academy and industry. However, there was no clear or widely accepted definition of QoE for decades as QoE is a multidisciplinary field including the influencing factors not only from the multimedia technology being evaluated but also from the human user. A clear and widely accepted QoE definition is emerging with the publication of the latest white paper from QUALINET. Here, several definitions are presented in

Table 2-1 The evolution of the definition of Quality of Experience

Time	QoE definition
2001	QoE is an extension of the traditional QoS in the sense that QoE provides information regarding the delivered services from an end-user point of view [29].
2004	QoE is closely related to the traditional concept of utility functions as high-level forms of requirement specification. Both QoE and utility functions allow to set degrees of desirability for some given levels of delivered QoS [30].
2006	QoE is how a user perceives the usability of a service when in use – how satisfied he/she is with a service in terms of, e.g., usability, accessibility, retainability and integrity [31].
2007	The overall acceptability of an application or service, as perceived subjectively by the end-user [26], [32]. NOTE 1 – Quality of Experience includes the complete end-to-end system effects (client, terminal, network, services infrastructure, etc.). NOTE 2 – Overall acceptability may be influenced by user expectations and context.
2009	QoE describes the degree of delight of the user of a service, influenced by content, network, device, application, user expectations and goals, and context of use [33].
2011	QoE is a set of human centric factors based on human subjective and objective cognitive aspects arising from the interaction of a person with technology and with business entities in a particular context [34].
2012	QoE is a blueprint of all human subjective and objective quality needs and experiences arising from the interaction of a person with technology and with business entities in a particular context [35].
2013	QoE is the degree of delight or annoyance of the user of an application or service. It results from the fulfilment of his or her expectations with respect to the utility and / or enjoyment of the application or service in the light of the user's personality and current state [27].

Table 2-1, which show the evolution of the definition of QoE and how the QoE differences from traditional Quality of Service (QoS).

One interesting observation is that the ITU-T define QoS as " *the collective effect of service performance which determine the degree of satisfaction of a user of the service*" which explicitly emphasizes the importance of "*the degree of satisfaction of a user*" in 1994 [47]. Although the original QoS definition is a user-centric concept which distinguishes with a purely technical concept, most of the work related to QoS solely focused on objective and technical performance measurements, such as data flow management of bit rate, delay, jitter, packet loss and bit error rate. This reduces the QoS to a technology-centric concept and a majority of work were conducted at the network or system-level to measure the quality. In order to reinforce the role of the user in the quality assessment, "*the user's perception",* "*degrees of desirability*", "*perceived usability*" are introduced in the first three definitions in Table 2-1 to link the 'user' with QoS. The ITU-T FG IPTV group and ITU-T Recommendation P.10/G.100 Amendment 1 define the QoE as user's "acceptability" and clearly note that client is a vital part of the QoE in the end-to-end system effects and the user's expectation and context will influence the QoE as well. So far, the QoE definitions are getting mature. As stated in the last four definitions in Table 2-1, The QoE is the degree of delight of a user for an application or service. Such "degree" is influenced by objective technical factors (e.g. content, network, device, and application), subjective human factors (e.g. user expectations and goals, context, personality and current state) and the interaction between these objective and subjective factors. The QoE will evolve over the time along with the overwhelming development of new applications and services, such as emerging MAVS applications. It raises the question of how to define a proper definition for

this particular application in addition to finding the significant objective technology factors and subjective factors to maximise the QoE for MAVS applications, which is a key focus of this thesis.

2.4.2 QoE models

The fundamental research question for QoE is how to operationalize the concept in terms of the assessment and application in a reliable, valid, efficient and objective way both for objective technical factors and subjective human factors. The question of "How can we quantify user perceived quality and how can we measure and maximise the quality" is a challenge since the QoE encompasses the user's perception, while quality measurement merely from traditional QoS related quality assessment in the technology level is not sufficient or completely applicable. In order to consider human factors such as a user's expectation, personality and context, quality assessment schemes not only measuring objective technology factors but also a user's perception or experience are needed on the basis of a proper QoE model. Prior QoE models which attempt to integrate the human factors and technical factors are illustrated in Figure 2.2 and elaborated upon below.

The ITU-T G.1080 defined a QoE model in 2008, which clearly divides the QoE into two groups, namely subjective human component and objective QoS factors as shown in Figure 2.2-(a), each has individual effect on QoE [48]. In this model, the QoS is used as measurable metrics for objective QoE measurement. There is no doubt that that the QoS factors will influence the QoE. Nevertheless, if the QoS factors could directly represent the objective factors of QoE, this is questionable as QoE is based on human centric factors, not technology centric factors.

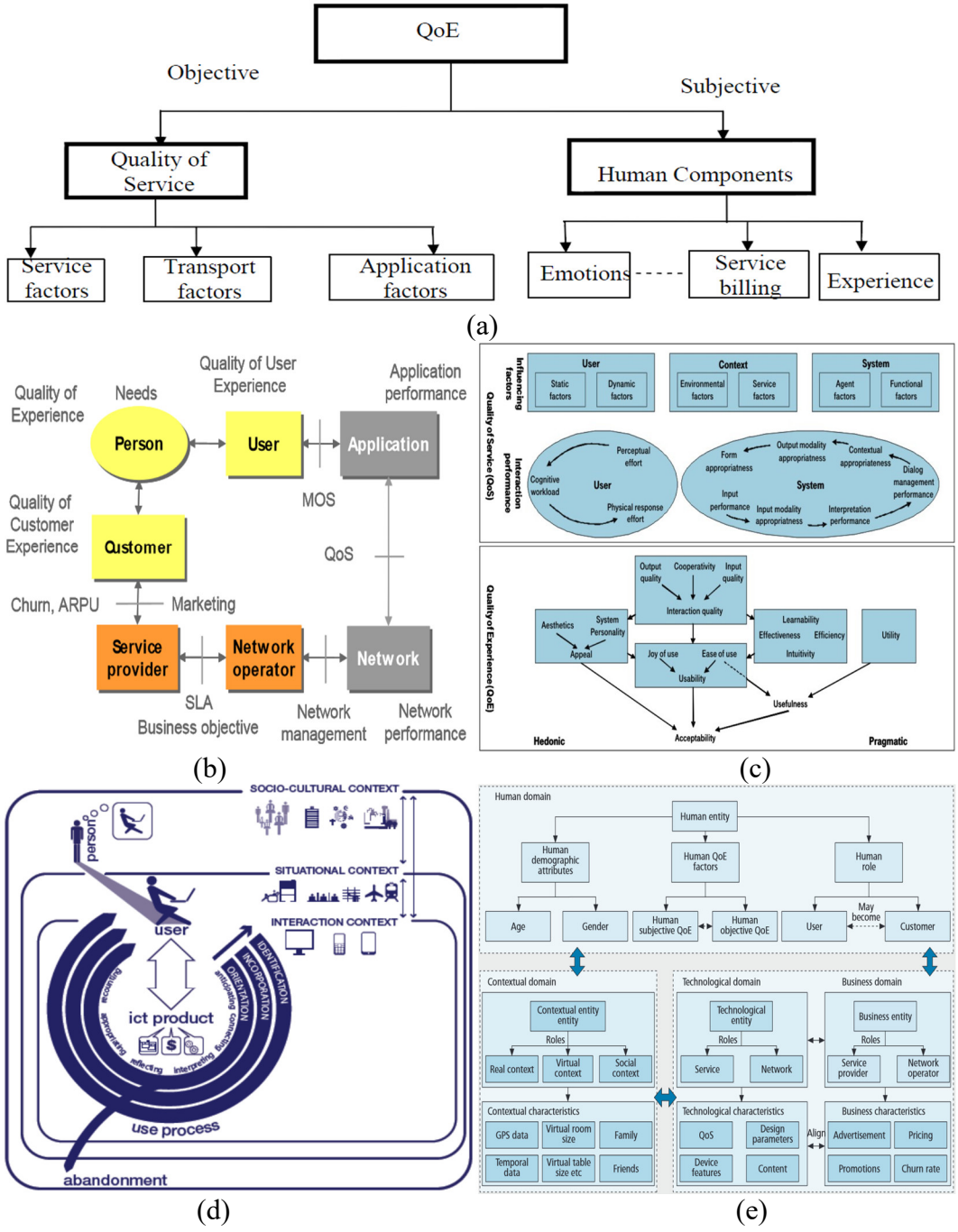


Figure 2.2 Prior attempts of QoE models. (a) The ITU-T G.1080 QoE model; (b) A QoE model in the communications ecosystem; (c) A QoE model based on the taxonomy of QoS and QoE of multimodal human-machine interaction; (d) A integrated architecture of measuring QoE; (e) A extended QoE model in the communication ecosystem by integrating HCI.

Kikki [49] proposed a common architecture for a QoE model in the context of communication ecosystem in 2008 as shown in Figure 2.2-(b). This model illustrated QoE as a description and management between the users and applications. In

addition, this model clearly and distinctly differed the role of customers from users. By introducing the customer module, the interaction between person, technology and business was integrated as an importance in QoE model. However, Kikki's model did not provide detailed subjective factors and objective factors for QoE measurement and the contextual information was also not considered.

Moller et al. developed a taxonomy from multimodal human-machine interaction to describe the relationship between QoE and QoS at three different levels in 2009 [50], which are QoS influencing factors including user, context and system; QoS interaction performance regarding user and system; QoE assessment related to perceived quality and acceptability, respectively as shown in Figure 2.2-(c). This model provides detailed metrics for QoE measurement and comparison between different human-machine interaction systems. However, only limited human computer interaction (HCI) was considered and the satisfaction of users perceived from the service was neglected in this model.

A QoE model including not only the technology but also business and context aspects of interaction, situation and socio-culture was proposed by David Geerts et al. in 2010 [51] as shown in Figure 2.2-(d). This model extended their prior work by integrating advanced research outcome from HCI, including, for example, users' expectations change over time, different contextual layers have different effects on users. But, the role of customer in term of the QoE quality was not described.

Laghari et al. presented a comprehensive extended QoE model on the basis of previous models in 2012 [52]. This model integrated many aspects of a communication ecosystem, including technology aspect, business model, human behaviour, and context as shown in Figure 2.2-(e). The subjective factors and objective factors were well classified in different domain as well as the mapping

interfaces cross different domain. But, this model is still a high-level QoE model. To use this model for emerging MAVS applications, an adaption is needed.

The Laghari QoE model is considered as efficient reference when studying the QoE for targeted MAVS application. The key influencing factors in different domain of Laghari QoE model were considered and be adopted in the subjective experiment in this thesis.

2.4.3 QoE measurement methods

The central question for QoE research is how to quantify the quality delivered to users and how the quality can be measured. This is a challenge question since QoE naturally encompasses the users' perspectives. A feasible mechanism is required to bridge or translate between traditional technology-centric QoS and user-centric QoE. Therefore, Quality assessment is needed not only from conventional end-to-end QoS parameters but also from key influencing factors reflecting users' requirement, perception, expectation and context. The QoE assessment methods can be broadly categorised into two groups: subjective quality assessment methods and objective quality assessment methods.

Subjective Quality Assessment Methods (SQAM) usually collects

Table 2-2 Rating scale of quantitative subjective quality assessment methods

ACR	DCR	CCR
5 Excellent 4 Good 3 Fair 2 Poor 1 Bad	5 Imperceptible 4 Perceptible but not annoying 3 Slightly annoying 2 Annoying 1 Very annoying	-3 Much worse -2 Worse -1 Slightly worse 0 The same 1 Slightly better 2 Better 3 Much better

information from test participants under specific experimental condition or stimulus by means of survey and user studies. In a subjective quality assessment, the participants generally are subjected to different levels of quality for test content (e.g. different delay, different image quality, different encoder parameters) which potentially or explicitly have direct or indirect effect on quality perceived by participants. To measure the users' response, quantitative or qualitative methods are mostly employed to get the users' perception and opinion. Quantitative methods gather users' perceived quality in the form of ratings. The ratings are numeric numbers, which can be used to perform non-parametric statistics to analyse the function of an influencing factor in term of the QoE. There are three well-known rating methods defined by recommendations like ITU-T P.800, ITU-T P.910, ITU-T P.913 and ITU-R BT.500-13, namely the Absolute Category Rating (ACR), Degradation Category Rating (DCR) and Comparison Category Rating (CCR) methods, respectively [53]–[56]. ACR is a kind of category judgement, where users encounter one stimulus and give a rating for that stimulus on a category scale each time. DCR is used to rate the impairment of the second stimulus with reference to the original stimulus. CCR is used to rate the impairment of two stimuli where these two stimuli are presented to the users in pairs. The rating scale of ACR and DCR is five-level rating scale while CCR is seven-level rating scale as shown in Table 2-2. The ACR scale, also known as the Mean Opinion Scores (MOS) scale, has become de-factor standard metric to capture the QoE in a majority of QoE researches. Alternatively, to complement the absolute scales used in MOS, relative Differential MOS (DMOS) or continuous rating scales are proposed and used in the literatures [57]–[59]. The MOS scale is employed in this thesis when conducting the subjective

QoE experiments. The equation to calculate the MOS value is shown in (2.2), which is the arithmetic mean of all individual ratings, and can range from 1 to 5.

$$MOS = \frac{1}{n}\sum_{i=1}^{n} a_i \tag{2.2}$$

Where a_i is the MOS rating from each experimental participant and n is the number of participants. In addition, because ITU-T P.10 defines QoE as overall acceptability [60], a binary scale (i.e. "1" and "0") is used to capture users opinions, for example, "accept or not, or like or dislike. Qualitative methods usually consist of observation and interview [61], [62]. These two kinds of qualitative techniques collect users' data through verbal communication between the experiment participants and researchers. The purposes of the observation and interview questions are explained to participants of subjective experiments beforehand to avoid the interruption to the participants during the test. The observations and answers of the interview questions are recorded by the researchers in their notes or electronically on a computer. The results of the qualitative methods will show the positive/negative responses from the users or portion/histogram of the users' opinions, which can be used to help researchers to identify the key influencing factors and how significant these factors are. Both quantitative MOS-based method and qualitative interview method are employed in the subjective experiment in this thesis.

Although SQAM is still the most sufficient, reliable and accurate way to measure the quality perceived by users and the only way to form the ground truths for quality assessment, due to its costly, time-consuming and complicated process, it is not suitable for in-service real-time quality assessment. Therefore, Objective Quality Assessment Methods (OQAM) are proposed for real-time quality assessment and monitoring by means of automatic algorithms, which process the input quality

related parameters to accurately predict user perceived quality. Nevertheless, only when the input parameters are closely related to the subjective quality, does this hold true. Therefore, the underlying requirement of OQAM is to discover the key influencing quality factors which can be quantifiable and then map the quantitative impact factors to ground truth MOS values which are derived from SQAM using optimum fitting technology. It is noted that OQAM is service dependent as different services have different quality influencing factors. For example, Peak Signal to Noise Ratio (PSNR), Structural SIMilarity (SSIM), Video Quality Metric (VQM) are widely used in the quality assessment of picture and video; the planning parameters are used in the ITU-T G.107 E-Model for transmission planning; perceptual Evaluation of Speech Quality (PESQ) is a commonly used intrusive mode to evaluate the quality of speech; waiting time is a significant parameter to monitor the quality in web browsing. Ultimately, these methods concentrate on the mapping approaches from QoS-related factors to QoE-related factors. The results derived from OQAM and SQAM can be mapped through curve fitting approaches by finding the maximum correlation function [63], [64]. However, despite its fast, easy and cost-effect advantages of OQAM, the objective metrics used in the OQAM can only capture partial aspects of users' perceived quality. The OQAM may provide inaccurate or dubious results when new conditions are encountered in the service. In that case, revalidation, renovation, redevelopment are required. Another kind of OQAM is based on the psychological signals combined with subjective measures can be used to assess the QoE, such as Magnetic Resonance Imaging, Galvanic Skin Response, EEG signal, and ECG signal. The advantages of such methods are that they collect the user's sensory information which can imply the opinion and perception of user. The disadvantage is that additional equipment needed in this kind

of measurement and the results are highly affected by the context. Such OQAM based on psychological model is out of the scope of this thesis.

2.4.4 QoE challenges in mobile devices for MAVS

MAVS systems blend virtual or related rich multimedia content with reality, interact with users in real time, and communicate with multimedia content repository on the air. Due to these unique properties of MAVS systems, there are challenges to overcome to deploy the MAVS systems in mobile devices:

1) MAVS systems commonly need to find the match from the database to the captured image. When capturing the image, there are two main variables that lead to poor quality images from mobile devices camera: the usage environment and the user. In terms of the usage environment, also known as photometric distortions, although current mobile devices equip sophisticated high resolution camera, they are still limited in their ability to capture photos under poor lighting condition. For example, the quality of the captured image is lost because of darkness and blurry during the shoot. It raises difficulty for MAVS algorithms to accurately process the image to extract useful information. Unfortunately, it is normally not available to access the low-level camera sensor to compensate the image quality during the image acquisition on mobile devices. Only limited compensation can be done in the high level application layer. Another key issue is the variability due to the users. Another factor is geometric distortions (e.g. rotation, out-of-plain) due to the amateur or arbitrary shoot normally occur which makes it more challenge to accurate extract useful information leading to correct image matching;

2) Mobile devices are promising platforms to deploy the MAVS systems as their rich software and hardware resources. However, the mobile devices are not designed specifically for MAVS systems. For example, the image codec in the

mobile phone is designed for image display, storage and transmission rather than image matching. The distortion occurred by image codec will influence the accuracy for image matching;

3) The MAVS systems have to process a large amount of data captured by camera (e.g. streaming frames captured by camera). However, not all the captured information is useful for processing. Advanced image processing and computer vision technologies are required to cope with these data to efficiently filter the capture information, find out useful information and discard useless information. This procedure requires huge computation;

4) Although smart devices have increasing computational power with the advance of high performance chips and high speed memory, the computation to process data and the transmission of data in real time is still challenge. Furthermore, the battery life and wireless network is still limited. Minimizing the data to be processed and transmission is definitely desirable;

5) The MAVS systems often have to search over a large database to target the predefined augmented multimedia content, for example, one hundred images in a dataset can contain more than tens of thousands of features for search. How to perform fast search in this kind of large scale dataset to get accurate result is challenge;

6) As a new emerging multimedia application in the mobile devices, the interaction with users has been changed from click-based-action to capture-based-action. This interface changing may significantly influence the users' perceived QoE. As well as considering the MAVS system usage context (e.g. indoor or outdoor), it raises a question whether previous QoE study result can be applicable to manage the QoE in the MAVS systems. This is also a challenge and need to be investigated.

Based on above challenges, the main research question of this thesis is "**How can QoE be ensured in an MAVS system which links print media with augmented multimedia content?"** This main research question is broken down to the following small questions:

1. What is the most efficient feature to extract the useful information from a user captured scene, which the computer can process easily and accurately?

2. What is optimal way to deal with distortion like optical distortion, geometric distortion and codec distortion to make sure the accuracy of MAVS systems?

3. Is there an efficient way to select the most significant information to minimize the computation and transmission of the MAVS systems to ensure real time performance?

4. What is the relationship between objective and subjective quality measures in MAVS applications and how can this relationship be used for maximizing the user QoE?

To answer these questions, studies are conducted in the later chapters focused on two key influencing factors, that are waiting time and accuracy.

2.5 Waiting time

Since the majority of modern multimedia applications employ request-response models to interact with users as well as MAVS applications, a waiting time refers to a time period starting from when a user's action until a user perceives a response. For example, a waiting time is from a user clicks a web link to the corresponding web page rendering in a web browsing service or a user requests a video playback until the video starts to play in a video streaming service. In an MAVS application, the waiting time can be defined from when a user triggers the camera capture to when

the user perceives the augmented multimedia content. The waiting time plays a vital role in the user perceived quality. As a direct quality perception stimulus, waiting time itself is a significant subjective quality factor. It is intuitive that users do not want to experience unnecessary delay as too long a delay will cause users' doubt whether the application works or not and results in user churn. As stated in [65], more than 8s page download time is unacceptable and results in user quit. Too long a waiting time in the context of a web shopping causes users' suspicion about the system and the safety of the online payment [66]. 0.1s is the limit to make user feel that the system responses instantaneously while 1s delay does not interrupt the users in an interactive system [67]. Memory effects will influence the users' satisfaction of perceived waiting time, a user experienced fast response in the past had less tolerance for long delay [68]. Thus, users may have different expectations about the waiting time in different services, that is the tolerance threshold of waiting time may vary [69]. In this section, the waiting time in QoE for web browsing and video streaming are reviewed.

2.5.1 Waiting time for web QoE

End user perceived waiting time is the key of QoE in web browsing [70], [71]. A web page is a text document with links to other multimedia content, such as images, videos, cascading style sheets, scripts, etc. Underlying users click a link to access a new page and receive new data, a HTTP request to load corresponding information is issued, the waiting time is the duration until the new page view renders in the browser. The waiting time directly influences the users' browsing experience and is known to depend on many factors both including subjective and objective factors, especially if it is closely related to QoS parameters of the network, such as large

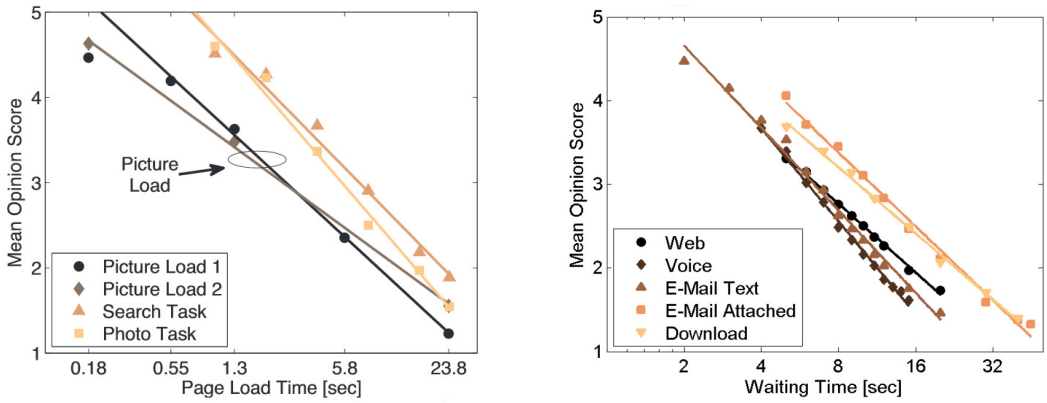


Figure 2.3 QoE study in different web services. a) User satisfaction for various constant page load time; b) Logarithmic relationship between MOS and waiting time for several services. The data of these figures are reported in [72]–[74].

packet delay and low bandwidth. Waiting time as a metric, which is also known as page load time (PLT) in the web service, is highly correlated with QoE as indicated

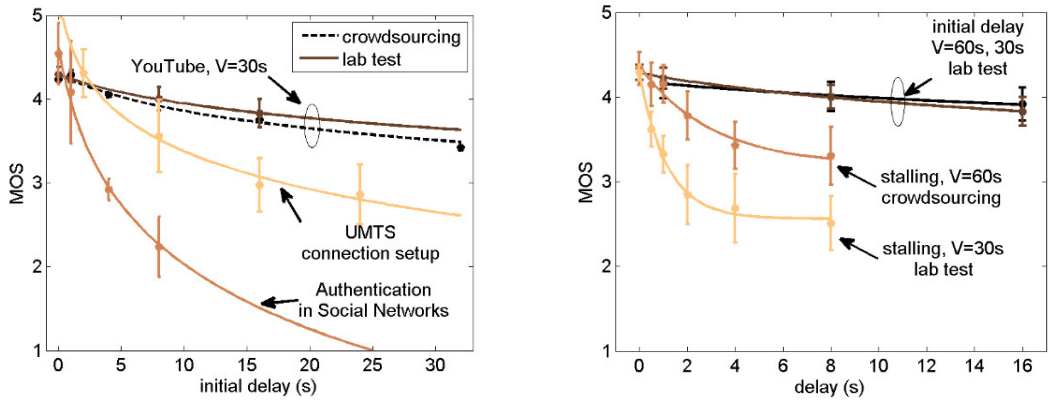
in [72] and is sufficient for predicting Web QoE. The study in [73]–[75] suggest that the relationship between waiting time and users' perceived QoE is logarithmic as shown in Figure 2.3. The discoveries in previous studies comply with the Weber-Fechner Law (WFL) in psychophysics [76], which states that human perception will diminish with the increase of the magnitude of a stimulus. A general equation describing the relationship between human perception and stimulus (e.g. waiting time) is below [76]:

$$P = k \cdot ln\frac{S}{S_0} \tag{2.3}$$

Where P stands for the magnitude of perception, S is current magnitude of stimulus, S_0 is a threshold of S (the minimum magnitude of a stimulus can be sensed), k is a constant depend on usage in different context.

2.5.2 Waiting time for video streaming

Besides surfing the web, video streaming is ubiquitous in current internet activities. The main difference of video streaming is that the video stream is compressed using a codec, such as H.264 and the encoded video is assembled into a container bitstream, such as MP4 and then the bitstream is transmitted to users using a streaming protocol, for example HTTP/TCP/RTSP. Additional waiting time for decoding the bitstream and buffering the media playout is required. Moreover, video streaming normally keeps active for a certain duration. During that period, re-buffering may occur and cause users to waiting for a while. These two kinds of waiting time are known as initial delay and video stalling, respectively. The study in [75], [77] suggests that the relationship between QoE and initial delay is still fitted with a logarithmic function but the influence of initial delay on QoE across different services is strongly diverse as shown in Figure 2.4-(a) (i.e. different logarithmic functions). The stalling is more annoying than initial delay even if the duration of waiting time is the same and the relationship between stalling and QoE is better fitted with an exponential function as shown in Figure 2.4-(b) because the stalling is more visible and noticeable to users and the memory effect have strong effects on users' perception.



(a) (b)

Figure 2.4 Influence of waiting time on video streaming. a) initial delay vs. MOS in different services b) comparison of the MOS between initial delay and video stalling.

According to previous researches, it is difficult to manage the waiting time in terms of the QoE. The human perception about time is not precisely corresponding to objective time. Users' feeling about the time is influenced by many factors, such as

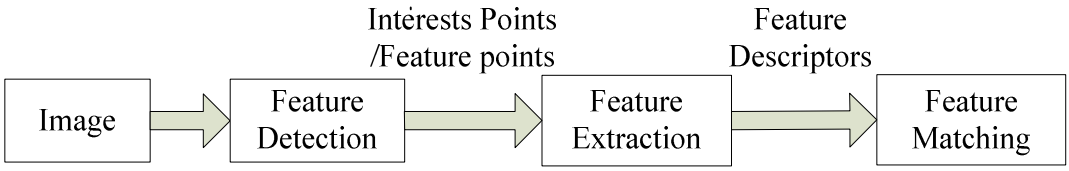


Figure 2.5 The diagram of feature detection and feature extraction when performing image matching.

the variety of application usage, personal factors (e.g. background, knowledge about the technology underlying the service), different interaction interface (e.g. mouse click vs. camera capture) and different expectations of desired response (i.e. users may expect different waiting time in different applications). As MAVS applications normally integrate with the web service and video streaming, it is questionable whether the logarithmic relationship is still applicable to such new emerging application? In additional, the ultimate goal of MAVS applications is to provide augmented information to users, the usage context is different from traditional multimedia service. The experience on mobile devices is obviously different from desktop computers. These factors may have different effects on QoE, which is worth to study. To answer these questions, a subjective experiment on a mobile device to study the waiting time on an MAVS application is conducted in this thesis.

2.6 Image feature detection and extraction

The focused MAVS applications in this thesis link print media with augmented multimedia content, such as web or video. One essential procedure is to perform image matching between two images which depict the same object. The diagram of image matching is shown in Figure 2.5. The two important procedures to generate a feature are feature detection and feature extraction. The feature detection finds the interests points, also known as feature points, in an image. Each feature point has the

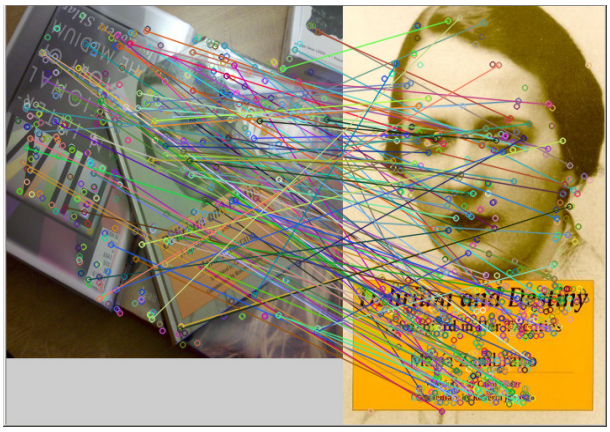


Figure 2.6 An example of matching image captured by camera to a clean image under complicated distortion.

location of that point in the image as well as some characteristic values depending on the feature detector (e.g. orientation, scale or response). Then, these feature points are used by the feature extractor to extract distinctive spatial information around each feature point to form feature descriptors. Each descriptor describes the information of the image patch around a feature point. However, the object in the image captured by a camera commonly appears in different position, scaling and orientation compared to image in the server as well as different image quality due to lighting change or blurring. These geometric and photometric distortions raise difficult for image matching in MAVS applications as shown in Figure 2.6 . In this section, the methods to detect and extract the most discriminative features for image matching are reviewed.

2.6.1 Feature detection

The purpose of feature detection is to find the image regions (i.e. a set of pixels) which are covariant with a class of transforms (e.g. viewpoint change). The requirement for these detected regions is that they have the ability to automatically adapt to the transformation and correspond to the same object in the image before

transform, which means a certain 3D projection function can be found from these regions between images from different viewpoints [78]. In addition, the detected features should be repeatable in the presence of varying image noise. Literatures in computer vision about affine region detectors have reached maturity [78]–[97].

Harris and Stephens proposed a combined corner and edge detector for feature tracking by means of auto-correlation function in a local region, which had good consistency on natural imagery [85]. Similar work had been done by Shi and Tomasi to propose a feature selection criterion to generate Good Features To Track (GFTT) [86]. Kadir, Zisserman and Brady designed an affine invariant salient region detector not only concentrating on viewpoint invariance but also considering insensitivity to image noise and repeatability under transformation [91]. These feature detectors mainly focused on viewpoint invariance but were short of scale invariance properties (one obvious example of scaling change is zooming). To solve the scale variance problem and detection efficiency, Lindeberg embedded an automatic scale selection algorithm into feature detection based on maxima over scales of normalized derivatives [87]. Lowe presented his initial work of Scale Invariant Feature Transform (SIFT) in 1994 for object recognition under varying image scaling, translation, rotation, illumination changes and 3D projection [88]. The SIFT detector uses Difference-of-Gaussian (DoG) to approximately detect the maxima and minima in the Laplacian-of-Gaussian (LoG) scale space [80]. Mikolajczyk and Schmid extended the work of Harris detector by adding affine transformations to the interest points and selecting the interest points which are local extremum in the laplacian scale space [89]. Maximally Stable Extremal Regions (MSER) was proposed by Matas et al. which discovered the extremes by

examination of an intensity function with varying intensity threshold in an image [81].

To achieve fast detection, Nister and Stewenius invented a linear Time MSER based on different computational ordering of the pixels to speed up the original MSERs detection [93]. Bay et al. proposed Speed-Up Robust Features (SURF) as a faster alternative to SIFT feature by using a box filter and integral image. Features from accelerated segmented test (FAST) corner detector is designed by Rosten et al. specifically for feature detection in real time feature tracking in video application [94], [97]. To solve rotation variation, ORB (Oriented FAST and Rotated BRIEF) feature adds a fast and accurate orientation component (i.e. intensity centroid moment) to FAST [82]. To achieve scale invariance, Binary Robust Invariant Scalable Keypoints (BRISK) detect FAST keypoints in scale space. Rotation-

Table 2-3 Summary of Important Feature detectors listed in the chronological order

Feature Detector	Year	Rotation Invariant	Scale Invariant
Harris detector [71]	1988	No	No
GFTT detector [72]	1994	Yes	No
affine invariant salient region detector [77]	2004	Yes	No
SIFT [66]	2004	Yes	Yes
MSERs [67]	2004	Yes	Yes
Harris-Laplace [75]	2004	Yes	Yes
FAST	2006	No	No
linear Time MSERs [79]	2008	Yes	Yes
SURF [69]	2008	Yes	Yes
ORB [68]	2011	Yes	No
BRISK [82]	2011	No	Yes
RIFF [81]	2013	Yes	Yes

invariant fast features (RIFF) achieves both rotation and scale invariant and fast speed based on an approximation of the LoG scale-space using differences between box filter responses [95].

Comparative studies of the aforementioned feature detectors were conducted in [78], [79] from the aspect of feature retrieval and precision under signal geometric and photometric distortion. From the evaluation results, different feature detectors have pros and cons. To extent previous work in this thesis, different detectors are evaluated with different descriptors in the context of MAVS application under combined geometric and photometric distortion.

The summary of the aforementioned feature detectors are shown in Table 2-3 in the chronological order of their first proposal.

2.6.2 Feature extraction

After detecting the feature keypoints, the features can be extracted from neighbouring image local patches around the keypoints and generate feature descriptors (e.g. vectors) which describe the structural information around the keypoints. The feature information encapsulated in the descriptor should be invariant to image noise such as geometric and photometric distortions and in the meantime are discriminative for performing distance-based descriptor matching to distinguish different image patches. An example of feature extraction and feature descriptor generation is shown in Figure 2.7 , where the descriptor describes structural information based on the histogram of gradient in a 4*4 spatial grid within a local image patch. Several famous feature descriptors are listed in Table 2-4.

SIFT descriptor is the most well-known and popular descriptor due to its' excellent rotation and scale invariance and robustness to image noise. SURF descriptor employs the similar concept to generate the descriptor which achieves

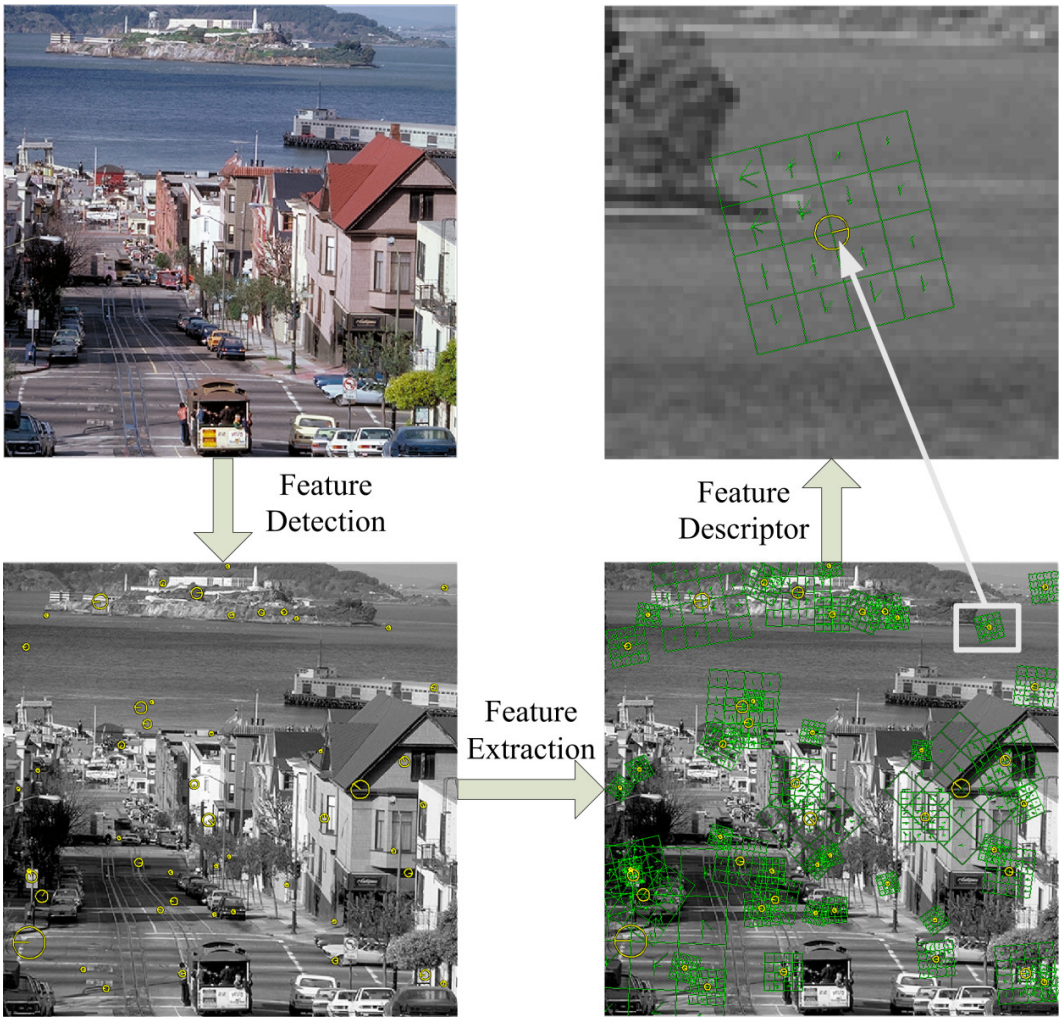


Figure 2.7 An example of feature detection and extraction in a spatial grid within an image patch

faster speed with comparative performance to SIFT. The common properties to construct these two kinds of descriptors are [80], [92]: 1) the orientation information is embedded in the descriptor by choosing the peak value in a local orientation histogram; 2) intensity normalization is employed to protect against brightness and contrast changes; 3) A Gaussian kernel weighted histogram of the gradients are used as components to construct the descriptor in a spatial space of 4*4 square grid within a local image patch. The gradients refer to the distribution of intensity gradients. These properties increase the descriptiveness of the SIFT and SURF descriptor. On

the basis of a performance evaluation of local descriptors, Mikolajczyk and Schmid concluded that gradient-based descriptors achieved the best performance in term of the recall-precision in the feature domain for image pairs [98](i.e. the retrieval rate of true positive gradient-based features is high which can results in a successful matching for an image pair). They also extended the SIFT descriptor and proposed a new descriptor named Gradient Location and Orientation Histogram (GLOH) by substituting square grid with polar grid.

The SIFT descriptor is high dimensional vector. To achieve compactness, Ke and Sukthankar employed Principal Components Analysis (PCA) to the normalized gradient patches around SIFT keypoints to create the PCA-SIFT

Table 2-4 Summary of several different type feature descriptor

Feature Descriptor	Data type	Characteristic
SIFT [66]	Float	Gradient
PCA-SIFT [84]	Float	Gradient
GLOH [85]	Float	Gradient
Compressed histogram of gradients (CHoG) [86]	Float	Gradient
SURF [78]	Float	Gradient
Dual-tree complex wavelet transform (DTCWT) [87]	Float	DWT
Local Polar DCT Feature (LPDF) [88]	Float	DCT
BRIEF [89]	Binary	Pixel intensity
ORB [68]	Binary	Pixel intensity
BRISK [82]	Binary	Pixel intensity
FREAK [90]	Binary	Pixel intensity
RIFF [81]	float	Gradient

descriptor while maintaining robustness against image deformation [99]. Aiming to implement both discriminative and compact features, Chandrasekhar et al. proposed the Compressed Histogram of Gradients (CHoG) descriptor by compressing the gradient distribution in each spatial cell and substituting the Euclidean distance with the Kullback-Leibler divergence for the distance measurement between features [100]. Takacs et al. proposed improved RIFF descriptor in [95] which also uses a histogram of gradients, but computes radial and tangential gradients in the polar coordinate system.

There are also the other feature descriptors which are not based on intensity gradients. Selesnick et al. designed a scale and rotation invariant descriptor by using the coefficients derived from the dual-tree complex wavelet transform (DTCWT) [101]. Song and Li presented a compact yet robust descriptor based on rearranging 2D-Discrete Cosine Transform (DCT) coefficients on a quantized local image patch [102]. Calonder et al. proposed an efficient feature descriptor called binary robust independent elementary features (BRIEF) from the point view of fast descriptor generation and feature matching by using simple intensity difference tests [103] to approximate the gradient calculation. By adding rotation component to BRIEF, Rublee et al. proposed ORB descriptor which improved the robustness against in-plane rotation [82]. Combining orientation normalization and simple brightness comparison tests, Leutenegger et al. designed the Binary Robust Invariant Scalable Keypoints (BRISK) descriptor[96]. Inspired by the human visual system, Alahi et al. employed retinal sampling pattern to extract Fast Retinal Keypoint descriptor by using intensity differences [104].

In principal, different applications have either strict demands in computation speed-up or precision. Binary feature descriptors are found to be

faster than gradients-based feature descriptor but with lower pairwise image matching accuracy as well as lower retrieval accuracy over a large database [105]. In this thesis, different types of descriptors are evaluated from the point view of fulfilling the key requirements of speed and matching accuracy for MAVS applications to maximize users' perceived QoE.

2.7 Pair-wise image matching

After feature detection and extraction, the distinctive information embedded in the feature descriptors can be used in an image pair to perform feature matching to decide if these two images contain some common objects or not. This procedure is known as pair-wise image matching. The pair-wise image matching should be robust against missing features and noisy features as well as be able to distinguish outlier features (e.g. exotic features from irrelevant objects or false features caused by image noise).

As shown in Figure 2.6 , given the features detected in the left image, the problem is to find the feature correspondences in the right image. Two sets of descriptors for the left image and right image in Figure 2.6 can be denoted as $\mathbf{F_1} = \{\mathbf{f_{1,1}}, \mathbf{f_{1,2}}, \cdots, \mathbf{f_{1,m}}\}$ and $\mathbf{F_2} = \{\mathbf{f_{2,1}}, \mathbf{f_{2,2}}, \cdots, \mathbf{f_{2,n}}\}$, respectively. For each descriptor $\mathbf{f_{1,i}} \in \mathbf{F_1}$, i=1 to m, it is desirable to find a corresponding descriptor $\mathbf{f_{2,i}} \in \mathbf{F_2}$, for which the distance between these two descriptor is the minimum distance compared to other descriptors in $\mathbf{F_2}$ as shown in (2.4):

$$\mathbf{f_{2,i}} = argmin_{\mathbf{j}} d(\mathbf{f_{1,i}}, \mathbf{f_{2,j}}) \tag{2.4}$$

where $d(\mathbf{f_{1,i}}, \mathbf{f_{2,j}})$ is a distance function measuring the distance between two descriptors in $\mathbf{F_1}$ and $\mathbf{F_2}$, respectively. There are different distance functions and distance measurement available [106], [107]. One of the most popular distance

measurements is the Euclidean distance. Although Euclidean distance is used in this section to discuss different matching methods, most methods are applicable to other positive distance metrics, for example, the hamming distance.

2.7.1 Threshold based matching method

Considering the Euclidean distance as a metric to measure the distance between two feature descriptors, threshold based matching distance is the simplest method to decide if these two features are matches or not [106]–[108]. By setting a threshold ε_T (i.e. the maximum distance of matches), the matches beyond the threshold are filtered out and the remaining features within the threshold are passed to the next step for further processing. It is hard to accurately set the threshold as setting too high threshold can lead to high false positives (i.e. many incorrect features are recognised as correctly matched features) while setting too low a threshold may result in high false negatives (i.e. many correct features are filtered out). High false positives are not efficient for next step processing as a descriptor can have too many matches while high false negatives can lose significant features and both can result in bad matching accuracy. Although the threshold can be learnt from the standard deviation of correctly matched features or a certain adaptive technology can be applied to generate a soft threshold [109]–[111], the threshold may be inappropriate when image noise occurs and varies in different feature space (i.e. different feature type and different image dataset require different threshold) as evaluated in [78], [98].

2.7.2 Ratio test matching method

As mentioned in the previous section, the problem of using a hard/fixed threshold is that it is difficult to set a proper value. To improve the feature matching accuracy, a Nearest Neighbour (NN) search which solves the objective function as shown in (2.4)

was proposed and studied in many literatures [112]–[117]. A simplest example of the NN search algorithm is that if descriptor B is the nearest neighbour of descriptor A (i.e. the distance between B and A is smallest compared to other descriptors) and while the distance between them is below a threshold δ, descriptor A and B can be considered as a match. In this case, for each descriptor, only one match can be found for each descriptor. However, as stated in [80], a global distance threshold did not perform well, a more effective measure based on ratio test method, known as Nearest Neighbour Distance Ratio (NNDR) is employed in [80], [118] to find more discriminative descriptors. NNDR compares the distance of the nearest neighbour to the second nearest neighbour as shown in (2.5).

$$\text{NNDR} = \frac{d_1(f_1,\ f_2)}{d_2(f_1,\ f_3)} = \frac{||f_1 - f_2||}{||f_1 - f_3||} < \varepsilon_T \tag{2.5}$$

where $d_1(f_1,\ f_2)$ is the distance between descriptor $f_1$and it's nearest neighbour f_2; $d_2(f_1,\ f_3)$ is the distance between descriptor $f_1$and it's second nearest neighbour f_3; ε_T is a threshold for the distance ratio. The performance of the threshold-based and ratio-test based matching method is evaluated in [98] which suggested that the ratio-test based matching method provided more reliable result.

2.7.3 Cross-check matching method

Although the NNDR algorithm can achieve high feature matching accuracy, one drawback of the method is that a threshold is still required. An alternative method named cross-check matching method was implemented in [119]. A toy example of cross-check matching method is illustrated in Figure 2.8. There are two descriptor feature sets A and B, each has 3 descriptors, A = $\{f_1, f_2, f_3\}$, B = $\{f_4, f_5, f_6\}$. The nearest neighbour of $\{f_1, f_2, f_3\}$ in B is $\{f_4, f_4, f_5\}$, respectively while the nearest neighbour of $\{f_4, f_5, f_6\}$ in A is $\{f_3, f_3, f_3\}$, respectively. f_3 and f_5 are both the nearest

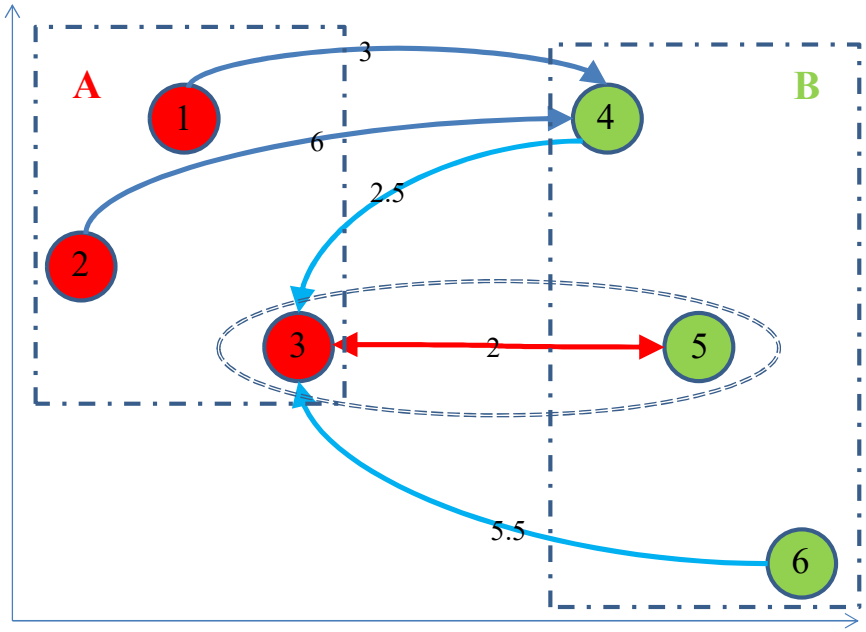


Figure 2.8 A toy example of cross-check matching method

neighbour for each other. Therefore, only the matches of f_3 and f_5 are taken as the true positive matches. Because of the usage of the NN criteria when measuring the distance between feature matching pairs, a fixed threshold is not required. Suggested by [120], the cross-check method provides better feature matching accuracy but demands double the level of processing for matching.

2.7.4 Efficient and fast matching method

Due to image geometric and photometric distortion, feature descriptors are often corrupted by noise which increases the difficulty of correctly finding the descriptor matches between two images. Additionally, it is not always guaranteed that feature correspondences can be found. Therefore, using exhaustive NN search is not efficient and is computationally and time-consuming. To tackle this problem, an efficient and

fast matching method is desirable which can optimally balance between computation and feature matching accuracy.

One efficient and fast solution is Approximate Nearest Neighbour (ANN) searching with indexing structure such as multidimensional searching tree structure or hashing table to rapidly find correspondences for given feature descriptors. The indexing structure can not only be used in a descriptor set extracted from an individual image (i.e. speed up feature matching and focus on the common objects in an image pair) but also is applicable for building a search structure of all the images in a dataset (i.e. fast search within a database).

By using ANN with K-Dimensional (K-D) tree, Arya et al. reduced the time complexity from linear time $O(n)$ to loglinear time $O(n \log n)$ to search in a partitioned K dimensional vector space (i.e. a binary tree structure in which each node is a k dimensional vector) [121], [122]. Lowe proposed a speed-up ANN feature search method called the Best Bin First (BFF) for SIFT descriptors [80]. Muja and Lowe subsequently released the fast library for ANN (FLANN) search which uses multiple randomized K-D trees for high-dimensional features such as SIFT, SURF and multiple hierarchical clustering trees for binary features such as ORB, BRISK [123]–[125]. In this thesis, FLANN is employed to construct a fast and accurate feature matching structure for targeted MAVS applications to maximise the QoE.

2.7.5 Geometric verification

After the feature matching, a Geometric Verification normally is followed. The principal of GV is that an object in an image generally has certain shape and the features extracted from the object should comply with a valid geometric model.

There are two aims to perform GV: 1) aiming to further filter out the outlier feature matches (i.e. incorrectly matched feature pairs); 2) estimating the geometric model of the inlier features. The features which do not fit the estimated geometric model are normally considered as outlier features.

There are two popular approaches for geometric estimation known as RANdom SAmple Consensus (RANSAC) [126] and Least MEDian of Squares (LMEDS) [127], [128]. These two methods start from randomly selecting a subset of observed data samples and then iteratively estimating the residual error between the selected subset data samples and entire dataset to find the data subset which has the minimum estimation error or below a certain predefined estimation error threshold. The geometric verification is a crucial process in image matching for filtering out the false features to improve the image matching accuracy. Variants of improved geometric verification approaches can be found in the literatures [129]–[134].

2.8 Feature selection

In a typical MAVS application, when using a mobile camera to scan a scene, the captured image normally is a rich content image as shown in Figure 2.6 and Figure 2.7 . Thus, hundreds of local features commonly can be detected and extracted from the captured image. As stated in section 2.7, to perform the matching to all the detected feature is not optimal as the computation is increased and the matching accuracy will be deteriorated because of false features extracted from irrelevant objects or background and generated by image noise. The features that do not contribute to a correct match are desirable to be filtered out before performing feature matching; otherwise such features will have a negative effect both on matching accuracy and processing delay. Hence, there are several benefits to

implementing feature selection in an MAVS system: 1) reducing the transmission bandwidth required for the features; 2) minimising memory resources required for feature storage; and 3) speeding up the feature matching system. As feature selection is desired to select the most essential and significant features before matching while keeping matching accuracy as high as possible, the criterion of feature selection is crucial and inappropriate feature selection would degrade the matching accuracy dramatically. Assuming a proper function can be found to evaluate the effectiveness of given feature dataset, feature selection can also be considered as a feature search problem, which searches for an optimal subset based on the selecting criterion.

Feature selection has gained attention for decades [135]–[148] and is an important component in an on-going MPEG-7 standardization known as CDVS [149]–[154] (the detailed information of CDVS is introduced in section 2.12), in which feature selection is investigated as a key technology for compact descriptor extraction aiming for low bit rate transmission and high matching accuracy. Broadly, the state-of-art local feature selection methods can be categorized into three groups: 1) threshold-based feature selection; 2) geometry-based feature selection; 3) relevance-based feature selection.

2.8.1 Threshold based feature selection

Threshold-based feature selection is normally implemented within the feature detection. After detecting a set of interest points in an image by a feature detector algorithm, a threshold is employed to determine whether an interest point is a keypoint or not. For example: the SIFT detector uses a contrast threshold to eliminate interest points extracted from low contrast regions and an edge threshold is employed to find keypoints extracted from an edge [80]; the SURF detector uses a hessian

threshold to filter out the interest points with a low hessian value [83]; the ORB detector uses a Harris corner response as a threshold to filter the features detected from a flat region [82]; and MSER uses an intensity threshold to select the features from the maximum or minimum intensity region. The threshold-based method is fast and efficient to filter weak interest points. However, this method is image dataset-dependent and easily influenced by image distortions [80], [82], [83].

2.8.2 Geometric information based selection

The principal behind geometry-based feature selection is that a rigid object in a natural scene image normally exhibits a certain shape with a closed contour resulting in a coherent spatial pattern. Keypoints tend to cluster on the basis of this geometric information and topological structure. This clustering characteristic of keypoints has been shown to be an effective constraint for a keypoint filter in [155]–[160]. By doing so, image spatial self-matching solutions have been proposed in [161], [162]. Such methods use four degrees of freedom transformation or six degrees of freedom transformation to generate an affine transformed image, for example, a flipped image or out-of-plane rotation image and then perform pairwise image matching. The matched local features are selected as useful features and can enhance the matching accuracy of image pairwise matching. But, such methods require a doubling of the feature detection, feature matching and geometric verification processing steps as well as additional image manipulation on the client side of the practical MAVS system, which consumes more computational resource and battery power resulting in longer processing delay in the client side. Additionally, it is also difficult to accurately determine the thresholds used in the feature matching and geometric verification stages used for feature selection for a wide variety of images. Hence, an

alternative approach that avoids this doubling of the process and additional image manipulation is desirable.

2.8.3 Feature relevance based wrapper selection

Relevance-based feature selection is developed on the basis of wrapper methods for feature selection [163]–[166] and takes advantage of the power of the image matching system. The wrapper methods of feature selection are unsupervised feature selection algorithms and have been widely used in machine learning for classification problems. Such methods wrap feature selection with classification algorithms that will ultimately be applied, and directly use the maximum likelihood output results of classification algorithms to select the useful subset features. Similarly, the relevance-based feature selection method selects the feature based on the results of a pairwise image matching system. But, instead of using the correctly matched features from a matching system directly, a posterior probability of a relevance parameter associated with correctly matched features is learnt off-line from a training dataset which contains both distorted images and corresponding clean images [167], [168]. This posterior probability of a relevant parameter associated with correctly matched features is known as the relevance. The MPEG-7 CDVS has adopted this method for SIFT feature selection [149]–[154]. The relevance of the output parameters of the SIFT detector, including the Difference-of-Gaussian (DOG) response θ_{peak} (denoted as peak in the following paragraphs), scale θ_{scale}, orientation $\theta_{orientation}$, location $\theta_{distance}$ (the distance from the keypoint to the image center), are used for SIFT feature selection. Parameter θ_{peak} is superior for identifying the most significant features for pairwise feature matching compared to other parameters as evaluated in [167], [168].

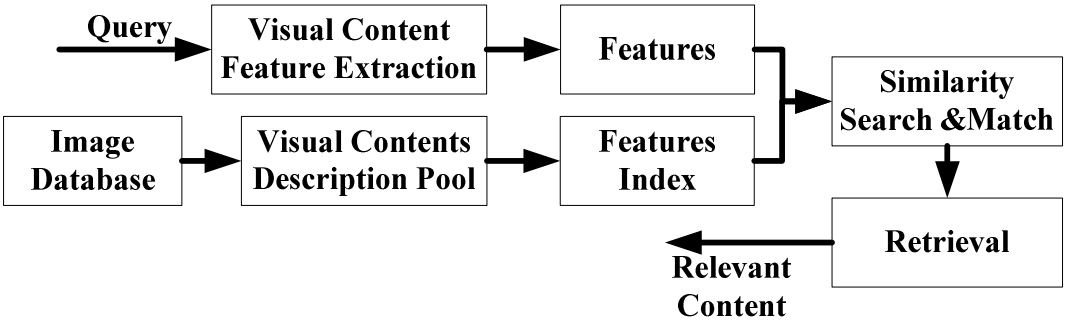


Figure 2.9 A diagram of CBIR system

Relevance-based feature selection can lead to better image pairwise matching accuracy compared to threshold-based feature selection and is faster than geometry-based feature selection as the relevance learning process is offline and only local feature reordering is required based on the learnt relevance. Hence, a relevance-based feature selection method to fast and accurately search the most significant features to maximize the QoE perceived by user is studied in this thesis.

2.9 Fast and accurate Content Based Image Retrieval (CBIR)

The MAVS applications targeted in this thesis typically operate with small-scale databases often containing no more than a few hundred images (e.g. images in a newspaper or magazine) [169], [170], [171], where users scan across images with their mobile device camera to find the related image on pre-select/pre-defined dataset and then trigger the linked augmented content to return to users as shown in Figure 2.9. As mentioned in the previous section, the waiting time caused by system delay should be as fast as possible to provide real time performance to users meanwhile keeping the matching accuracy high. Hence the key challenge is how to choose an efficient method to generate compact visual information for processing and transmission in mobile devices whilst achieving real-time and highly accurate image retrieval against distortion. This has previously been shown to be essential to

maximizing user Quality of Experience (QoE) for such applications [169], [170], [172], [173].

To achieve low computation and fast retrieval, many approaches have been proposed for compressed domain retrieval during the past decades using DCT and wavelet coefficients. Such methods were applicable for duplicated image retrieval. However, they have not achieved high matching accuracy under optical and geometric distortion [174]–[177]. To improve retrieval rate and matching accuracy, low level feature extraction based on DCT and wavelet coefficients in spatial domain were proposed. However, the matching accuracy of such methods still degraded under joint optical distortion and geometric distortion [102], [178]. Alternatively, as reviewed in section 2.6 and 2.7, many studies have focused on developing repeatable, distinctive, and discriminative high quality local feature to achieve highly accurate matching rates for image matching and image retrieval [78], [79], [98]. The most popular and proven high discriminative features are histogram-based features, such as SIFT [80]. Although SIFT achieved excellent matching accuracy, the main drawback of using SIFT on mobile devices is high dimensionality resulting in complex computations for feature matching and a significant amount of bandwidth for feature transmission (often more than the compressed JPEG format of the image [172]). To reduce the dimensionality and alleviate the transmission issue, the SIFT compression schemes including hashing, transform coding, vector quantization have attracted much attention [98], [179], [180]. Such methods normally compromised between bit-rate and matching accuracy such as CHOG which achieved low bit rate transmission by directly compressing the gradient histogram while maintaining high matching accuracy [100]. However, the extra computational complexity and power consumption are added to feature detection and extraction. If

such processing was performed on mobile devices, it will cause significant delay and result in decreased QoE for the users [172], [173]. Especially, in practice, more features are required to maintain high accuracy against real world distortions, which further increases the demand of resources and delay for computation and transmission. Therefore, a low bit rate, low complexity, and low latency image matching architecture with high accuracy for MAVS system is studied in this thesis for fast and accurate CBIR.

2.10 Performance evaluation measurements for CBIR

Recall, precision and mean average precision are commonly used in information retrieval research for performance and correctness measurement. In this section, the definition of these measures is discussed in the context of targeted MAVS applications.

2.10.1 Recall

The general form of Recall can be described as (2.6). This measures the ratio of the retrieved items that are relevant to the query item. In the targeted MAVS applications, Recall measures the fraction of images that are corresponding to the captured image (i.e. images contain the view of object in the captured image).

$$Recall = \frac{Number\ of\ retrieved\ Relevant\ items}{Number\ of\ relevant\ items\ in\ the\ dataset} \quad (2.6)$$

2.10.2 Precision

The generally form of Precision can be described as (2.7). Precision is used to measure if the retrieved items are relevant to user's need or not. This is an important measurement in the targeted MAVS applications. Ideally, the retrieved images should correspond exactly to the captured image. It is trade-off to balance between

recall and precision, which is known as Receiver Operating Characteristic (ROC) analysis [181]–[183]. The ROC curve shows the true positive rate against false positive rate at a threshold in the system.

$$Precision = \frac{Number\ of\ retrieved\ Revelant\ items}{total\ Number\ of\ retrieved\ items} \tag{2.7}$$

2.10.3 Mean Average Precision

The recall and precision are a single measurement based on the retrieval list returned from the search system. However, it is also vital to consider the Rank (i.e. the position) of the relevant content in the returned list as it is desirable that the first returned image in the targeted MAVS applications is the correct correspondence to the captured image so that the pre-defined augmented content can be precisely triggered and delivered to users.

As presented in [181], [183], by counting the precision and recall of each relevant content according to its order in the returned list, the precision at a certain rank r for a captured image can be described by (2.8):

$$P(r) = \frac{(number\ of\ relevant\ images\ of\ rank\ r\ or\ below)}{r} \tag{2.8}$$

The average precision can be defined as shown in (2.9)

$$\begin{gathered} Average\_P = \frac{1}{R}\sum_{r=1}^{N} P(r)rel(r) \\ rel(r) = \begin{cases} 1, & if\ the\ image\ at\ rank\ r\ is\ relevant \\ 0, & otherwise \end{cases} \end{gathered} \tag{2.9}$$

where N is the number of retrieved images, R is the number of relevant images. This is the measurement for a single query. To evaluate the accuracy of a retrieval system, Mean Average Precision (MAP) is proposed for system measurement based on average precision as shown in (2.10):

$$MAP = \frac{1}{Q} \sum_{q=1}^{Q} Average\_P(q) \tag{2.10}$$

where Q is the number of queries.

As mentioned above, it is crucial for targeted MAVS applications that the first returned image is the exact correspondence to the captured image. Therefore, precision @ 1 MAP is proposed to evaluate such performance as a special condition

$$rank1\ MAP = \frac{1}{Q} \sum_{q=1}^{Q} P(q)$$
$$P(q) = \begin{cases} 1, & the\ rank\ 1\ matched\ image\ is\ correct \\ 0, & otherwise \end{cases} \tag{2.11}$$

of (2.10) as shown in (2.11):

2.11 MAVS application system

In this section, three system architectures that possibly can be employed in MAVS applications are discussed: a) sending compressed image mode; b) sending compact features mode; c) on-device process mode [172], [184]–[186]. The architectures of these three modes are shown in Figure 2.10.

The first mode uses the existing image processing technologies available on mobile devices to send the compressed image to the remote server. All the time consuming computations, such as feature extraction, feature matching, relevant content searching and etc., are performed on the server side by taking the advantages of the much more powerful computation capabilities. The second mode is emerging because of more powerful mobile devices are available on the market and some complicated processing such as image analysis and feature extraction can be deployed on the client side to reduce computation burden on the server side when large numbers of users are active simultaneously. The third mode puts the entire processing mainly on the mobile devices with a local database. The local database

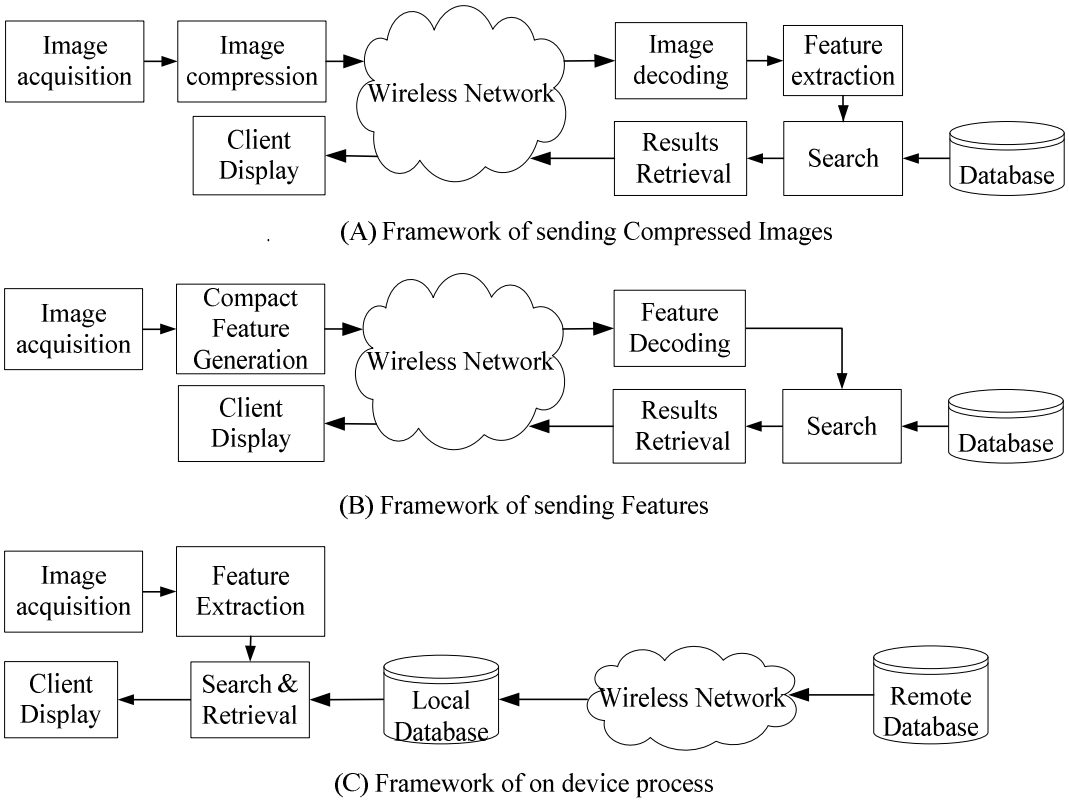


Figure 2.10 Possible MAVS application architecture

can be updated by a remote server to refresh the data. No matter what architecture was employed, the first primary concern would be to maximize the user's QoE. The system latency should be minimized. The retrieval results should be accurate and satisfy the user's expectation. More importantly, the system should be robust to a variety of realistic distortions. In this thesis, the ultimate goal is to develop an efficient MAVS system which is fast, accurate and robust against noise to provide good QoE to users.

2.12 MPEG-7 Compact Descriptor for Visual Search

The work conducted in this thesis is partially motived by an on-going MPEG-7 standard called the Compact Descriptor for Visual Search (CDVS). A brief review of

CDVS is given in this section and some technologies and evaluation methods in CDVS are absorbed in this thesis.

The aim of CDVS is to standardize technologies, in order to enable an interoperable, efficient and cross-platform solution for internet-scale visual search applications and services [153], [187]. It will ensure interoperability of visual search applications and databases, simplifying design of descriptor extraction and matching for visual search applications. It will also enable low complex, low memory hardware support for descriptor extraction and matching in mobile devices and sensibly reduce load on wireless networks carrying visual search-related information. The main characteristics of the current CDVS white paper are summarized in Table 2-5.

Table 2-5 Summary of the goal of CDVS

Goal	characteristics
Compression efficiency	512 bytes to 16Kbytes (512B, 1KB, 2KB, 4KB, 8KB, 16KB)
Scalability	different size by a different number of local features
Web-scale databases search	Global descriptor embedded for fast search by generating a limited set of candidates for further refinement
Hardware implementation efficiency	low computational complexity, small memory footprint, low-power hardware implementations(SoC)
Generality	Any textural rigid objects, such as books, CDs, landmarks, printed documents, DVDs, paintings, buildings
Robustness	An overall amount of 30,000 query images and additionally 1 million images as a distractor set
Sufficiency	Self-contained, easy to combine with other relevant metadata aiming at narrowing the search scope and improving retrieval efficiency

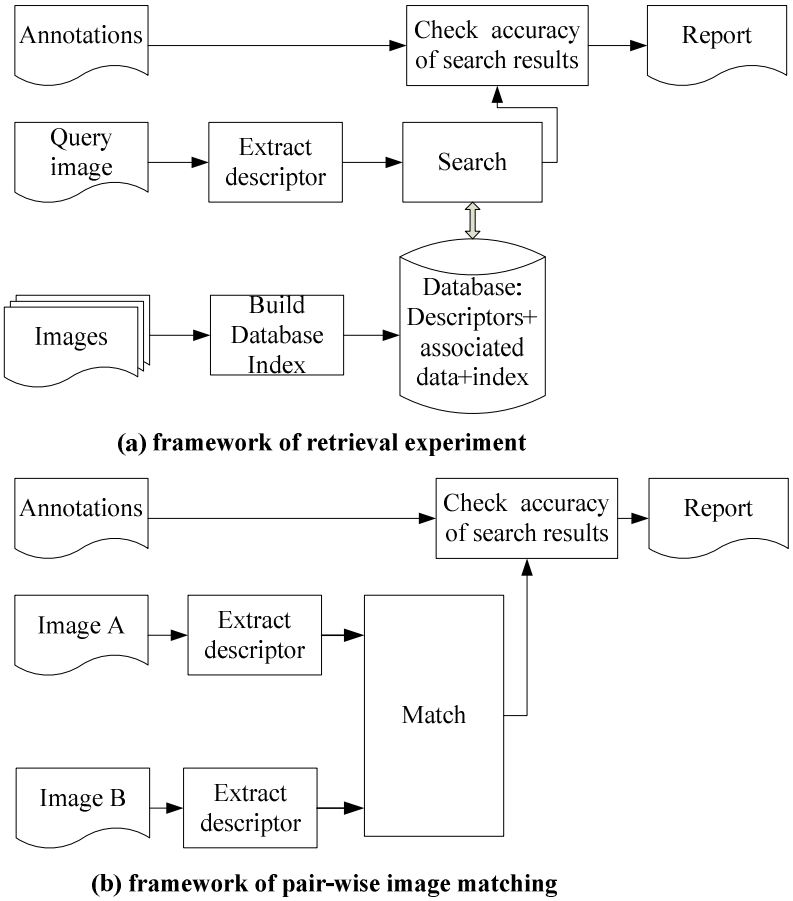


Figure 2.11 The evaluation schemes of CDVS

To achieve the goal mentioned in Table 2-5, the SIFT is chosen as the key feature in the CDVS which achieves good matching accuracy [149]–[152], [154]. The output parameters of the SIFT detector is used to perform feature selection to find the most significant local features. The Principal Component Analysis (PCA) and Gaussian Mixture Model (GMM) is used for feature aggregation while the arithmetic coding is employed for feature location compression. The SIFT feature is also used in proposed MAVS application due to its high discrimination leading to high matching accuracy. The feature selection method proposed in this thesis is compared to the method used in the CDVS as well.

To evaluate the performance of the CDVS system, the evaluation schemes of image retrieval and pair-wise image matching is proposed in [188]. The diagram of

the evaluation is shown in Figure 2.11. The evaluation scheme as used in the CDVS standardisation activity is incorporated to evaluate the matching accuracy of the proposed MAVS system.

2.13 Summary

In this chapter, the background of digital imaging on the targeted MAVS applications was firstly introduced. Considering compact visual information representation and transmission, the image compression processors were then discussed. Aiming to maximizing the QoE for the targeted application, the QoE definition, modelling, measurements method and challenges were discussed. Then, the research around two key QoE influencing factors, namely waiting time and matching accuracy, were reviewed. The review started from the waiting time management from the point view of QoE in different multimedia services that are web QoE and streaming service QoE. Then the fundamental technologies to ensure the matching accuracy were reviewed. Feature detection and extraction algorithms were reviewed from the aspects of feature discrimination and computation efficiency. Followed by pair-wise image matching methods, the merits and demerits of three different matching methods were discussed. Considering both the fast processing speed and improving the matching accuracy to ensure the QoE perceived by users, efficient and fast matching method and geometric verification were reviewed. Moreover, the feature selection which potentially can be beneficial for improving matching accuracy and reducing processing delay was discussed. Following the reviews of image matching technologies, the content based image retrieval was discussed focusing on fast and accurate performance. Then the performance evaluation methods were reviewed. Sequentially, several possible MAVS system architectures were reviewed. Finally,

an on-going MPEG-7 standard called CDVS which is closely related to the work in this thesis was reviewed.

In the next chapter, an extensive evaluation of the performance of different features under image compression and joint photometric distortions are conducted in the context of targeted MAVS applications using the new performance metric known as precision @ 1. This aims to find the most discriminative solution for realistic distortion in the targeted MAVS application.

3 MATCHING ACCURACY OF STATE-OF-THE-ART LOCAL FEATURE ALGORITHMS UNDER REALISTIC DISTORTIONS

3.1 Introduction

As the MAVS applications targeted in this thesis directly trigger the display of the first returned relevant multimedia content, the matching accuracy is a crucial QoE factor to provide good QoE to users. It is obvious that users' perceived quality would deteriorate if incorrect or irrelevant content was delivered to users. Therefore, the study starts from the question of how to ensure the matching accuracy of a MAVS application is maximised that directly links print media viewed through a camera with augmented multimedia content. This chapter presents an extensive evaluation of state-of-the-art local feature detectors and descriptors utilised by MAVS applications under realistic distortions. Different combinations of various local feature detectors and descriptors are investigated by means of matching accuracy as measured by precision @ 1 to study the matching accuracy in a print media dataset.

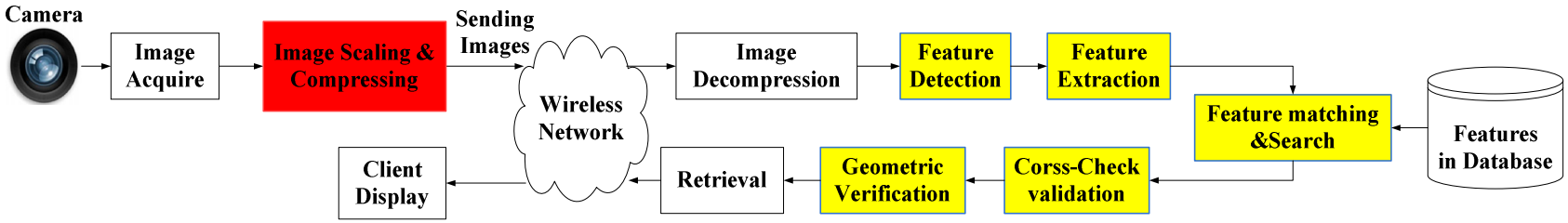
Mobile phone cameras have entered the multi-mega-pixel era. The acquired image is becoming larger and larger. As reviewed in section 2.3, the raw captured image is normally compressed by an image coder (e.g. widely used JPEG) for efficient data storage and transmission. Image fidelity deteriorates during compression, especially for low bit rates. In section 3.2 and section 3.3, to study the influence of image compression on matching accuracy of various local feature algorithms, three image compressors are employed to compress the query images captured by mobile phone cameras from a high bit rate to a very low bit rate. Then, a retrieval evaluation of a matching system using various local feature algorithms on these compressed images is performed. Based on the evaluation results, the trade-off

of deploying various local feature algorithms in the MAVS application from the aspects of matching accuracy and processing time for maximising QoE is discussed.

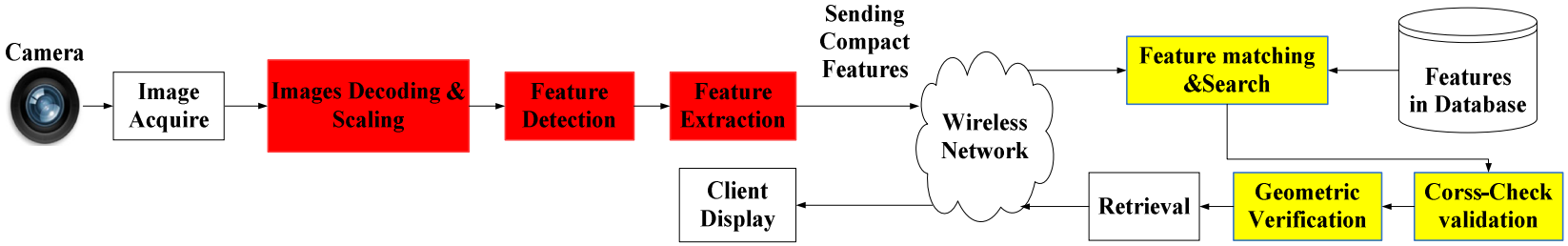
Recall from the review in Section 2.2 and 2.4.4, the image distortions created during camera shot can cause problems for image matching and retrieval on mobile devices. From section 3.4 to section 3.7, two common photometric distortions, namely image blur and illumination distortions are examined to study their joint effects on image matching accuracy for the MAVS. The investigation starts from the analysis of a photometric distortion model for a mobile phone camera. Then, keypoints clustering and fast retrieval based on a KD-tree search is considered in the evaluation system with various state-of-the-art local feature algorithms on the controlled distortions. Finally, the influence of different image types and different cameras on matching accuracy is discussed.

3.2 Image Compression for MAVS Applications

There are several potential system architectures for MAVS applications as reviewed in section 2.11. The first scenario is to transmit a compressed query image to the server. The whole time-consuming matching process is performed in the server by taking advantage of the powerful computation and memory resource of the server. This architecture can make use of the existing image acquired and processing technologies on mobile devices and incurs the least processing on the device at the client's side. However, this scenario requires a large transmission bandwidth without efficient image compression. Meanwhile, an optimal solution to balance the image compression ratio and image matching accuracy is required. The second scenario is to perform the feature detection and extraction on the mobile device and then transmit the features to the server for image matching and relevant content retrieval.



(A) Framework of sending Images



(B) Framework of sending Features

Figure 3.1 System architecture of A) sending compressed images and B) sending compact features. Red block indicates the process on the client's mobile device; Yellow block indicates the process in the server.

In this case, feature detection and extraction are performed on the client side after the image is shot. And then, a compact local feature descriptor is used to transmit a smaller amount information to the server compared to sending the original captured image. Extra processing is required on the client's mobile devices. The third scenario is to download a database of images (or features representing the images) from the server and all processing is performed on the mobile device. The third scenario is only suitable for small scale databases and frequent updates are required if the database refreshes daily. Considering that the print media updates frequently and includes a medium to large dataset, the first two scenarios are discussed in this thesis.

Considering the first two scenarios, the whole process can be subdivided into several modules which are shown in Figure 3.1:

(a) capturing the image;

(b) detecting and extracting the image features using computer vision algorithms;

(c) transmitting the features or compressed image between mobile devices and server by wireless technologies (e.g. 3G network);

(d) searching and matching to find corresponding relevant content;

(e) sending retrieved content to the user.

Recall the review in section 2.4.4, from the perspective of the user perceived quality of experience, overall system latency (i.e. the time taken from capturing the image to displaying relevant content to user) and the matching accuracy are of prime importance. Firstly, there is a latency associated with each part of the process. The latency associated with different parts of the system are determined by the performance of the mobile camera, speed of the wireless network and size of the transmission, computation of the employed algorithms and processing capacity of the devices on which algorithms are processed. On the other hand, the performance of

employed local feature algorithms are influenced by distortions (i.e. lighting, blurring and rotation) that occur when capturing the image as well as distortion caused by image compression. The matching accuracy is therefore influenced by the robustness of the local feature algorithms and matching methods to these distortions. Ideally, the ultimate goal is to develop an MAVS application with low system latency and high matching accuracy. Hence, the technological deployment strategy is to minimize the processing time on the client's mobile device, minimize the size of the transmitted data and minimize the overall latency while at the same time maximizing the matching accuracy. To achieve such a goal, there are different trade-offs to be made between the system constraints.

The first primary concern is to find an optimal local feature for MAVS application, which can achieve high matching accuracy. To study the performance of various local feature algorithms, previous evaluations in [79], [98] focused on the precision and recall of the local features under varying image transform distortions, such as lighting changes, image blurring, rotation, scaling and JPEG compression. But, these evaluations solely investigated whether the same local features could be repeatedly detected and be matched correctly in a local feature sets derived from image pairs. It is more crucial whether the correctly matched feature pairs can result in accurate relevant content retrieval in a MAVS application as reviewed in section 2.10. Therefore, an extensive evaluation is conducted using precision @ 1 to measure the ability of various local feature algorithms of finding a corresponding image from a realistic print image dataset with varying distortions.

Another concern is to break through the bottleneck of wireless transmission limitation. In [172], Chen et al. have employed the scenario of sending image features (Figure 3.1 (B)). They examined the performance in terms of retrieval

accuracy, system latency and power consumption when using features specifically designed for low bit rate transmission (i.e. CHOG features) rather than transmission of the JPEG image. They achieved over 94% accuracy by sending compressed CHOG features with a total size of 5 KB. They also discovered that the accuracy of sending the JPEG image (Figure 3.1 (A)) deteriorated to 89% as long as the query image was compressed to 10 KB in their image matching system. In their application, the time to extract features of one image on a typical smart mobile device required approximately 1s, which, when added to the feature transmission time, is still two times smaller than sending the JPEG image in a typical 3G network. But in the same work in [172], the situation reversed when the data is transmitted via a typical Wireless Local Area Network (WLAN) due to the processing delay on the mobile device becoming more significant than the transmission delay. A hypothesis is that if compressed image sizes can be made similar to compressed features sets (e.g. CHOG features) whilst maintaining matching accuracy when employing the scenario of sending compressed image, a significant reduction in processing delay on the mobile devices, battery power consumption, system latency and bandwidth can be achieved by taking advantage of server to do the majority of processing. A further advantage of this approach is that it allows flexibility in choosing the image features used within the matching process e.g. to enable adaptive algorithms that use the most optimal features for a given type of image. In [172], only JPEG is considered, while in this work, two more efficient and standard image compression algorithms namely JPEG2000 [43], [44], [189] and HDPhoto [38], [40] are employed for investigation. Several state-of-art local feature algorithms, including various feature detectors and descriptors available in the OpenCV computer vision library [190], are discussed in the next section for investigation.

3.2.1 Keypoint detection and feature description

Local feature algorithms are composed of a keypoint detector and a feature descriptor. Today, the most successful algorithms for content-based image retrieval all aim to detect salient interest points in the image. Recall from the review in section 2.6, a keypoint detector detects the salient interest points, known as keypoints in the image. The detected interest points should be repeatable, distinctive, of sufficient quantity and efficient [79]. The property of repeatability requires that a high percentage of accurate keypoint detections under different perspective changes and deformation, such as scale down, rotation or compression artefacts and photometric deviations. Many authors have addressed the problem of keypoint detection in [79], [82], [98], [100] to name a few. A variety of interest point detectors have been implemented and are available in the OpenCV library which utilize intensity based algorithms as reviewed in section 2.6.1, including FAST [94], [97], STAR [84], SIFT [191], MSER [81], HARRIS and GFTT [85], [86], SURF [92], AGAST[192], ORB [82], BRISK [96]. It is known that these different detectors give different performance in terms of computation and accuracy. SIFT is robust to the scale changing and rotation but is extremely slow in processing speed with high computation and memory storage requirements. SURF is faster than SIFT but with worse performance for rotation. FAST is extreme fast in interest point detection while it offers worse performance in repeatability. ORB is based on FAST and has better performance by adding a fast and accurate orientation component to FAST [82]. AGAST can achieve faster corner detection compared to FAST while BRISK added scale invariance by filtering the detected keypoints in the scale space [96], [192].

After keypoint detection, the feature descriptor is used to describe the region of the image around each keypoint, also known as the image patch, based on certain characteristics. Generally, such feature descriptions are used to perform feature matching in a reference local feature set to find the matching feature pairs. Five descriptors available in the OpenCV library are employed for investigation, which are the SIFT descriptor, SURF descriptor and ORB descriptor, BRISK descriptor and FREAK descriptor. SIFT and SURF descriptors are floating-point descriptors while ORB, BRISK and FREAK descriptors are binary descriptors. The reasons for choosing these descriptors are:

1) SIFT and SURF descriptors combined with corresponding keypoint detectors are well known for high discrimination and the robustness against scale variation and rotation variation, which are two main geometric distortions when capturing a query image in an MAVS application;

2) ORB, BRISK and FREAK are binary descriptors, which are fast and computationally-efficient when performing feature pair matching as only an XOR operation is required. These three binary descriptors are robust to image noise. The ORB descriptor is especially designed for rotation invariance [82]. BRISK achieved scale invariance [96]. FREAK achieves low memory and computation complexity while remaining robust to scale, rotation and image noise.

The problems of deploying SIFT and SURF on the mobile devices are that these two descriptors require a high computational complexity and consume a large bit rate for representation and transmission. The SIFT descriptor proposed by Lowe [191] is a 128-dimensional distinctive descriptor, which uses 1 Kbit to represent each feature. The SURF descriptor proposed in [92] is a 64-dimensional descriptor using 512 bit per feature, which is much smaller than SIFT. However, considering a rich

content image with hundreds of features, sometimes, this causes the total size of all feature descriptions to be larger than the original compressed image. Several compression schemes have been proposed to reduce the bit rate of SIFT and SURF descriptors, such as transform coding and vector quantization, while Principal Component Analysis (PCA) has been employed to reduce the dimensionality of SIFT and SURF descriptors [99], [179]. However, these methods are complex and require a high computational load which is not suitable to deploy on mobile device. But, if using the scenario of sending compressed images, these problems can be solved as long as it can still achieve high matching accuracy. Binary descriptors, like ORB, BRISK and FREAK, were designed to make descriptors faster to compute, match and more compact and suitable to deploy in a mobile devices. However, how to achieve high accuracy is still a concern.

In this work, different keypoint detectors and descriptors in the OpenCV library are combined to study the performance under varying image compression schemes when using aforementioned image compressors. The experimental evaluation methodology is introduced in the next section.

3.2.2 Evaluation system

This section describes how the image matching based on local features is performed in a typical MAVS application. Two types of images are defined in the matching system of a MAVS application. One type of image is denoted as the query image, which is the image captured by a camera of mobile device and then compressed and transmitted to the server to perform image matching. Another type of image is denoted as the reference image, which is a bunch of target images pre-stored in the

server. Ideally, a query image is corresponding to a reference image in the server in a MAVS application.

It is assumed that a query descriptor set derived from a query image A_k, where k denotes the image number, can be denoted as:

$$A_k = \{\alpha_i^k\}, i \in M \tag{3.1}$$

Where M is the dimension of the query descriptor set.

The reference descriptor set from reference image B_l, where l denotes the reference image number, can be denoted as:

$$B_l = \{\beta_j^l\}, j \in N \tag{3.2}$$

Where N is the dimension of the reference descriptor set.

$F_{A\to B}$ is the set of minimum Euclidean distance of each descriptor from A to B:

$$F_{A\to B} = \{f_{A\to B}^i\} = \left\{\min_j \|\alpha_i - \beta_j\|\right\}, i \in M, j \in N \tag{3.3}$$

Meanwhile, $F_{B\to A}$ is the minimum Euclidean distance set of each descriptor from B to A:

$$F_{B\to A} = \{f_{B\to A}^j\} = \left\{\min_i \|\beta_j - \alpha_i\|\right\}, j \in N, i \in M \tag{3.4}$$

As reviewed in section 2.7.1~2.7.3, to achieve the best feature matching accuracy, the cross-check feature matching method is employed. The matching collection derived from the intersection of $F_{A\to B}$ and $F_{B\to A}$ can be denoted as:

$$M_{AB} = F_{A\to B} \cap F_{B\to A} \tag{3.5}$$

The matching collection only corresponds to descriptor pairs where the Euclidean distance is the nearest both in $\{f_{A\to B}^i\}$ and $\{f_{B\to A}^j\}$. After that, as reviewed in section 2.7.5, Geometric Verification is performed on M_{AB} using RANSAC [126] as shown in (3.6) to filter the outliers from M_{AB}, which produces the final matching collection S:

$$\|dstpt_i - (\mathbf{N} * srcpt_i)\| > T_d \quad (3.6)$$

where $dstpt_i = \begin{bmatrix} x_i \\ y_i \\ 1 \end{bmatrix}$ is the coordinates (x_i, y_i) of the keypoints in the query image; $srcpt_i = \begin{bmatrix} x'_i \\ y'_i \\ 1 \end{bmatrix}$ is the coordinates (x'_i, y'_i) of the points in the reference image; $\mathbf{N} = \begin{bmatrix} n_{11} & n_{12} & n_{13} \\ n_{21} & n_{22} & n_{23} \\ n_{31} & n_{32} & 1 \end{bmatrix}$ is the matrix which represents perspective transformation such as rotation [80]; **N** is estimated starting from identity matrix by RANSAC method, which minimises the error of $\sum_i \left(x'_i - \frac{n_{11}x_i + n_{12}y_i + n_{13}}{n_{31}x_i + n_{32}y_i + 1}\right)^2 + \left(y'_i - \frac{n_{21}x_i + n_{22}y_i + n_{23}}{n_{31}x_i + n_{32}y_i + 1}\right)^2$. T_d is the maximum allowed re-projection error to treat a keypoint pair as a true positive. T_d is set as 3 in the experiment which distinguishes the major inliers from outliers across the given image dataset [193].

In this experiment, each query image is matched to 100 reference images using the above procedure. Thus, there are collections W_k after image matching for each query image:

$$W_k = \{S_l\}, l = 1,2,\dots,100, k = 1,2,\dots,100 \quad (3.7)$$

The *k*-th query image is judged to match to *l*-th reference image only if the cardinality of S_l, denoted as $|S_l|$ (i.e. the total number of matched descriptors), is the maximum as shown in (3.8):

$$|S_l| = max\{|W_k|\} \quad (3.8)$$

According to the ground-truth annotation list, when the matched reference image is exactly corresponding to the query image, it is correctly matched image as shown in (3.9).

$$P(k) = \begin{cases} 1 : k = l \\ 0 : k \neq l \end{cases} \quad (3.9)$$

Subsequently, the matching accuracy is calculated using precision @ 1 as reviewed in section 2.9.3:

$$precision\ @\ 1 = \frac{1}{Q}\sum_{q=1}^{Q} P(q) \tag{3.10}$$

Where Q is the number of query images in the database, which is 100 in this case.

3.2.3 Experimental dataset and methodology

Using the matching algorithm described in Section 3.2.2, experiments were conducted to evaluate the performance of different combinations of feature detectors and descriptors chosen from the OpenCV library when extracting local features from images compressed to different bit rate using different image compressors. The performance was measured in terms of matching accuracy of precision @ 1 and processing time (i.e. includes image compression time and image matching time) for a dataset of images captured from a mobile (cell) phone. The experimental machine used for these experiments equipped with an Intel Core i7 2.93GHz CPU and 4GB RAM.

Focusing on print media, like book covers, a subset of the Stanford Mobile Visual Search (MVS) dataset [194], which is composed of a wide range of images captured by a high-end mobile phone camera, was used for evaluation. A few examples of the images are shown in Figure 3.2. In the experiment, 100 images from the book cover image sets captured by a 5.0 Megapixel Android mobile phone camera were chosen. These images are captured under a variety of realistic distortion conditions such as varying light conditions, random rotation, foreground and background clutter, which are common distortions that will occur in a MAVS application. The MVS dataset also provides corresponding ground-truth images which are clean versions downloaded from book publication websites with an image

Figure 3.2 Examples of query and reference image pair from dataset. Clean version pictures are matched against captured image with various distortion.

resolution of 400*400 pixels. To prepare the dataset for experimental usage, the clean versions are taken as reference images in the evaluation system. The images chosen from the "book_cover" catalogue in the MVS dataset are scaled to 400*400 resolution, which is comparable to the reference image resolution, and then further compressed with varying compression ratios by using different image compressor to generate query image dataset. For each compression ratio, a scaled book cover image is compressed by applying JPEG, JPEG2000 and HDPhoto compression, individually, which generate 3 versions of compressed images:

1) For JPEG compression, the JPEG conversion program developed by Independent JPEG Group is used to compress the image to JPEG format. The default lossy compression mode in compliance with JPEG standard [34] is used to compress the image to different compression ratio by setting different image quality factors using '-q' option as shown in Table 3-1;

2) For JPEG2000 compression, considering JPEG2000 offers better rate-distortion performance than JPEG under the same compression ratio [189], the kakadu software in compliance with JPEG2000 standard is employed to

compress the image to JPEG2000 format [195]. By using irreversible compression mode (i.e. lossy compression) and different bit/sample factors, the image is compressed to different compression ratio as shown in Table 3-1;

3) HDPhoto, also known as JPEGXR, is dedicated to a high dynamic range image codec developed by Microsoft as part of the Windows Media family, which offers better performance than JPEG [38]. The default lossy compression mode is used to compress the image to different compression ration by setting different quantization level factors (similar to JPEG) as shown in Table 3-1;

In summary, the query image dataset consists of various compressed versions of the original "book_cover" images in JPEG, JPEG2000 and JPEGXR formats, with file sizes ranging from 3 KB to 30 KB. For each specific compression ratio, 300

Table 3-1 Summary of Image Compression Parameters for Different Image Compressor

Image bit rate (KB)	3	4	6	8	10	12	15	20	25	30
JPEG quality factors	3	7	14	24	35	47	63	77	85	89
JPEG2000 Bit/sample (10^{-3})	150	203	304	407	510	613	765	1023	1278	1537
HDPhoto quantization factors	77	69	61	53	48	42	36	28	21	16

Table 3-2 Summary of Investigated Different Combinations of Local Feature Detectors and Descriptors. (Detector is indicated by italic)

Detector-Descriptor	*Detector*-Descriptor	*Detector*-Descriptor
SIFT-SIFT	*ORB*-ORB	*AGAST*-FREAK
SURF-SURF	*MSER*-ORB	*SIFT*-FREAK
FAST-ORB	*GFTT*-ORB	*MSER*-FREAK
STAR-ORB	*HARRIS*-ORB	*ORB*-FREAK
SURF-ORB	*BRISK*-BRISK	*GFTT*-FREAK

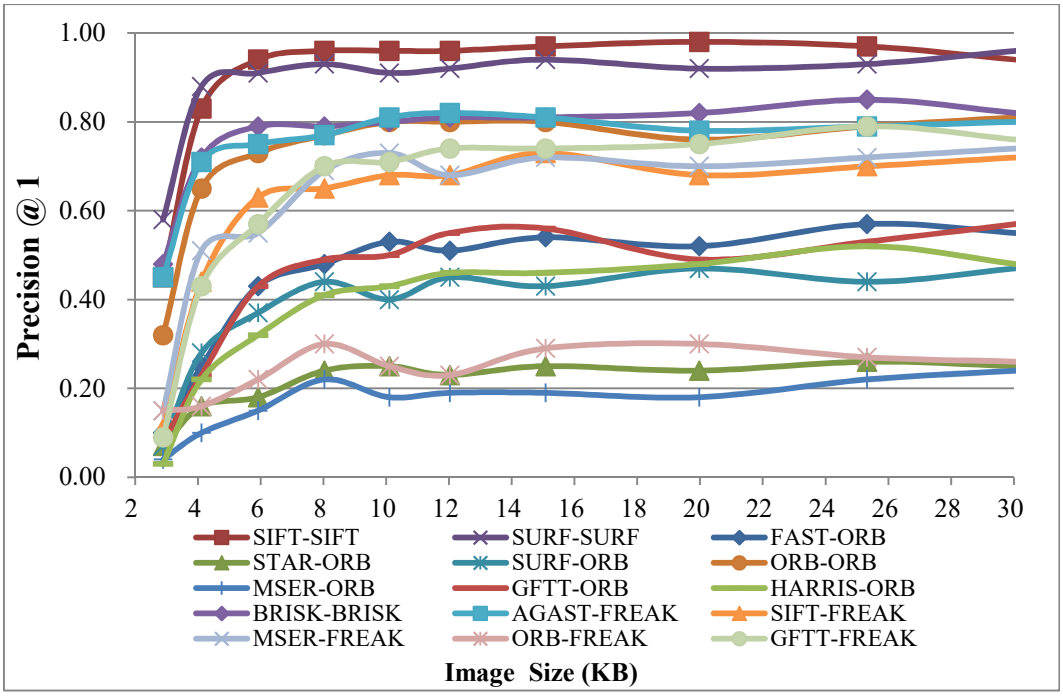
query images are studied with the combination of various feature detectors and descriptors. The compression control factors of different compressors are summarized in Table 3-1. Subsequently, after preparation of query images and reference images in the test dataset, different combinations of local feature detectors and descriptors, as shown in Table 3-2, are employed to detect and extract the local features and then create descriptors both in the query image and reference image for use within the matching system described in section 3.2.2.

3.3 Comparison of various combinations of local feature algorithms when applying different image compressor in a MAVS application

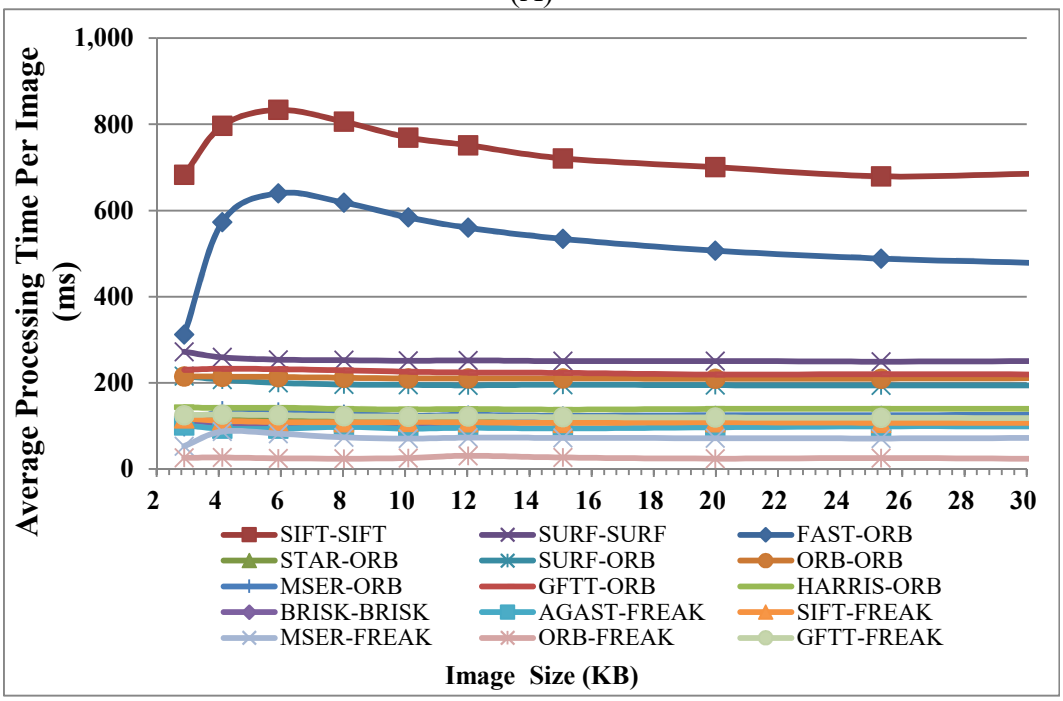
The effects of image compression using JPEG, JPEG2000 and HDPhoto on MAVS application when employing the system architecture of sending compressed image are studied in this section by means of precision @ 1 and processing time.

3.3.1 The influence of JPEG lossy coder

Figure 3.3-(A) shows the precision @ 1 whilst Figure 3.3-(B) shows the processing time with JPEG compression variation, respectively. Different combinations of various feature detectors and descriptors demonstrate different performance. When the compressed image size is greater than 10 KB, the precision @ 1 of all local feature algorithms only has a little variation, which indicates that minor image compression using JPEG does not influence matching accuracy a lot. Especially, SIFT and SURF algorithms achieve high matching accuracy of around 96% and 93% precision @ 1, respectively. With further compression to 6 KB, the precision @ 1 of all local feature algorithms starts to decrease by an average 2%. In another words, the moderate JPEG compression starts to affect the matching accuracy. With further compression to low bit rate, the decrement in matching accuracy becomes more



(A)



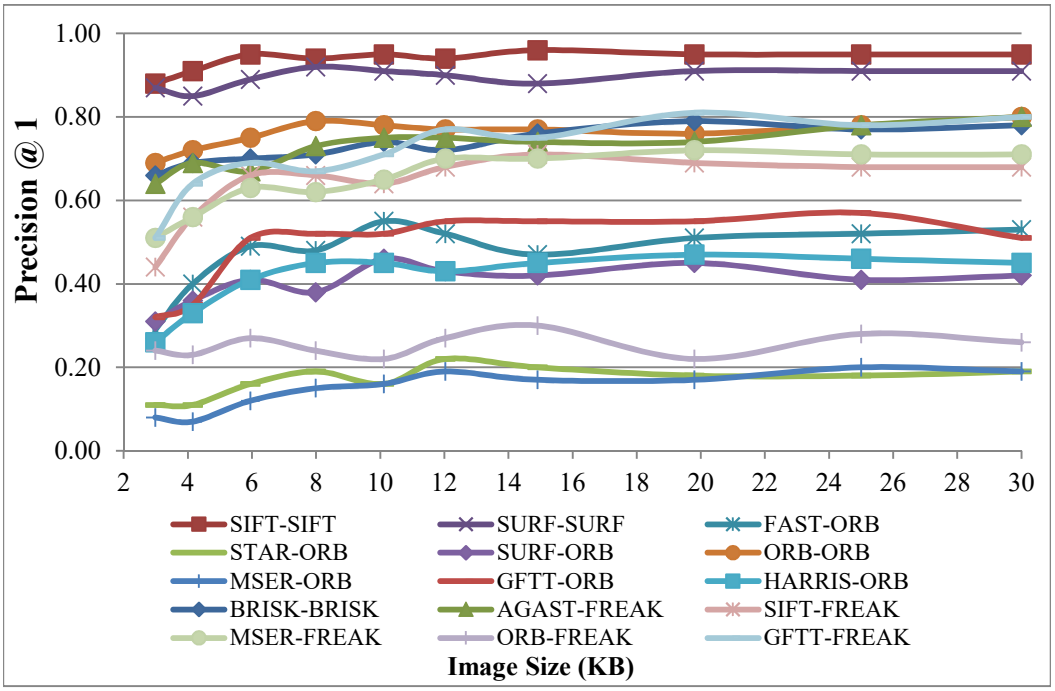
(B)

Figure 3.3 The performance of various combinations of feature detector and descriptor under different JPEG compressed image bit rate: (A) Precision @ 1 (B) Processing time.

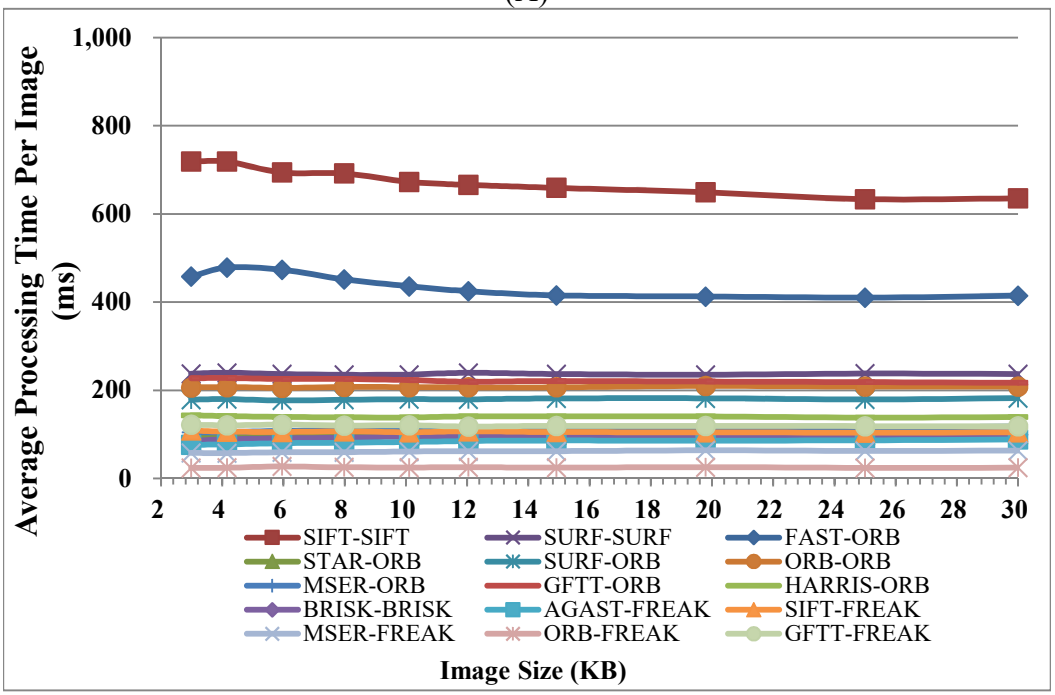
dramatically. Another vital aspect needed to be pointed out is that the matching accuracy is affected by two combined influences under high JPEG compression; one factor is the distortion arisen from the image compression and another factor is the robustness and stability of the feature algorithms applied to the image. For example, when the HARRIS detector is applied to JPEG compressed image in Figure 3.3-(A), the precision @ 1 at 25KB is a little bit higher because of the instability of the HARRIS keypoint detection. The similar situations can be found when using GFTTORB at 15KB, MSERFREAK at 10KB and ORBFREAK at 8KB, respectively. The floating descriptors, such as SIFT and SURF, achieve superior matching accuracy than binary descriptors. However, the processing time of the floating descriptor is much longer than the binary descriptor as expected. The only exemption is that when using FASTORB, too many keypoints are detected from the compressed image, which results in long processing for feature pair matching and geometric verification. From Figure 3.3-(B), overall, the processing time increases slightly with the increase of compression until 6KB, mainly due to the number of features growing as well as a little more geometric verification time. After that, the number of features decreases dramatically, which reduces the time. The most time consuming algorithm is SIFT, which is still less than 1s on average to processing one image.

3.3.2 The influence of HDPhoto lossy coder

Figure 3.4 shows the precision @ 1 and processing time with HDPhoto compression variation. When the compressed image size is greater than 6KB, the matching accuracy of all local feature algorithms has an approximate reduction of 1% while SIFT and SURF are around 95% and 90%, respectively. With further compression to 4KB and below, the decrement in matching accuracy becomes more dramatically.



(A)



(B)

Figure 3.4 The performance of various combinations of feature detector and descriptor under different HDPhoto compressed image bit rate: (A) Precision @ 1 (B) Processing time.

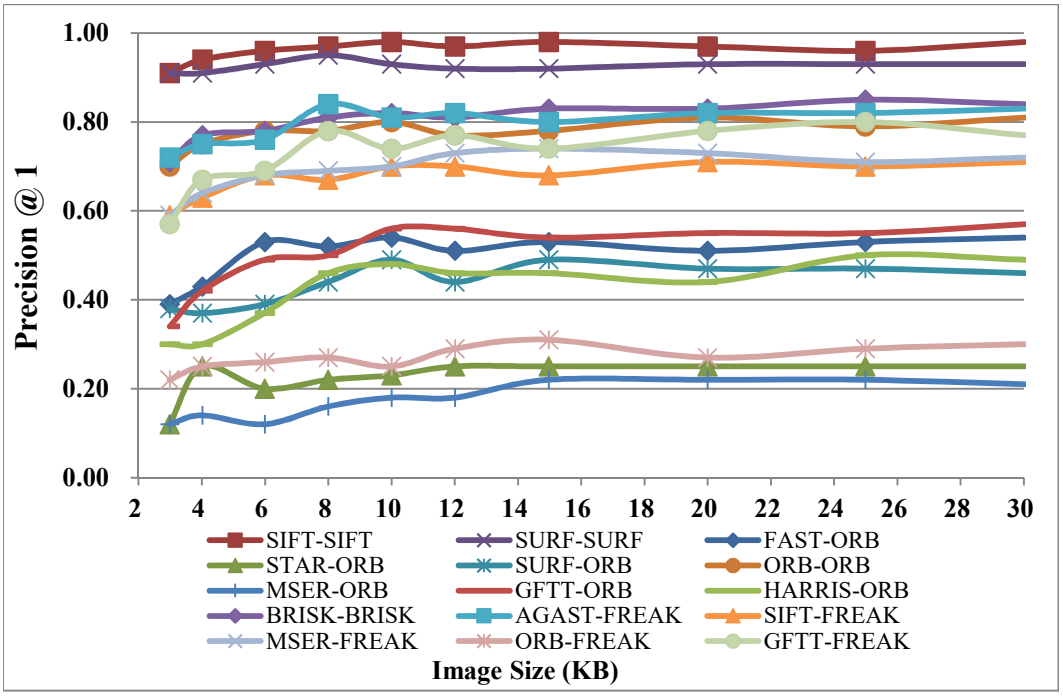
The reason for processing time variation when applying various local feature algorithms to HDPhoto compressed images is similar to JPEG. The number of detected features is reduced along with the increase in compression ratio below 4KB. But the performance both in terms of the precision @ 1 and processing time are more stable compared to JPEG, which indicates that HDPhoto compression has less impact on local feature algorithms compared to the JPEG compressor because of the better rate-distortion.

3.3.3 The influence of JPEG2000 lossy coder

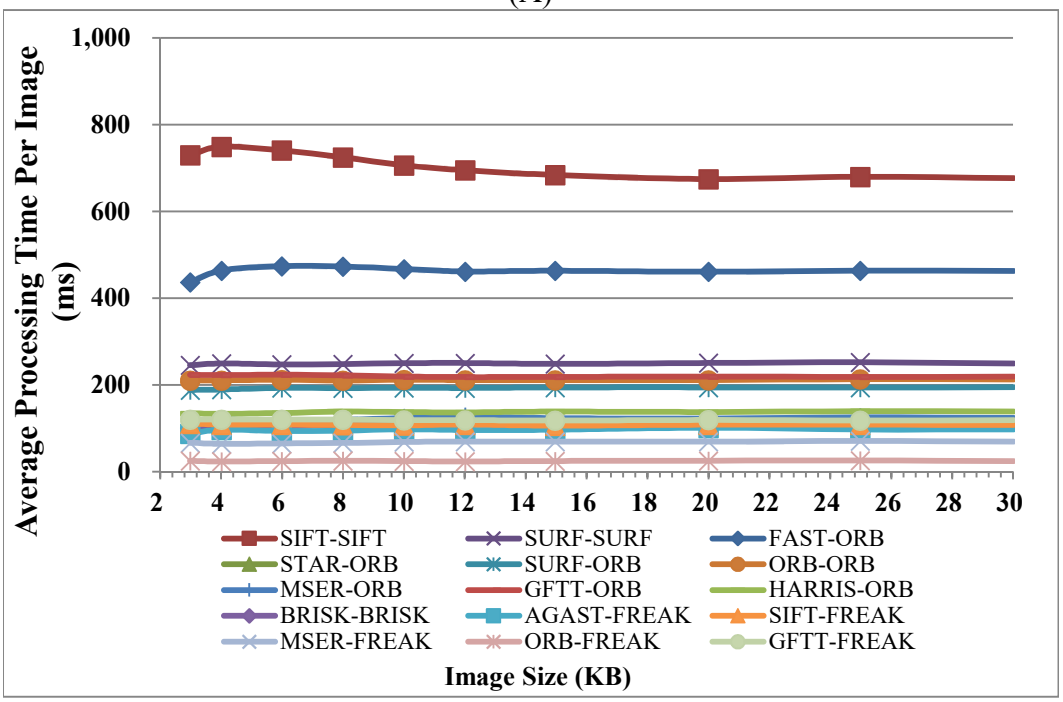
Figure 3.5 shows the precision @ 1 and processing time when using the JPEG2000 compressor to compress the image of varying compression ratios. When the compressed image size is greater than 4KB, the matching accuracy of all local feature algorithms also has an approximate change of 1% while SIFT and SURF are around 97% and 93%, respectively. With further compression to 3KB, the decrement of precision @ 1 becomes more obvious. The variation of processing time when employing JPEG2000 compressor is quite similar to that for HDPhoto.

3.3.4 Discussion of the impact of various image coders

To illustrate the relative relationship among these three compression schemes, the precision @ 1 of the SIFT feature with the increasing compression ratio is shown in Figure 3.6 as the SIFT feature achieved the best matching accuracy. Based on the experimental results, it can be inferred that the image size can be compressed to approximate 4KB to 10KB using JPEG2000 and HDPhoto without significant sacrifice of precision @ 1. However, for such size, it is equal to send 32~80 SIFT descriptors or 64~160 SURF descriptors or 128~320 ORB descriptors or 546~1364 CHOG descriptors [100]. Normally, for a 400*400 resolution image with medium



(A)



(B)

Figure 3.5 The performance of various combinations of feature detector and descriptor under different JPEG2000 compressed image bit rate: (A) Precision @ 1 (B) Processing time.

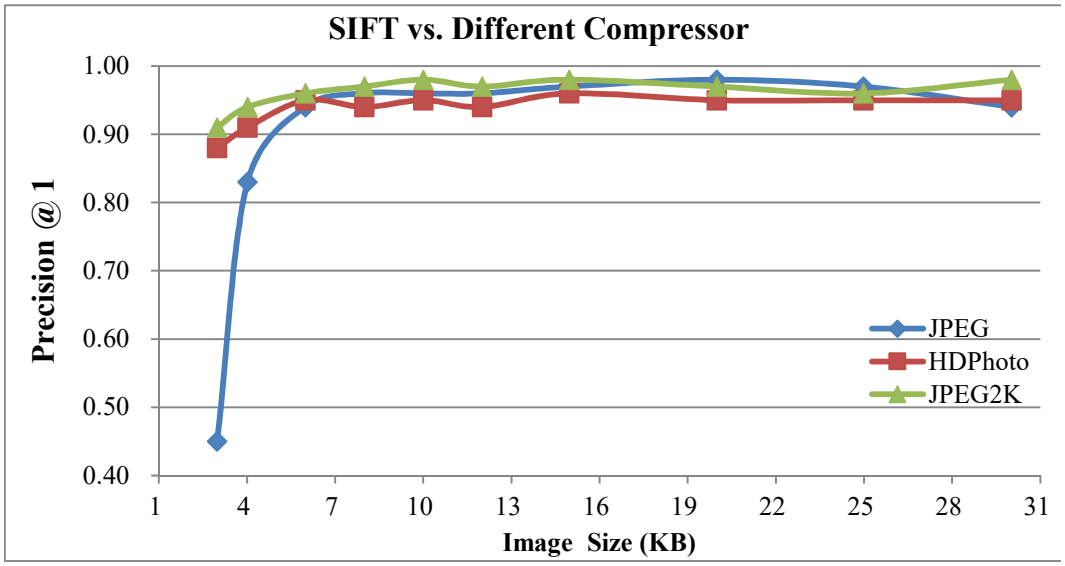


Figure 3.6 The precision @ 1 of SIFT feature algorithm with different compression scheme.

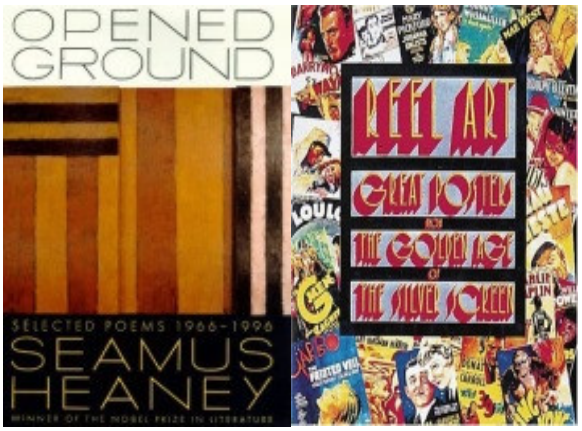


Figure 3.7 Strong interference image in the dataset.

complex content as used in this experiment (see example in Figure 3.2), it will generate over 200 SIFT descriptors or 400 SURF descriptors or 500 ORB descriptors. If all features were transmitted without optimal feature selection, this results in significantly more data than compressed images and hence the transmission time for the sending feature mode (Figure 3.1 (b)) will be larger than sending the compressed image mode without efficient feature compression (Figure 3.1 (a)) as well as battery life. Although the data size can be reduced by only transmitting a subset of detected

features on the basis of employing a certain selection criterion (e.g. Hessian response), this risks reducing the matching accuracy. It should also be noted that even if the whole feature sets are used to within the matching algorithm, the accuracy of the best SIFT feature algorithm cannot reach 100% due to some strong similarities between different images. Figure 3.7 shows an example of two images which have similar visual features (i.e. vertical and horizontal lines), which lead to SIFT features that are similar for each image. Therefore, another method may be needed to tackle the problem when it is hard to find a correct match. In this case, by employing the color histogram, these two images can be treated as different. But this needs to be further studied. For example, a mechanism is needed to decide when to enable the color histogram comparison as it will introduce extra processing delay and hence affect the overall QoE. Moreover, finding an appropriate threshold for color histogram comparison is required.

While transmitting all features leads to unsatisfactory transmission time, an alternative is to send compressed features [172]. For the CHOG descriptor, a 400*400 resolution image will produce 4KB of compressed feature sets. If the upload speed of a 3G network is assumed to approximate 300Kbps which is the worst case as surveyed in [196], it will take 107ms to transmit CHOG features whilst 107ms~266ms to transmit 4KB~10KB JPEG2000 or HDPhoto images. Compared to nearly 1s processing time in a mobile device as evaluated in [172], it is better to send the compressed image than to send features. The only doubt is how long it will take to compress the image with HDPhoto and JPEG2000. Comparing the processing time of the three compressors used in this experiment it was found that they were of the same order of magnitude. Thus, it can be inferred that the complexity of these compressors is also of the same order of magnitude. Hence, the time to conduct the

JPEG2000 and HDPhoto compression on a mobile device will be similar to JPEG and this time (i.e. less than an average 5.3ms in the experiment) is much less than the time required for feature extraction on the mobile device.

It is also can be seen from Figure 3.3 to Figure 3.5 that the most time consuming algorithms are SIFT and SURF while they achieve the best matching accuracy as expected. But the maximum time is still less than 900ms, which contains the overall time for feature extraction and feature matching. While the processing time will be longer on a mobile device compared to the experimental machine, the relative relationship between the processing times of these feature algorithms will be similar for mobile devices. Furthermore, if a more powerful server or cloud is used, the time will further be reduced to achieve the real time requirement by accelerating the processes of feature pair matching, cross-check matching and geometric verification. This makes it feasible to apply SIFT or SURF to a MAVS application.

3.3.5 Conclusion

This section evaluated the effects of image compression within a MAVS application. The performance of different combinations of various local feature detectors and descriptors is investigated under various image compressions from the aspects of precision @ 1 and processing time and the trade-off between sending compressed images and sending features is discussed. Another potential benefit of sending the compressed image is that it does not restrict the feature to a specific algorithm in a MAVS system. Flexible selection of local feature algorithms or combinations can be employed in the server or cloud regardless of the computational constraint in the client side to improve the retrieval performance.

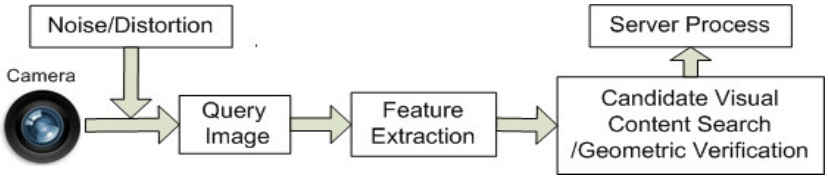


Figure 3.8 A general architecture for a MAVS application. The noise or distortions such as blurring, illumination changes, viewpoint changes and JPEG compression are introduced during the process of capturing. These distortions will influence the matching accuracy when finding candidate visual content from server.

3.4 Joint effect of image blur and illumination distortions for MAVS application

Recall the review in Section 2.6, the discrimination of local features will be affected by distortions that occur when capturing and transmitting images. The matching accuracy is influenced by the robustness of local feature algorithms against these distortions. A lot of research has been conducted to develop robust local feature detectors and descriptors in computer vision like SIFT [80], MSER [81], ORB [82], FREAK [104], AGAST [192]. Many evaluations have been done to study the performance of different local feature algorithms against optical and geometric distortions. For example, SIFT is robust to scale and rotation changes; ORB achieves computational-efficiency and rotation invariance; FREAK is fast to compute and robust to scale rotation and noise; The MSER feature is robust to view changes in edge remarkable datasets such as buildings. However, the existing research has only focused on the repeatability, distinctiveness, and efficiency of local feature algorithms used in visual search under a specific single distortion. Considering a practical MAVS applications, joint distortions will inevitably occur when capturing an image as shown in Figure 3.8: 1) illumination changes due to the ambient lighting condition; 2) image blurring due to object motion or camera motion and out-of-focus;

3) rotation and scale; 4) partial occlusion. These distortions combine together to affect the matching accuracy of a MAVS application. The geometric distortions such as rotation and scale, and partial occlusion are easier to overcome if efficient instructions are provided to users when they are capturing images. However, the illumination changes and blurring are harder to avoid when capturing images from mobile phone cameras because of the uncontrolled lighting environment; poor shooting quality; limited inbuilt image enhancement functionality; and quality of the mobile phone camera.

The effect of image compression on matching accuracy has been studied in Section 3.2 and 3.3. Focusing on the influence of the optical distortions, the effect of joint illumination changes and blurring that inevitably happen during the camera shot is studied with various combinations of state-of-art local feature algorithms. In the next section, a problem description of joint distortion is provided.

3.5 Joint optical distortions

A general MAVS application is to use a mobile phone camera to capture print media and then find the matching reference visual content in a remote server as shown in Figure 3.8. The joint effect of illumination change and blurring when capturing an image of print media is analyzed in this section. Then, a joint distortion model is proposed to add the distortion to the experimental images to study the effect on the matching accuracy of various state-of-art feature algorithms over a wide range of joint distortions.

3.5.1 Effects of global illumination changes during camera shot

Assuming a stable light source in a MAVS application, the illumination will not dramatically change across a specific capture scene, no matter what kind of capture

modes are used in a mobile device camera (i.e. still image mode or video mode). According to the study in [197], it can be assumed that the trichromatic (RGB) color I(x, λ) of pixel x of a certain image or frame under a certain wavelength λ is computed as:

$$I(x,\lambda) = W_d(x)S_d(x,\lambda)E(\lambda) + W_s(x)E(\lambda) \qquad (3.11)$$

$W_d(x)$ and $W_s(x)$ are geometric parameters of the diffuse and the specular reflections resulting from the 3D surface variations, $S_d(x,\lambda)$ is the diffuse reflectance function and $E(\lambda)$ is the spectral energy distribution function of the illumination. Assuming a single illuminant, λ can be approximately treated as invariant across the surface of the print media. In addition, it is assumed that common print media examined in this work has flat surfaces, such as book covers. Therefore, $W_d(x)$, $W_s(x)$ and $S_d(x,\lambda)$ can be treated as constant for the print media investigated in this work. For the image sensors used in a mobile phone camera, the output resulting from $I(x,\lambda)$ at each sensor (for each pixel) can be modeled as follows [198]:

$$Q_{i,j}(x,t) = \int_0^t (i_{ph}(t) + i_{dc})dt + U(t) + V(t) + C \qquad (3.12)$$

Where $Q_{i,j}(x,t) \leq Q_{sat}$, Q_{sat} is the saturation value determined by the hardware device, $Q_{i,j}(x,t)$ is the output pixel at the end of exposure time *t*, $i_{ph}(t)$ is the response of $I(x,\ \lambda)$, i_{dc} is the reference signal predefined in the hardware, $U(t)$ is shot noise, $V(t)$ is the readout circuit noise and *C* is reset noise. Because of the high quality CMOS camera sensors typically used in current mobile phone cameras and this work, these last three types of noise can be assumed minimal and insignificant. Therefore, according to (3.11) and (3.12), when the ambient lighting condition is changing, $I(x,\lambda)$ varies as a function of $E(\lambda)$, this in turn induces a shift of the

histogram of image intensities (i.e. increase/decrease the amplitude of all pixels by certain magnitude) unless $Q_{i,j}(x,t)$ reaches Q_{sat}. Such influence is considered as the global illumination changes applied to print media in this work. But, because of camera saturation, this affects the captured surfaces to a certain extent [80]. After the pixel level analog-to-digital converter (ADC), the output of the CMOS image sensor $Q_{i,j}(x,t)$ is digitized to F which represents the digital image output. Theoretically, there is quantization loss in the process of ADC. However, such quantization loss becomes a minor influence because of the high precision ADC (more than 8 bit ADC [199]–[201]) used in contemporary digital cameras. Therefore, this influence is not considered.

The global illumination changes normally affect the performance of feature detectors and feature descriptors. The global illumination variation will influence the detected keypoints in terms of their position and number. It will also decrease the matching accuracy of feature descriptors (i.e. the correct feature matching pairs decrease with increasing distortion), although several local descriptors employ local illumination normalization to achieve robustness against illumination changes [98][78].

3.5.2 Effects of blurring during image shot

Image blurring can generally be categorized as: 1) out-of-focus blurring; and 2) motion blurring. The former blurring can be caused by user's amateur operation or autofocus of the camera while the latter blurring can be caused by camera and object motion. In a MAVS application, the positions of object and camera are relatively still or have minor motion as a user will point the mobile device camera to an image. The boundary of the image becomes an implicit constraint of capture area. Hence,

blurring caused by out-of-focus is the main consideration in this work. The image blurring can be modelled as [202], [203]:

$$G = H * F + N \tag{3.13}$$

where H is the blur function representing the above mentioned types of blur and where matrix G, F, N and * represent the blurred image, original image, noise and the convolution operation, respectively. For out-of-focus blurring, H can be approximated using a Gaussian kernel [203], [204]. The additive noise, N is typically considered to have a zero mean and white distribution and is orthogonal to the original image [204] and can hence be neglected. The image blurring has a great effect on image edges sharpness. In some cases, for a given set of images, the edge information will disappear and the sharpness will be mitigated. As a consequence, since the most common content-based local feature keypoint detectors used in computer vision and explored in this work rely on image edge information, the number and position of the keypoints detected by local feature detectors will be influenced and the feature descriptors will be affected as well.

3.5.3 Joint lighting variation and blurring distortion model

Based on the effects of lighting changes and image blurring as discussed in the previous two sections, when the global illumination changes $\Delta I(x, \lambda)$ during a certain exposure time, the response $i_{ph}(t)$ has corresponding changing $\Delta\tilde{\imath}_{ph}(t)$. Thus, the illumination changes can be described as:

$$\widetilde{Q_{\iota,J}}(x,t) = \int_0^t \left(\tilde{\imath}_{ph}(t) + i_{dc}\right) dt + \tilde{U}(t) + \tilde{V}(t) + C \leq Q_{sat}$$

$$= Q_{i,j}(x,t) + \int_0^t \Delta\tilde{\imath}_{ph}(t)dt + \Delta\tilde{U}(t) + \Delta\tilde{V}(t) \tag{3.14}$$

Where $\tilde{\imath}_{ph}(t) = i_{ph}(t) + \Delta\tilde{\imath}_{ph}(t)$, $\Delta\tilde{U}(t) = \tilde{U}(t) - U(t)$, $\Delta\tilde{V}(t) = \tilde{V}(t) - V(t)$.

Thus, the digital output of $\widetilde{Q_{\iota,J}}(x,t)$ is defined as:

$$\tilde{F} = F + \delta \tag{3.15}$$

Where the δ is the digital approximation of the summation of $\int_0^t \Delta\tilde{\imath}_{ph}(t)dt$, $\Delta\widetilde{U}(t)$ and $\Delta\tilde{V}(t)$.

When the out-of-focus blur occurs simultaneously, the joint effect of illumination changes and blurring can be modelled as [203], [204], [205]:

$$\tilde{G} = H * \tilde{F} = H * (F + \delta) \tag{3.16}$$

This equation is used to generate the distorted images in the experiments.

3.6 Matching methods based on the feature clustering

In this section, the matching method and evaluation criterion is introduced. Compared to the matching method in Section 3.2.2, the matching method based on clustering to find the closest matching image in a given dataset for a query image is employed in this work. The whole matching method can be divided into two steps. Firstly, a local feature cluster is found in a query image and then matched to a cluster derived from a given reference dataset to find corresponding matching pairs. Secondly, an image pair-wise cross-check matching and geometric verification as used in Section 3.2.2 are performed to validate the retrieved image. The detailed matching steps are described as follows.

3.6.1 Discovery of local feature clusters in query images by keypoint clustering

Keypoint clustering is motivated by the fact that the important keypoints of an object in an image have the tendency to cluster [172], [206], [207]. Such clustered keypoints are typically more useful for matching than isolated keypoints. In addition, isolated keypoints have a great possibility to be outlier keypoints because the feature

descriptors generated from these isolated keypoints often produce incorrect feature matching pairs. Thus, using keypoint clustering on the query image is helpful for removing these isolated keypoints resulting in improved matching accuracy as well as faster performance. The Mean Shift algorithm is used to find the keypoint clusters. Assuming a given query image, the n detected keypoints are $X = \{x_1, x_i, \cdots, x_n\}, x_i \in R^2$, where x_i is the coordinate of the i-th keypoint. Then the Mean Shift algorithm, which can be modeled as a nonparametric kernel density estimator [208], is performed on X to find the clusters:

$$f(x) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \tag{3.17}$$

where h is the bandwidth (i.e. kernel size), d is the dimension of variable (i.e. coordinate of keypoint, which is 2 in this work), $K(\cdot)$ is the Gaussian kernel function. In the experiment, the kernel size, h, is set to 20 pixels: 1) It is $1/20$ of the resolution of the processed image; 2) the histogram of detected region size is mainly distributed around 20 [78]. This process results in k clusters, and every x_i is indexed with corresponding closest cluster center $\hat{X} = \{\hat{x}_1, \hat{x}_2, \cdots, \hat{x}_k\}, \hat{x}_k \in R^2$ as judged by the mean squared error.

The keypoints around the cluster center have a large probability of belonging to the same object [1][17]. Therefore, a 2-D location histogram of X is used to study the neighboring relationship around each cluster centre as used in [160]. The image is divided into spatial bins and the number of keypoints is counted within each spatial bin. The bin containing the cluster center is also located. In the experiment, to provide sufficient keypoints for GV, the size of the bin is set to 4 pixels and the bin with less than 3 keypoints is filtered because 3 keypoints are the minimum requirement for GV [80]. Then, the adjacent bins around the bin containing the

cluster center are reserved. The keypoints within these bins are merged to form the final keypoint set $\widetilde{X}$, which is then used to generate local feature descriptors for feature matching. The size of $\widetilde{X}$ varies for different images.

3.6.2 Clustering in a reference image dataset and K-Nearest Neighbour (KNN) search

Recall the review in Section 2.7.4, comparing the query image against each image in a large image dataset using exhaustive pair-wise feature matching is time-consuming and unsuitable for MAVS applications. To speed up the feature matching, the Fast Library for Approximate Nearest Neighbors (FLANN) is employed to train the extracted local features from reference images dataset and then get the FLANN index using a KD-tree [120]. The KNN search is performed by using this index to find the nearest neighbor for each query descriptor. Here, K is set to 1 as only one corresponding feature matching pair is desired. After examining the feature matching pairs, the matched images are ordered by the number of feature matching pairs. The first 5 images are chosen for further pair-wise matching and geometric verification as the same in Section 3.2.2. The results are reported using precision @ 1.

3.6.3 Experimental image dataset construction

A subset of the Stanford Mobile Visual Search (MVS) dataset [209] was created containing 291 clean versions of reference images to be matched, which consist of 100 book cover images, 100 DVD cover images and 91 museum painting images (i.e. 2D flat surface print media). The images are captured from heterogeneous low and high-end camera phones under real world conditions. These camera-phone images contain rigid objects, widely varying lighting conditions, perspective distortion, foreground and background clutter, which reflect the real-world situation. More

detailed descriptions about the characteristic of the dataset can be found in [209]. The corresponding images captured by the Motorola Droid mobile phone camera (5 MP) under varying indoor lighting conditions are used to generate the query image dataset. However, the lighting condition will have more variation when outdoors. Therefore, the joint distortion model introduced in Section 3.5 was used to generate more distorted images to study the relative performance of different local feature algorithms from slight distortion to severe distortion. Although each image in the dataset was not taken under the same conditions, they still form a baseline set that can be adjusted using the proposed distortion model introduced in Section 3.5. A second distorted dataset was also created using the images captured by the higher quality Canon PowerShot G11 digital camera (10 MP) to compare with the Motorola Droid images.

The following distortions reflecting the joint blurring and lighting changes were added: 1) blur: images are filtered with Gaussian blur of kernel sizes ranging from 1 to 13. The interval is 2. Increasing size increases image blurring. 1 means no image blurring. If the image is blurred using kernel size more than 13, the detected keypoints do not change significantly or just disappear; 2) Illumination changes: 6 different levels of illumination changes ±50, ±100, and±128 are added to the pixel values of the images. These manipulations result in 28,518 different distorted query images.

3.7 Experimental results of various feature algorithms under joint optical distortions

The experiment is conducted on the aforementioned dataset by using the matching method introduced in the previous section. Four local feature algorithms are evaluated as shown in Table 3-2 that indicates different combinations of keypoint-descriptor pairings labelled here as AGASTFREAK, MSERSIFT, ORBORB, and SIFTSIFT. These feature detectors and descriptors are chosen because they produced good matching accuracy in the previous investigation. The number of features detected from images prior to clustering is set to at least 300, which has shown to ensure accurate matching results [169], [210]. Figure 3.9 shows the precision @ 1 of

Table 3-2 Summary of evaluated local feature algorithms

Keypoint detection	**Descriptor extraction**
AGAST	**FREAK**
MSER	**SIFT**
ORB	**ORB**
SIFT	**SIFT**

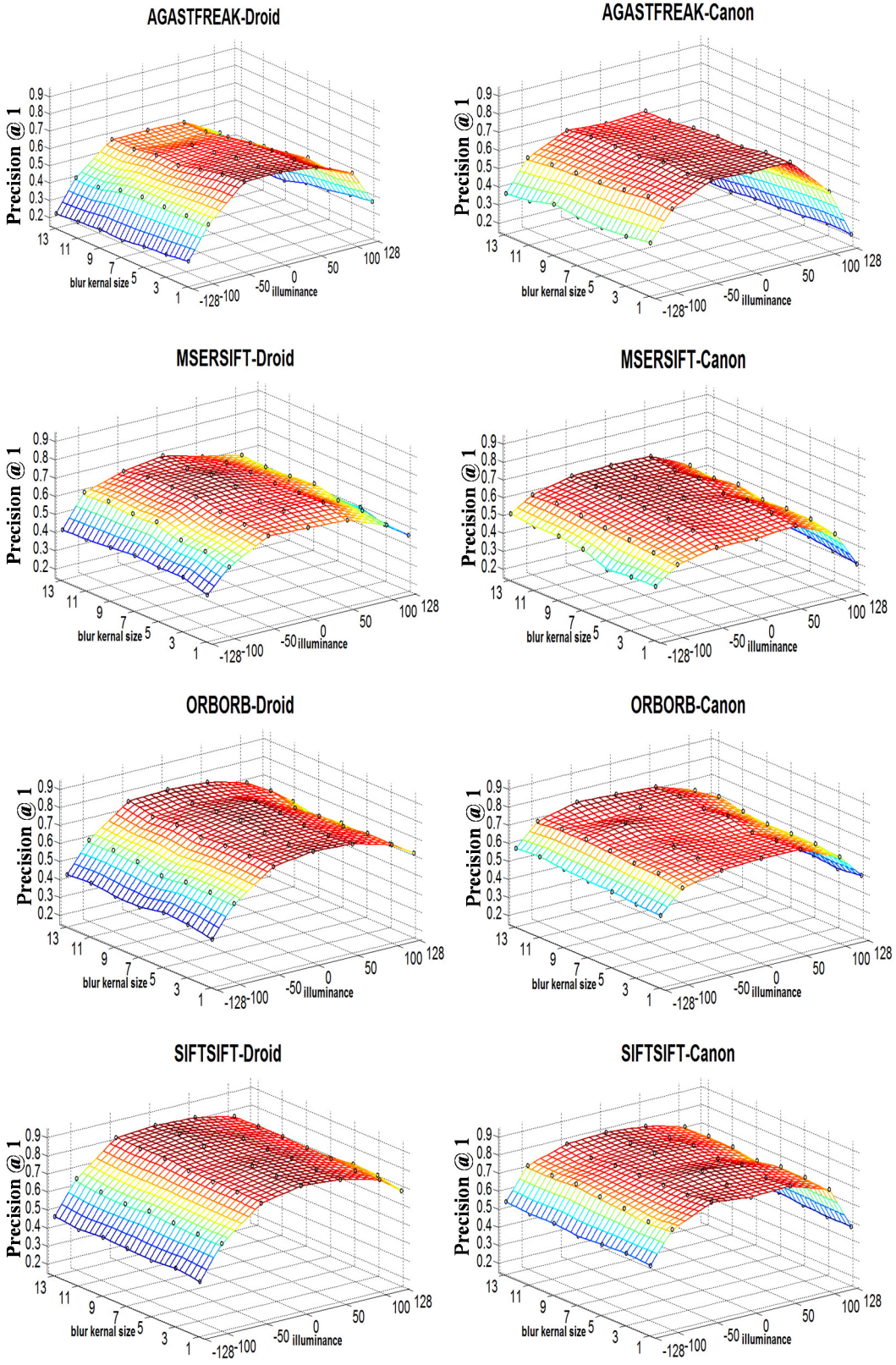


Figure 3.9 The precision @ 1 of various local feature algorithms under joint distortions.

Table 3-3 The values of the precision @ 1 of various local feature algorithms from slight distortion to severe distortion. I0: original illumination; I1, I2: slight and severe illumination increase; -I1, -I2: slight and severe illumination reduction. B1: no blurring; B2, B3: slight and severe blurring.

AGASTFREAK-Droid						AGASTFREAK-Canon					
I / B	-I2	-I1	I0	I1	I2	I / B	-I2	-I1	I0	I1	I2
B1	0.29	0.67	0.67	0.64	0.36	B1	0.44	0.70	0.69	0.69	0.21
B2	0.26	0.63	0.64	0.63	0.32	B2	0.41	0.69	0.67	0.66	0.20
B3	0.23	0.57	0.57	0.57	0.23	B3	0.37	0.63	0.62	0.63	0.16
MSERSIFT-Droid						MSERSIFT-Canon					
I / B	-I2	-I1	I0	I1	I2	I / B	-I2	-I1	I0	I1	I2
B1	0.40	0.64	0.63	0.62	0.45	B1	0.45	0.59	0.58	0.60	0.30
B2	0.44	0.66	0.67	0.66	0.5	B2	0.44	0.63	0.63	0.63	0.34
B3	0.42	0.66	0.69	0.63	0.47	B3	0.52	0.65	0.66	0.65	0.34
ORBORB-Droid						ORBORB-Canon					
I / B	-I2	-I1	I0	I1	I2	I / B	-I2	-I1	I0	I1	I2
B1	0.41	0.73	0.76	0.75	0.62	B1	0.55	0.72	0.73	0.73	0.49
B2	0.45	0.74	0.75	0.74	0.60	B2	0.57	0.74	0.71	0.70	0.49
B3	0.43	0.76	0.77	0.75	0.59	B3	0.58	0.76	0.75	0.73	0.47
SIFTSIFT-Droid						SIFTSIFT-Canon					
I / B	-I2	-I1	I0	I1	I2	I / B	-I2	-I1	I0	I1	I2
B1	0.44	0.80	0.84	0.82	0.67	B1	0.54	0.80	0.78	0.79	0.48
B2	0.45	0.81	0.84	0.83	0.71	B2	0.54	0.79	0.81	0.79	0.48
B3	0.46	0.82	0.84	0.83	0.68	B3	0.55	0.79	0.80	0.78	0.49

various local feature algorithms under joint illumination and blurring distortions across the entire query dataset. The values of precision @ 1 derived from slight distortion to severe distortion can be found in Table 3-3.

3.7.1 Influence of camera on precision @ 1 for local feature algorithms

The precision @ 1 of AGASTFREAK on query images captured by Motorola Droid and Canon has two major differences: 1) the images captured by the Droid camera have better precision @ 1 at the positive illumination change compared to negative change while the images captured by the Canon camera have an opposite trend; 2) the maximum precision @ 1 reduction with illumination change for Droid images is nearly 35% while it is nearly 49% for the Canon image; 3) the maximum precision @ 1 reduction with blurring changes for Droid images by approximately 13% while it is approximate only 7% for Canon images.

The precision @ 1 of MSERSIFT shows symmetrical distribution with the positive or negative shift of the illumination for Droid images while the Canon images result in better precision @ 1 at the negative illumination change compared to positive change. The maximum precision @ 1 reduction of MSERSIFT for Droid images is nearly 25% following illumination change while it is approximate 32% for Canon images. The precision @ 1 of MSERSIFT fluctuates with blurring change both for Droid images and Canon images. The maximum variation is nearly 8%. Blurring can slightly improve the precision @ 1 under certain degree of blurring in different illumination.

The precision @ 1 of SIFTSIFT and ORBORB is better at the positive illumination shifts compared to negative shifts for Motorola Droid while the images captured by Canon camera still show opposite trends. The maximum reduction on

precision @ 1 of ORBORB and SIFTSIFT is about 32% and 38%, respectively, for Droid images while it is 23% and 31%, respectively, for Canon images. Blurring has less effect on SIFTSIFT and ORBORB compared to illumination change where the matching accuracy only varies nearly 3%.

From the curve and the value of variation of precision @ 1, the joint distortions of illumination changes and blurring all have influences on the matching accuracy for the studied feature algorithms. Additionally, the illumination variation influences matching accuracy more than blurring, especially for severe distortion conditions. The severe darkness often occurs when the ambient light is too dark while the severe brightness often happens when overexposure or the ambient light is too bright. Therefore, image enhancement technologies may be employed in such situations or proper instruction (e.g. displaying color histogram) can be given to help users capture good quality images. In addition, different cameras have different optical performance which also influences the precision @ 1 of local feature algorithms under different joint distortions.

3.7.2 Influence of image type on precision @ 1 for local feature algorithms

Results in Figure 3.10 show the average precision @ 1 of different print media under different joint distortions. It shows that different local feature algorithms have varying matching accuracy for different print media under joint distortions and different cameras. AGASTFREAK achieves better precision @ 1 for book covers on the Canon images compared with Droid images. All four local feature algorithms have better performance for DVD covers in Droid images compared to Canon images. ORBORB has better precision @ 1 than SIFTSIFT for museum paintings in Canon images while ORBORB and SIFTSIFT have the similar performance for

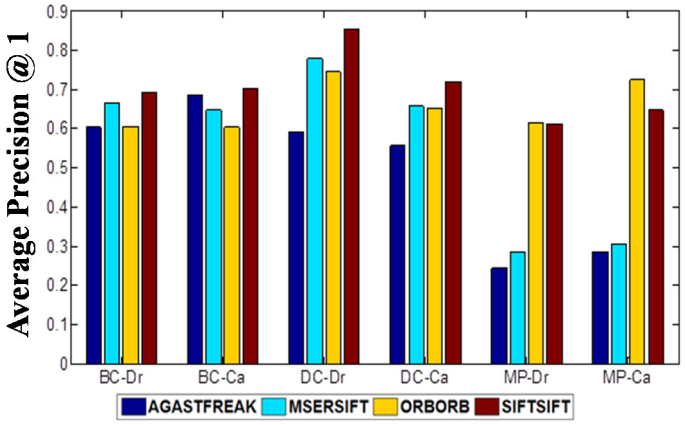


Figure 3.10 Average Precision @ 1 vs. different image types. "–Dr": Motorola Droid; "-Ca": Canon PowerShot G11; "BC": book cover; "DC": DVD cover; "MP": museum painting.

museum paintings in Droid images. MSERSIFT has slightly better accuracy for book covers and DVD covers than ORBORB, while ORBORB has much better precision @ 1 for museum paintings than MSERSIFT.

3.7.3 Conclusion

Evaluations have been done in this section to study the joint effect of illumination changes and image blurring in the context of a MAVS application. By examining the precision @ 1 of four local feature algorithms under various joint distortions with two different cameras, it has been found that illumination changes have more influence on matching accuracy compared to image blurring for the studied local feature algorithms under tested image datasets. Different cameras also influence the performance of local feature algorithms. Thus, flexible feature selection or combinations may be required to improve the precision @ 1 for a specific MAVS application within a heterogeneous camera phone environment. For example, if the MSERSIFT and ORBORB are chosen, it is better to use MSERSIFT for DVD covers and book covers while ORBORB is better for museum paintings.

3.8 Summary

An extensive evaluation of various combinations of state-of-art local feature detectors and descriptors are presented in this chapter. The performance in terms of precision @ 1 and processing time of various local feature algorithms are studied in the context of MAVS applications from the aspects of realistic distortions, including compression artifacts, optical and geometric distortions. The influences of three image compressors, and joint illumination and blurring changes are investigated on a practical print media image dataset captured by mobile device cameras. In the next Chapter, low bitrate and accurate MAVS systems based on these evaluations are presented.

4 ACCURATE AND LOW BIT RATE MAVS SYSTEM

4.1 Introduction

An extensive study of various local feature algorithms under realistic distortions has been presented in Chapter 3. Based on the evaluation, the SIFT feature algorithm presented superior performance compared to the other local feature algorithms. The SIFT feature algorithm is robust against various realistic compression artefacts as well as optical and geometric distortions. It achieved the best precision @ 1 in a dataset captured by a mobile device camera. However, the SIFT feature algorithm is the most time consuming local feature algorithm compared to alternatives such as SURF and ORB. To utilize the high discrimination and robustness of the SIFT feature algorithm in a MAVS application, how to achieve a low bitrate representation of the feature meanwhile achieving a high matching accuracy is the key. In addition, considering a heterogeneous mobile device environment with low-end mobile devices to high-end mobile devices, two low bitrate and accurate MAVS schemes are presented in this chapter for the purpose of adaptive MAVS system design.

One approach is to find the essential visual information which is significant for image matching from a query image and then transmit such information at a low bitrate to a server to perform feature detection, extraction and accurate image search and retrieval. Recall from the review in Section 2.9, to achieve fast and accurate content based image retrieval, the retrieval approaches in the compressed domain have been studied for decades. DCT or wavelet coefficients extracted from the compressed domain have been proven to be efficient for accurate image matching [32], [33], [211], [212]. In other words, these coefficients encompass the most significant visual information in an image. Considering that the JPEG coder is the

most popular image coder in the contemporary mobile devices, a fast and accurate low bit rate solution for an MAVS application is proposed based on extracting SIFT local features from images reconstructed from the low spatial frequency components in the DCT compressed domain.

The alternative approach is to select as few and robust features as possible such that the matching accuracy is invariant to distortions caused by camera capture whilst minimising the bit rate required for their transmission. Recall from the review in Section 2.8, the feature selection can be employed in an MAVS system to achieve such an aim. In this chapter, novel feature selection methods are proposed, based on the entropy of the image content, entropy of extracted features and the DCT coefficients to achieve good retrieval accuracy under low bit rate transmission.

The remainder of the chapter is organised as follows: Section 4.2 investigates the low bit rate transmission MAVS system using low frequency DCT coefficients, starting from the low frequency response of the SIFT features and then followed by the system design and the experimental result. Section 4.3 presents the low bit rate transmission MAVS system using relevance-based feature selection. Based on the modelling of feature selection in an MAVS application, three selection metrics are proposed and studied for different image dataset. Conclusions are drawn in Section 4.4.

4.2 Low bit rate transmission using low frequency DCT coefficients

4.2.1 Overview and novelty

When users scan across images with their mobile device camera to receive related content in an MAVS system, the system latency, which is defined as the total time between image capture on the device and matching result returned from the server,

must be minimized. To achieve such an aim, processing time on the mobile device in particular for low-end mobile devices and the transmission time to a server of the resulting image information must be minimized. While wireless networks have limited but improving bandwidth and mobile devices have increasing processing power, a limiting constraint is battery life and any reduction in the amount of data to be transmitted or processed locally is a benefit. In addition, as evaluated in Chapter 3, the optical distortion and geometric distortion occurred when capturing images on a mobile device camera by an amateur is an added problem, which will influence the matching accuracy of an MAVS application. Hence, the key challenge is how to minimize the transmission of visual information and reduce processing in the mobile device whilst achieving real-time and highly accurate retrieval against distortion. This has previously been shown to be essential to maximizing user Quality of Experience (QoE) for such applications [169], [170], [172], [173].

Aiming for low computation and fast processing, DCT and wavelet coefficients have been proposed for compressed domain image retrieval for decades. Such methods were applicable for duplicated image retrieval. However, they have not achieved high matching accuracy under optical and geometric distortion as reviewed in Section 2.9 [174]–[177]. Although low level feature extraction based on DCT and wavelet coefficients in the spatial domain were proposed to improve retrieval rate and matching accuracy, the accuracy of such methods is still degraded under joint optical distortion and geometric distortion [102], [178]. Alternatively, recall from the review in Section 2.6 and Section 2.9, due to the development of repeatable, distinctive, and discriminative high quality local features, local feature based image content retrieval has gained a lot of attention[78], [79], [98]. The most popular and proven high discriminative features are histogram-based features, such

as SIFT [80]. But, the main drawback of using SIFT on mobile devices is high dimensionality resulting in complex computations for feature matching and a significant amount of bandwidth for feature transmission (often more than the compressed JPEG format of the image [172]). To reduce the dimensionality and alleviate the transmission issue, the SIFT compression schemes including hashing, transform coding and vector quantization have attracted much attention [98], [179], [180]. Such methods normally compromise between bit-rate and matching accuracy. Recently, CHOG achieved low bit rate transmission by directly compressing the gradient histogram while maintaining high matching accuracy [100]. However, extra computational complexity and power consumption are required for feature detection and extraction. If such processing was performed on a low-end mobile device, it will cause significant delay and result in decreased QoE for the users [172], [173]. Especially, in practice, more features are required to maintain high accuracy against real world distortion, which further increases the demand of resources and delay for computation and transmission.

Two observations derived from the evaluation in Chapters 2 and 3 are: 1) some feature algorithms still achieve high matching accuracy on highly compressed JPEG images [213] ; 2) the nature of dimensionality reduction or feature compression is to discard unnecessary redundancies while preserving spatial information essential for accurate feature matching. Hence, it is hypothesized that the most compact spatial information in the DCT domain can be found and such information can be encoded and transmitted at a much lower bit rate than the original image to a sever or cloud to achieve high matching accuracy by using highly discriminative features meanwhile solving aforementioned challenge.

To verify this hypothesis, firstly, the SIFT feature associated with the loss of spatial frequency information in the DCT domain is investigated. Secondly, a new low bit rate, low complexity, low latency architecture with high accuracy for MAVS applications is proposed. The proposed system uses SIFT features extracted from images reconstructed from low spatial frequency components, which are represented by the encoded block based 2D DCT coefficients. Thirdly, the robustness of this approach is studied for various image distortions under controlled experimental environments, including additive white Gaussian, global illumination changes, out-of-focus blur, rotation and scaling, which commonly occur in practice. In the next section, the sensitivity of SIFT features to spatial frequency is analyzed.

4.2.2 Analysis of the relationship between SIFT features and DCT coefficients

The SIFT feature is detected in the Gaussian scale invariant space by finding the Difference-of-Gaussian [80]:

$$D(x, y, \sigma) = \big(G(x, y, k\sigma) - G(x, y, \sigma)\big) * I(x, y) \qquad (4.1)$$

where $D(x, y, \sigma)$ is the extrema found by the difference in the Gaussian scale-space; $G(x, y, \sigma)$ is a Gaussian function; k is constant multiplicative factor to generate different Gaussian scales; * is the convolution operation; and $I(x, y)$ is the input image. To find the most stable and repeatable keypoints under different image distortions [80], the image is smoothed by a Gaussian kernel σ which is equivalent to strongly attenuating all but the lower spatial image frequencies in the DCT domain using a low-pass filter. After finding the keypoint, the SIFT descriptor is formed by accumulating the gradient orientation histograms and magnitudes within an image patch around a keypoint. The gradient magnitude $m(x, y)$ and orientation $\theta(x, y)$ in the SIFT descriptor is computed using pixel differences:

$$m(x,y) = \sqrt{\begin{matrix}(L(x+1,y) - L(x-1,y))^2 \\ +\big(L(x,y+1) - L(x,y-1)\big)^2\end{matrix}} \quad (4.2)$$

$$\theta(x,y) = tan^{-1}\left(\frac{L(x,y+1)-L(x,y-1)}{L(x+1,y)-L(x-1,y)}\right) \quad (4.3)$$

L is Gaussian smoothed image which is defined as:

$$L(x,y,\sigma) = G(x,y,\sigma) * I(x,y) \quad (4.4)$$

Before calculating the $m(x,y)$ and $\theta(x,y)$, the high image spatial frequencies are filtered out. Therefore, it can be inferred that the low spatial frequencies are most important for constructing the SIFT feature. Since the value of σ typically chosen in the SIFT algorithm is 1.6 (i.e. cutoff frequency Ω_c of Gaussian filter is 0.117) which offers good matching accuracy and computation efficiency [80], the discrimination of SIFT features is mainly determined by spatial frequencies which are below 0.117.

Considering the captured image is stored in the JPEG file format, the image is decomposed into a set of coefficients by 2D DCT transformation. These DCT coefficients represent different image spatial frequencies. Thus, the question is: "What is the most essential DCT coefficients required to achieve high matching accuracy when using the SIFT feature?" As mentioned in the studies of [98], [213], SIFT features still achieved high matching accuracy under high JPEG compression. Thus, it can be inferred that some spatial frequency components in the DCT domain are imperceptible for the SIFT descriptor. Using the 2D DCT, considering the standard 8-by-8 DCT block in JPEG, an image $I(x,y)$ is represented as [214]:

$$\mathrm{I}(x,y) = \sum_{u=0}^{7}\sum_{v=0}^{7} C(u,v)F(u,v) \quad (4.5)$$

Thus, (4.4) can be rewritten as:

$$\hat{\mathrm{L}}(\mathrm{x},\mathrm{y},\sigma) = \sum_{\mathrm{u}=0}^{7}\sum_{\mathrm{v}=0}^{7} \mathrm{C}(\mathrm{u},\mathrm{v})(\mathrm{G}(\mathrm{x},\mathrm{y},\sigma) * \mathrm{F}(\mathrm{u},\mathrm{v})) \quad (4.6)$$

$C(u,v)$ are the DCT coefficients, $F(u,v)$ are the DCT basis functions. It is noted that $\hat{L}(x,y,\sigma)$ is different from $L(x,y,\sigma)$ due to the use of a block-based 2D-DCT

resulting in spectral leakage caused by the finite length filter window and spectral subsampling. The $F(u,v)$ is smoothed by a low-pass Gaussian filter. The high frequency DCT coefficients beyond the cut-off frequency of the Gaussian filter Ω_c are discarded before calculating the gradient magnitude $m(x,y)$ and orientation $\theta(x,y)$. Therefore, the gradient magnitude and orientation of SIFT descriptors are mainly determined by low frequency DCT coefficients.

4.2.3 Evaluating the spatial frequency sensitivity of SIFT

This section evaluates the minimum spatial frequency information in the DCT domain for the SIFT feature to achieve high matching accuracy. Edge information causes sharp variance along a certain direction and has the most effect on the values of $m(x,y)$ and $\theta(x,y)$. Moreover, edge information is primarily represented by the DC component and the first 3 to 8 AC components in the 8-by-8 block DCT [23][24]. Hence, DCT coefficients beyond 8 are not analysed as they were found to be near zero and are expected to have minor contribution to the gradient magnitude $m(x,y)$ and orientation $\theta(x,y)$.

The algorithm of [80] is used to extract the SIFT features. The image dataset used in the experiment comprises 316 uncompressed query images from the clean version images of Tampere dataset [215], CSIQ dataset [216] and UCID dataset [217] and 1397 reference images including corresponding reference images and interference images, which cover a wide range of image complexity.

Evaluating the matching accuracy of SIFT features associated with different spatial frequency information in the DCT domain is conducted as follows:

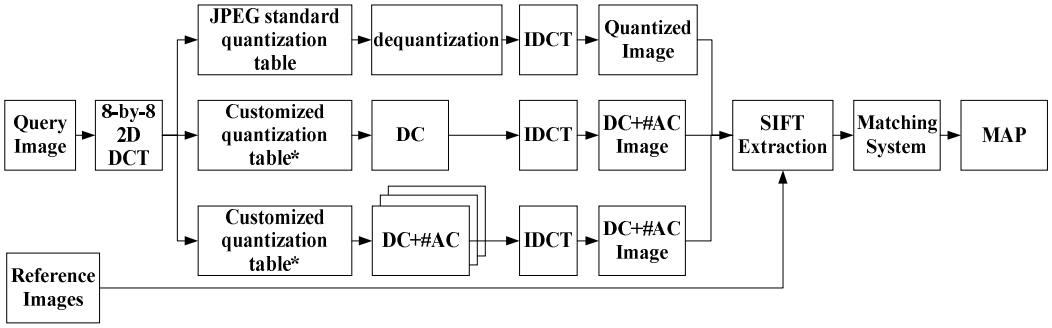


Figure 4.1 The architecture of studying SIFT feature associated with varying spatial information.

1) The DCT coefficients are extracted from the query images and then a dataset of query information with different spatial frequency components is created using different combinations of DCT coefficients as shown in Figure 4.1. Getting DCT values of DC, DC+#AC is equal to quantize the DCT values with a customized quantization table. For example, the quantization table for extracting DC is [1 0 0 0 0 0 0 0; 0 0 0 0 0 0 0 0; 0 0 0 0 0 0 0 0; 0 0 0 0 0 0 0 0; 0 0 0 0 0 0 0 0; 0 0 0 0 0 0 0 0; 0 0 0 0 0 0 0 0; 0 0 0 0 0 0 0 0];

2) This query information is encoded and transmitted to the server-side (i.e. quantization, DPCM and entropy coding as used in JPEG compressor is used in the encoder to generate the query information for transmission.);

3) The SIFT features are extracted from these decoded images to be used within the feature matching system;

4) The matching system employs the Fast Library of Approximate Nearest Neighbours Search to build KD-Tree index from the SIFT features extracted from reference images and then KNN (K=1) search, cross check matching and geometric verification as reviewed in Section 2.7 are used to perform pair-wise image matching [123], [126], [218]–[222];

5) The matching accuracy is then evaluated using the precision @ 1 of (2.11).

The evaluation results of the spatial frequency sensitivity of SIFT feature is presented in this section with the comparison results of using the CHOG feature and LPDF feature within the same database and matching algorithm introduced in the previous section.

Four parameters, including coordination of the feature, the angle of the feature and L2-norm between the feature and corresponding reference feature are studied as these parameters are significant for the matching system. The angle as shown in (4.2) is a vital parameter when calculating the feature descriptor. If the angle changes significantly with the reduction of the high frequency DCT coefficients, the feature descriptor would also change significantly and result in a totally different feature descriptor which was supposed to be able to match with the reference feature. The location (x, y) is important for geometric verification. If the location varies too much, the accuracy of geometric verification would be degraded. The L2-norm measures the distance between the query feature and the reference feature. If a large variance is found in the L2-norm of the feature, there would be high possibility to filter out the feature mistakenly in the matching system. Table 4-1

Table 4-1 Changes in location, orientation and L2-norm of SIFT features associated with loss of DCT coefficient and increase of quantization. (DC component plus different AC values; Q5, Q10 indicate 5, 10 quantization values respectively).

	x	**y**	**Angle(°)**	**L2**
uncompressed	50.32	356.77	232.89	NA
DC	49.99	357.05	233.11	19.053
DC2AC	50.36	356.81	232.89	9.899
DC3AC	50.37	356.98	233.03	8.485
DC8AC	50.34	356.80	232.93	7.28
Q5	51.06	357.25	232.36	41.869
Q10	51.10	357.19	232.10	20.688

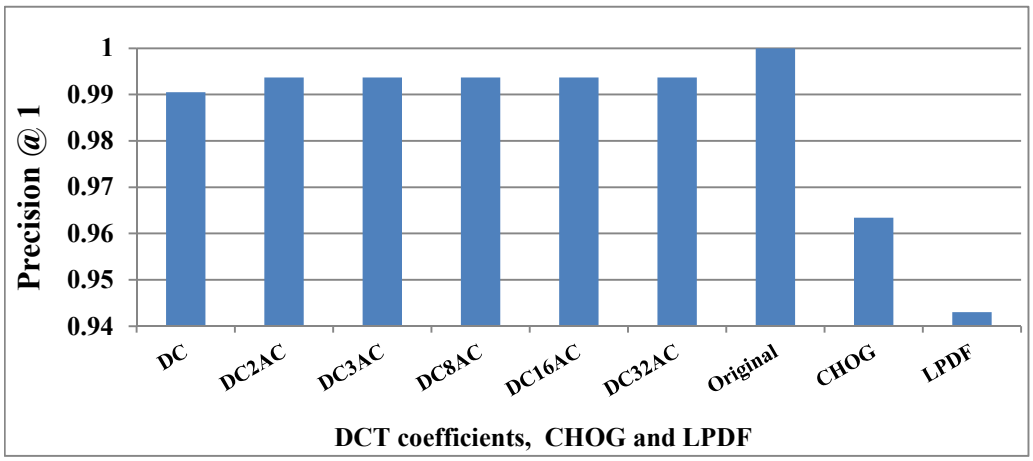


Figure 4.2 The precision @ 1 of SIFT feature associated with different spatial information and precision @ 1 of CHOG and LPDF feature; DC: the query image only contains DC coefficient; DC2AC: the query image contains DC coefficient and first 2 AC coefficients. DC3AC, DC8AC, DC16AC and DC32AC have the similar definition (i.e. DC with 3, 8, 16, 32 AC coefficients, respectively. Original: original image.

shows a typical example from the SIFT descriptor values set extracted from the image dataset which shows the influence of filtering out high frequency DCT coefficients. With the loss of high frequency DCT coefficients, the location (x, y) and angle has slightly changes while the L2-norm increases. It is observed that the high frequency AC values only have minor influence on these parameters thus have little effect on matching accuracy. The result for the impact of quantization of the DC components (as typically used in JPEG compression) is also investigated and shown in Table 4-1. With increasing quantization, the difference of the features extracted from quantized DC frequency components becomes more apparent and may result in a false feature matching. A distance of 19 for DC is acceptable as it achieves 96% matching accuracy as shown in Figure 4.2 which matches the distance of 20 for Q10 with 95% matching accuracy as evaluated in previous chapter [213].

The matching accuracy as measured by the precision @ 1 of SIFT features associated with filtering out different DCT coefficients is presented in Figure 4.2 as

well as the precision @ 1 of CHOG and LPDF features. Results indicate the DC component has the major effect on matching accuracy when using the SIFT feature. Although AC components reflect spatial variation information, such information has a minor impact on the SIFT feature matching accuracy. This result complies with the discussion in Section 4.2.2. These results are comparable to precision @ 1 results obtained for the same database using CHOG and LPDF features, which achieved matching accuracy for precision @ 1 of approximately 96% and 94%, respectively. It is proposed that the weaker performance of LPDF may be due to similar texture patterns appearing in non-identical query images, which lead to similar DCT coefficients.

4.3 Proposed low bitrate MAVS system using low spatial frequency DCT coefficients under realistic distortions

Motivated by the results in Section 4.2.3, this section proposes and evaluates a low bit rate, low complexity, low latency MAVS system architecture which encodes only the DC component in the query information and reference images for image matching. The proposed system diagram is shown in Figure 4.3. To evaluate the performance of the proposed architecture under realistic varying image distortions, different image distortions are applied to the query images in the dataset described in Section 4.2.3. The matching method is the same as used in Section 4.2.3. The corresponding results of the precision @ 1 measured for the proposed system under various realistic distortions are presented in the following subsection. The comparison results of using all spatial frequencies of the image, CHOG features and LPDF features are also presented and discussed.

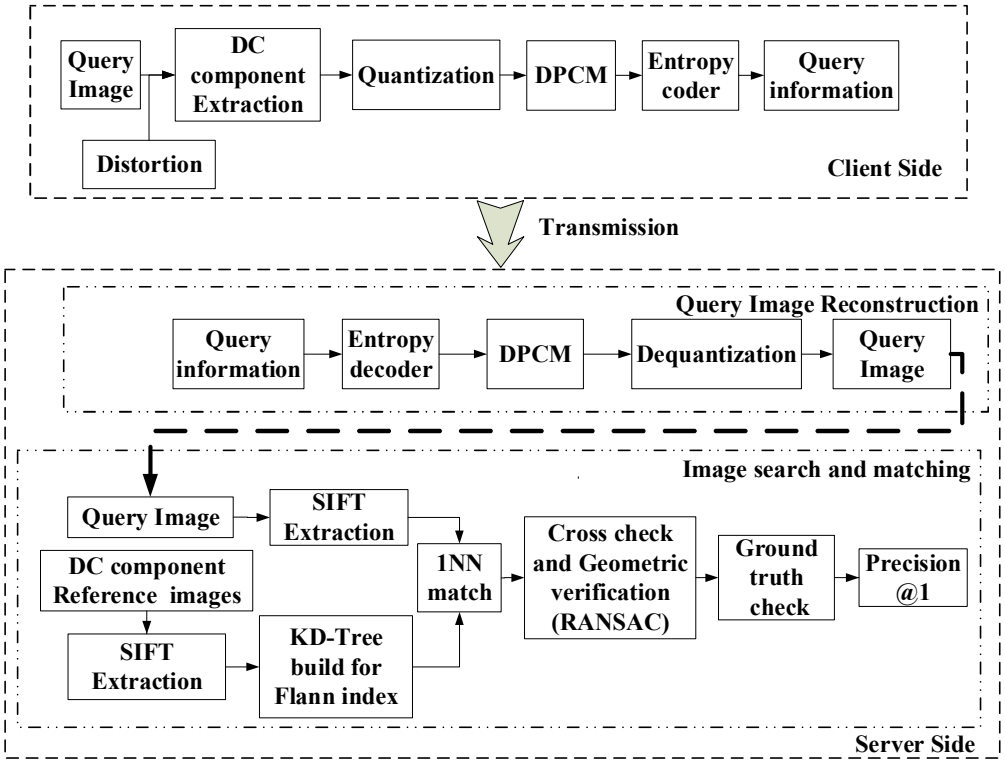


Figure 4.3 Proposed MAVS architecture for using only DC coefficient under distortion including Additive white Gaussian noise, Global illumination change, Out-of-focus blur, rotation and scaling.

Several typical distortions typically caused by the photographic environment and amateur camera operation are examined in this section, including Additive White Gaussian Noise (AWGN), global illumination change, out-of-focus blur, rotation and scaling as shown in Table 4-2. The precision @ 1 results are shown in Figure 4.4 to Figure 4.8, where "Ori" indicates the precision @ 1 of using original image (i.e. all

Table 4-2 Different image distortions applied to query images

Distortion Type	Distortion parameters
AWNG	Mean = 0 variance $\mathbf{V_\sigma}$ = 0.1, 0.2, 0.4, 0.6, 0.8, 1
Global illumination change	Illumination shift = -128, -110, -80, -60, -30, -5
Out-of-focus blurring	Sigma = 0.8, 1.4, 2, 2.3, 2.9, 3.2
Rotation	Angle = 15°, 30°, 45°, 60°, 75°, 90°
Scaling	Factor = 1/2, 1/4, 1/8, 1/16

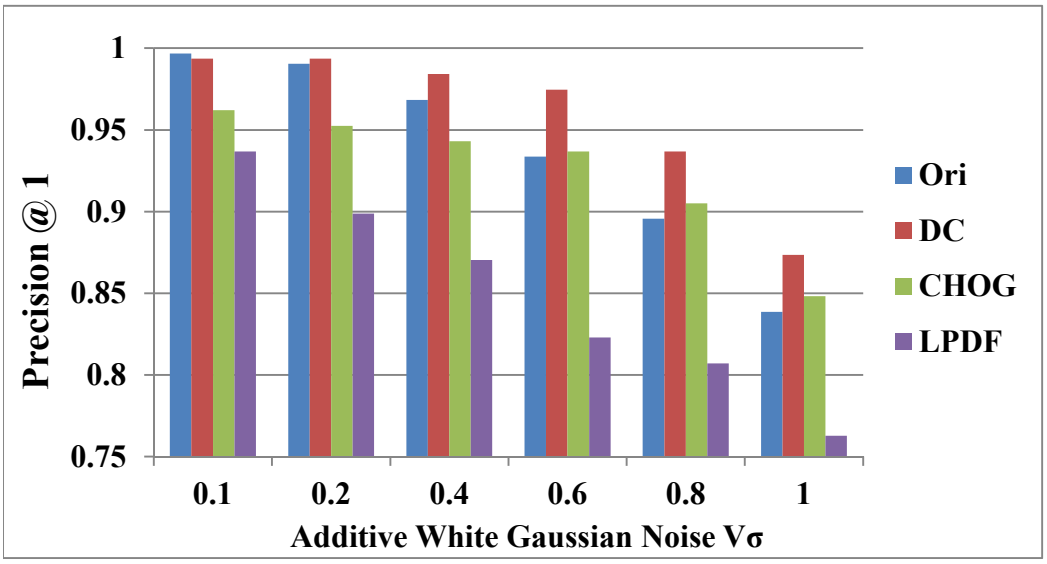


Figure 4.4 The precision @ 1results under varying AWGN

spatial frequencies are used) and SIFT feature for matching; "DC" indicates precision @ 1 of proposed system when using only the DC component; "DC3AC" indicates precision @ 1 of proposed system when using the DC and the first 3 AC components; "DC8AC" indicates precision @ 1 of proposed system when using the DC and the first 8 AC components; "CHOG" indicates the precision @ 1 when using the original image and CHOG features for matching; "LPDF" indicates the precision @ 1 when using the original image and LPDF features for matching.

4.3.1 Experimental results of proposed system under the distortion of AWGN

Additive white Gaussian noise is used to simulate environmental thermal noise (e.g. arising from shot noise, warm objects or the sun) during image acquisition [215]. The additive white Gaussian noise with zero mean and variance V_σ is added to the query image before extracting the DC component and extracting the SIFT, CHOG and LPDF features. Precision @ 1 results under varying AWGN are shown in Figure 4.4. By using the image reconstructed from only the DC component, the precision @ 1 is approximately 5% better than CHOG for noise variances of 0.4 or less. Results show

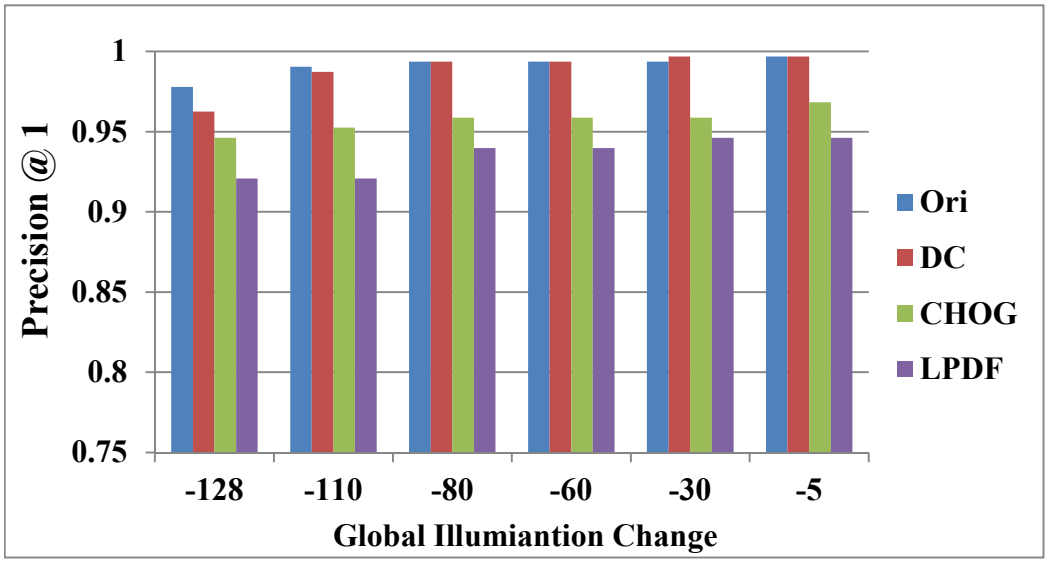


Figure 4.5 The precision @ 1results under varying global illumination change

the proposed method is robust to strong thermal noise (variances greater than 0.4), with precision @ 1 values significantly better than those for SIFT extracted from the original noisy image.

4.3.2 Experimental results of proposed system under the distortion of global illumination change

Global illumination changes can cause the illumination shift on the shot image, which impacts on feature matching. This is simulated using a brightness shift to the negative direction in the RGB channel of the query image before extracting the DC component and local features. The precision @ 1 results under illumination change are shown in Figure 4.5. Due to the illumination normalization employed in the SIFT feature, the illumination change has a minor impact on the matching accuracy of SIFT feature, with the DC component image achieving matching accuracies equal to or better than the original image for all but the two largest illumination changes. In addition, the performance of proposed method is superior to CHOG and LPDF feature by approximately on average more than 3% and 5%, respectively.

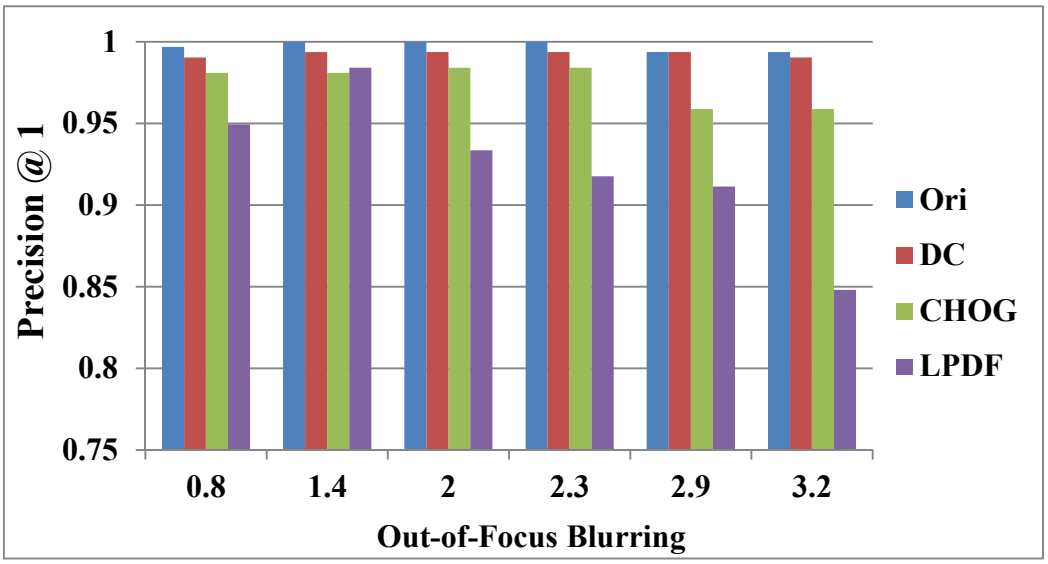


Figure 4.6 The precision @ 1 results under varying out-of-focus blurring

4.3.3 Experimental results of proposed system under the distortion of out-of-focus blurring

Out-of-focus blur can be modelled as the convolution of the image with a Gaussian kernel [202], [203]. Precision @ 1 results under varying out-of-focus blurring amounts are shown in Figure 4.6. The matching accuracy when using the DC component image is comparable to those obtained using SIFT extracted from the original image and superior to CHOG and LPDF for large out-of-focus blurring (above 2.3).

4.3.4 Experimental results of proposed system under the distortion of rotation

Another problem for MAVS applications is the rotation caused by arbitrary position of the camera. Figure 4.7 shows the precision @ 1 of using the DC component with varying rotations applied to the query images (reference images have original rotation). The results show that the SIFT feature extracted from the DC component image still achieves more than 97% matching accuracy across tested rotations from 0 to 90 degrees and achieves much better accuracy than LPDF.

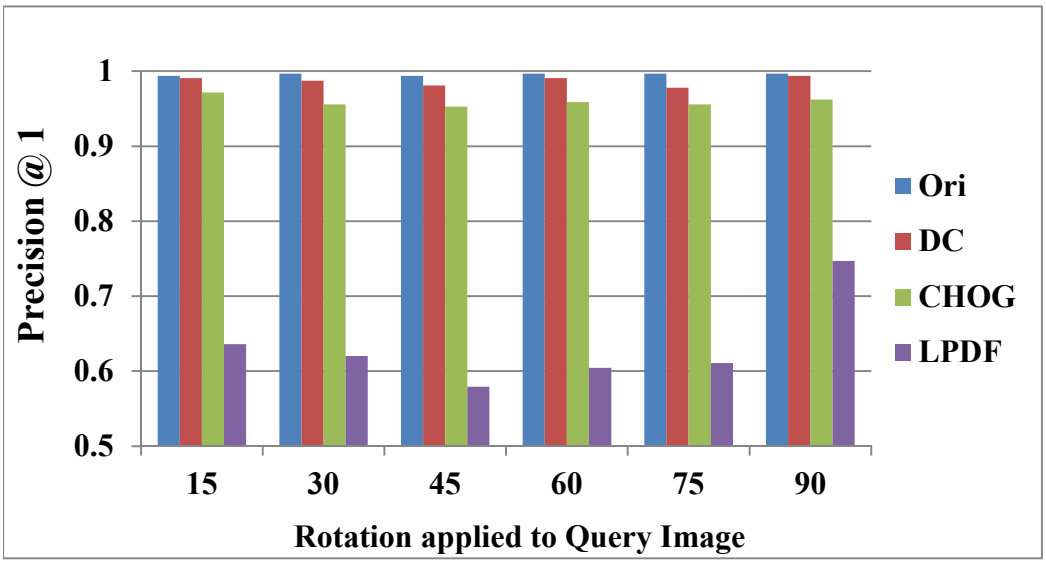


Figure 4.7 The precision @ 1results under varying image rotation

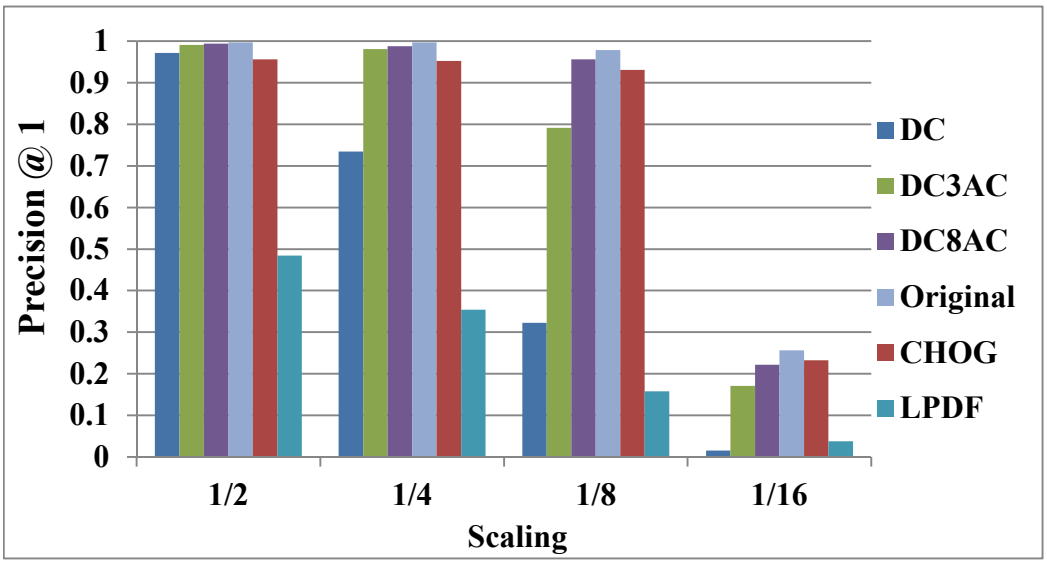
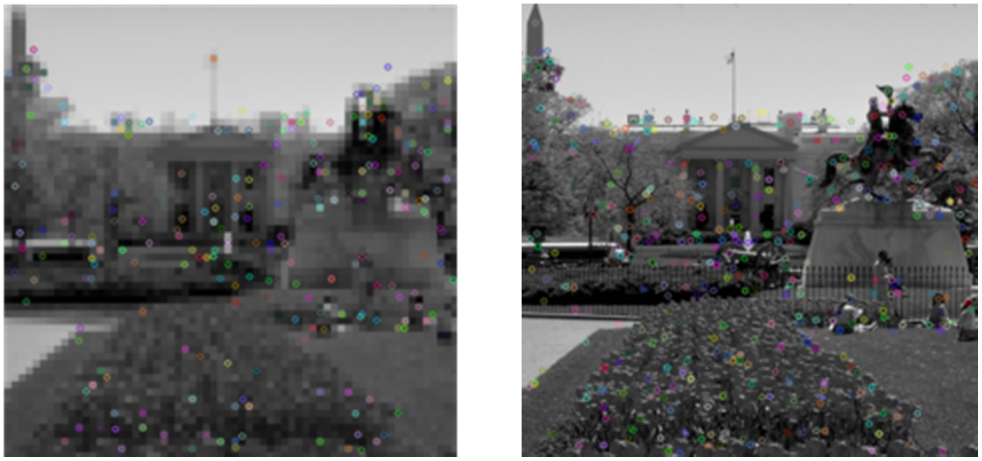


Figure 4.8 The precision @ 1 results under varying image scaling

4.3.5 Experimental results of proposed system under the distortion of scaling

The original images are scaled to low resolution images by using image pyramid reduction. There are two potential benefits to employ image scaling in an MAVS system: 1) significantly reduce the transmission bit rate; 2) results in fewer features, hence reducing the computational complexity of feature extraction. Precision @ 1



(A) Reconstructed DC image (B) Original image

Figure 4.9 An example of reconstructed image compared to original image.

results under different scales are shown in Figure 4.8. Results show that scaling the low frequency images with only DC and the first few AC components to half scale doesn't result in a significant decrease in matching accuracy because of the good scale-invariant character of SIFT. The low frequency images with the first 3 AC components or the first 8 AC components achieve better matching accuracy than images only containing DC component in particular when severely reducing the image resolution below 1/4. It can be inferred that the DCT AC coefficients become more important for improving matching accuracy under low resolution as they provide more fine spatial information. Compared with the full scale results of Figure 4.2, the precision @ 1 results reduce to approximately 97% and 92% for half scale DC image and CHOG feature transmission, respectively.

4.3.6 Bandwidth saving and system latency reduction

The bandwidth saving of using DC image comparing to compressed CHOG features is shown in Figure 4.10. Compared to transmitting the original image, the bandwidth required to transmit the DC image in the proposed system is reduced by 97% and 99%, while maintaining precision @ 1 of 99% and 97%, for full scale and half scale,

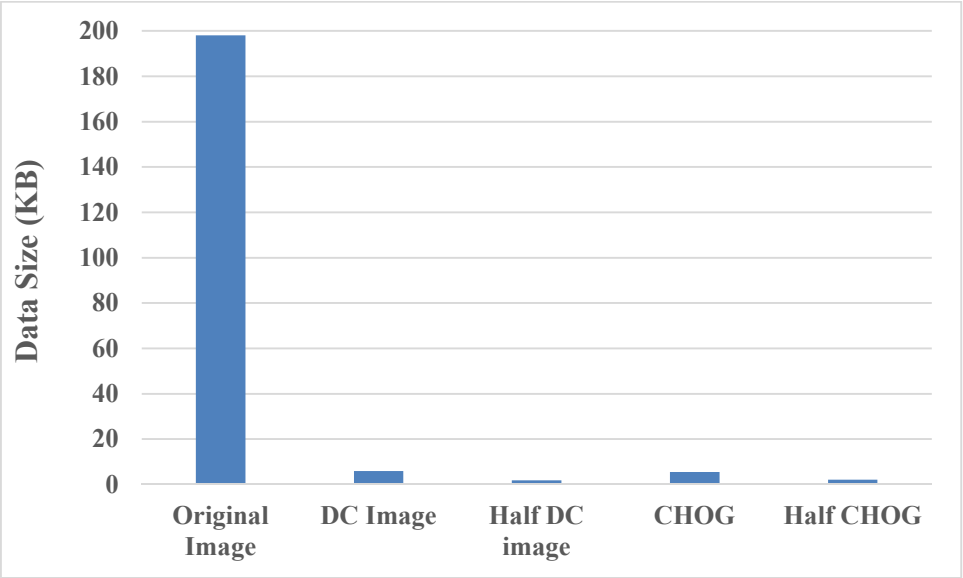


Figure 4.10 The bandwidth saving of using proposed method compared to compressed CHOG feature.

respectively. An example of a reconstructed DC image compared to the original image is shown in Figure 4.9. These results are comparable with transmission of the low bit rate CHOG features, which reduce the bandwidth by 97% and 99% for full scale and half scale images, respectively, while the precision @ 1 of 96% and 92%, respectively. LPDF requires more than 10 times the bandwidth compared to transmitting the DC image. Direct transmission of SIFT features requires an even larger data rate than the original image [172].

For system latency comparison, reference implementations of JPEG, CHOG [100], LPDF [102] and OpenCV implementations of SIFT and the matching system on an Intel i7 processor with a 2.9GHz processor were used in the simulations (client side processing time would be higher on a typical mobile device). For the proposed system, client side processing was approximately 63ms (i.e. including the DC extraction and compression time) compared to 120ms (i.e. including the CHOG feature detection, extraction and compression time) for a CHOG-based system, 89ms for detecting and extracting SIFT features (i.e. including SIFT feature detection and extraction time) and 182ms (i.e. LPDF detection and extraction time) for LPDF

extraction. Hence, the proposed system achieves an approximate 50% reduction in processing time compared to an existing low-bit rate system using CHOG feature [100]. Server side processing time is similar for all systems due to using the same matching algorithms.

4.3.7 Conclusion

A new MAVS system is proposed based on SIFT features derived from DC coefficients in the 2D block-based DCT domain to enable a low complexity, low latency and accurate implementation on mobile devices whilst requiring low bit rate transmission by using a powerful remote server to complete the most time-consuming processing. The method achieved more than 97% precision @ 1 matching accuracy while reducing the transmission bandwidth requirement by more than 97%, whilst reducing client side processing time by approximately 50% compared to an existing low-bit rate CHOG feature matching system. The DC image can be reduced to half scale to further reduce transmission bandwidth under poor transmission situations without obvious loss of matching accuracy. Alternatively, the JPEG encoder can use a customized quantization table which discards the AC coefficients, such that there is no need to modify current image codecs in mobile devices and can be easily deployed in a low-end mobile device. There is also a limitation for the proposed method, in that it is not applicable to images which have similar low frequency DCT coefficients. Therefore, when constructing the pre-defined image dataset in the server, images with similar contents and structures should be avoided.

4.4 Low bitrate transmission using feature selection

4.4.1 Overview and novelty

Recall the review in Section 2.8, due to the richness of the captured image scene (e.g. complex visual objects in a scene), capture distortions (e.g. viewpoint and scale changes, environmental lighting variation, shadow, etc.), and foreground and background clutter, hundreds of local features can be detected and extracted in a rich content image. In addition, not all the features are necessary for matching, especially features leading to false positive matches due to distortions. Therefore, it is not the optimal way to transmit and perform searching and matching by using all the detected features, otherwise huge computation and larger transmission bandwidth are required in the MAVS applications. To tackle the problem, one solution is to employ feature selection technologies to select the distinctive visual features as few as possible. In addition, the selected distinctive features should be robust to realistic distortions and have adequate characteristics to perform highly accurate similarity visual matching to find the predefined corresponding multimedia content in remote server or local repository. However, due to the richness of the captured image scene (e.g. complex visual objects in a scene), capture distortions (e.g. viewpoint and scale changes, environmental lighting variation, shadow, etc.), and foreground and background clutter, makes it difficult to extract the most discriminative features from the captured camera scenes. The criterion of feature selection is crucial and inappropriate feature selection would degrade the matching accuracy dramatically.

By extending the feature selection work in MPEG-7 CDVS to tackle the feature selection problem of MAVS applications for low bit rate transmission and high matching accuracy, novel SIFT feature selection methods are proposed based on three new metrics: 1) the entropy information of the image content in the keypoint

domain; 2) the entropy information of the feature descriptor in the descriptor domain; and 3) the Discrete Cosine Transformation (DCT) coefficients in the compressed domain. The proposed approaches are suggested as efficient methods for selecting the most significant and robust features in terms of their ability to result in accurate matching within a MAVS system under different bit-rate constraints and realistic complex capturing distortions. The details of the proposed feature selection methods are provided in the following section.

4.4.2 Methodology of Proposed Feature Selection Method

The proposed feature selection method is also a relevance-based feature selection method. Compared to the selection methods as reviewed in Section 2.8, the hypothesis is that a more efficient relevance-based method can be found to utilize not only the output parameters of a feature detector but also the implicit information embedded in the local image patch and feature descriptor to select the most significant and robust features, which are always necessary for correctly matching meanwhile minimizing the transmission bit-rate requirement.

In this section, the problem of selecting the key features for matching a captured frame to a reference image in a MAVS application is firstly defined and then the learning procedure to get a posterior probability of a relevance parameter for local feature selection is introduced. Finally, the proposed feature selection method based on the entropy information of the image content in the keypoint domain and SIFT features in the descriptor domain as well as the DCT coefficients in the compressed domain are proposed.

The problem of selecting the key features extracted from a captured image to match a predefined image in a remote database containing N candidate images can be

formulated as follows (the images used in this work are grayscale images for content-based image matching):

1) Assume the captured image is represented by the feature set matrix $X = \{x_1, x_2 \dots x_L\}$, where $x_i \in R^m$ is a feature vector of length m; the N candidate images in the database are represented by the feature set matrix $\{Y_1, Y_2 \dots Y_N\}$, $Y_i = \{y_1, y_2 \dots y_K\}$, where $y_j \in R^m$ is a feature vector of length m as well;

2) Assume that the probabilities of the captured image being correctly matched to each candidate are $H = (h_1, h_2 \dots h_N)$; $h_i = f(X, Y_i)$ where $f(\cdot)$ measures the similarity between X and Y_i;

3) If the *j*-th candidate in the database corresponds to X, the objective is to find a proper relevance metric θ associated with h_i to select the key features that make $P(X|h_j) > P(X|h_i), where\ i \epsilon [1, N], i \neq j$, for example, $\theta_{peak}, \theta_{orientation}, \theta_{scale}, \theta_{distance}$ as used in [153], [154], [167].

The key stage of the proposed method is to learn a posterior probability of a relevance metric θ to measure how well a feature can be correctly matched from the dataset, which is denoted here as 'matchability' of a feature. On the basis of the learning method in [153], [154], [167], the K Nearest Neighbour search and cross-check method [119] which achieved better performance for finding true positive matched features in the previous work are incorporated into the learning process. The procedure to learn the 'matchability' is summarised as follows:

For all the features extracted from the candidate images of the dataset, the relevance metric θ is calculated and assigned for each feature, respectively. Then, the correctly matched features are learnt from the supervised pair-wise image matching, which is performed on image pairs according to the ground-truth image list. The

images in a pair are the distorted image and corresponding ground-truth image. Each image pair undergoes the following process:

1) Detect features from both images in a pair. For each feature, a relevance metric θ is computed and recorded for each feature;

2) Perform the Nearest Neighbor search (i.e. KNN search where k=1 [123]) within each image pair to find the nearest neighbour for each feature. In contrast to using the distance ratio test [153], [154], [167], the cross-check method as reviewed in Section 2.7 is employed in this work. The cross-check method only returns feature matching pairs (*p, q*) where the *p*-th feature from an image is nearest to the *q*-th feature from another image in a pair in the matched local feature collection and vice versa [119].

3) Perform Geometric Verification using RANSAC and the remaining features are taken as the correctly matched features and labelled as c=1;

4) Calculate the 'matchability' of local features using (4.7).

To calculate the 'matchability' of local features associated with θ, define the region of θ as G. The θ is calculated for all the local features detected from the dataset. And then, the histogram of all the features for θ is calculated and denoted as $\mathrm{h}(\theta_{DE} \in \mathrm{G})$ using S bins while the histogram of correctly matched features for θ is denoted as $\mathrm{h}(\mathrm{c} = 1 \cap \theta_{DE} \in \mathrm{G})$ using same bin number. The bin number S can be different for different relevance metrics. Then the 'matchability' of features associated with θ is defined as:

$$Matchability(\theta) = p(c = 1|\theta \in G) = \frac{h(c=1 \cap \theta \in G)}{h(\theta \in G)} \qquad (4.7)$$

To find a proper relevance metric θ, three metrics in different domains are considered for feature selection: 1) Keypoint domain using Local Patch Entropy (θ_{LPE}); 2) Descriptor domain using Descriptor Entropy (θ_{DE}); and 3) Compressed domain using DCT coefficients of a local patch around a keypoint (θ_{DCT}). The definitions of these metrics are described in the following subsection.

4.4.3 Feature selection using local region entropy in the spatial domain

The local entropy is used to determine the local complexity of an image [223]. Intuitively, the local entropy is an efficient metric to select the local features. After the feature detection, given a detected feature point *x*, a local neighborhood R_x around that feature point, which takes on pixel values $\{r_1, \dots, r_a\}$, the local patch entropy can be calculated as:

$$\theta_{LPE} = -\sum_i P_{R_x}(r_i) log_2 P_{R_x} \tag{4.8}$$

where $P_{R_x}(r_i), i \in [1, a]$ is the probability of r_i based on the histogram of pixel values. Thus, each detected feature point x can be assigned a relevance metric θ_{LPE}. Although the bin number influences the precision of estimated density $P_{R_x}(r_i)$, it has little effect on the ranking of the local features using θ_{LPE}. Since grayscale images are used, 0~255 (i.e. 256 bins) are used for histogram computation.

4.4.4 Feature selection using descriptor entropy in the descriptor domain

A local feature is represented by a descriptor which is a vector that normally encapsulates certain high level characteristics extracted from pixel values. For example, a SIFT descriptor encapsulates the gradient and orientation information around a SIFT keypoint [80]. The assumption is that the more entropy the descriptor has, the more distinctive information is encapsulated in the descriptor and thus the more important the descriptor is. Therefore, the entropy of a descriptor is considered

as a relevance metric for feature selection. Given a detected feature point x and a corresponding n-dimensional descriptor, $D_x \epsilon R^n$ takes a value on each dimension $\{d_1, \dots, d_b\}$ and encapsulates the high level information around a keypoint. The descriptor entropy can be calculated as:

$$\theta_{DE} = -\sum_i P_{D_x}(d_i) log_2 P_{D_x} \quad (4.9)$$

where $P_{D_x}(d_i), i \in [1, b]$ is the probability of d_i based on the histogram of descriptor values (0~255 or 256 bins since the SIFT descriptor is used and each dimension of the SIFT feature is represented by 8 bits). Other descriptors can use a different entropy function to calculate the entropy value based on the data type and region. Therefore, each detected feature point x can be assigned a θ_{DE} computed from the corresponding descriptor.

4.4.5 Feature selection using DCT coefficients in the compressed domain

The DCT coefficients associated with SIFT feature have been studied in Section 4.2 and it is known that DC and AC components are related to the matching accuracy of the SIFT feature. In addition, the DCT coefficients have been widely used for compressed domain retrieval and the DC component and first two AC coefficients contain the main structural information of the image as studied in [32], [175], [177]. Therefore, DCT coefficients could be a potentially efficient relevance metric for feature selection. Given a detected feature point x, a 16*16 local image patch around the keypoint (as the region of a SIFT descriptor is 16*16), a 16*16 2D-DCT transformation is applied in the local patch to calculate the coefficients θ_{DCT}:

$$\theta_{DCT}(u, v) = \alpha_u \alpha_v \sum_{x=0}^{I-1} \sum_{y=0}^{J-1} f(x, y) \cos \frac{\pi(2x+1)u}{2I} \cos \frac{\pi(2y+1)v}{2J} \quad (4.10)$$

where

$$\alpha_u = \begin{cases} 1/\sqrt{I}, u = 0 \\ \sqrt{2/I}, 1 \le u \le I-1 \end{cases}; \alpha_v = \begin{cases} 1/\sqrt{J}, v = 0 \\ \sqrt{2/J}, 1 \le v \le J-1 \end{cases}$$

Here, I=J=16. The following DCT coefficients θ_{DCT} are mainly considered: $\theta_{AC1} = \theta_{DCT}(0,1)$; $\theta_{AC2} = \theta_{DCT}(1,0)$ as these components contain the main edge information of the local patch compared to the DC component and higher frequency AC coefficients [32], [175] . Therefore, each detected feature point x can be assigned a series of relevance metrics θ_{DCT}.

4.4.6 Feature Selection using hybrid selection method

The hybrid feature selection method is proposed in the manner of naive Bayesian by multiplying the aforementioned selection metrics $\{\theta_{LPE}, \theta_{DE}, \theta_{AC1}, \theta_{AC2}\}$ to produce the final hybrid relevance metric θ_{LDAC}:

$$Matchability(\theta_{LDAC}) = Matchability(\theta_{LPE}) * Matchability(\theta_{DE}) * Matchability(\theta_{AC1}) * Matchability(\theta_{AC2}) \quad (4.11)$$

4.4.7 Experimental dataset

Focused on MAVS applications of matching print media images, such as cover images, newspaper images and natural scene images, three datasets covering a wide

Table 4-3 Summary of image dataset used for evaluation

Dataset	Image Type	Distortion	# distorted images	# clean images
MVS dataset [35]	CD, DVD, Book cover	Combined Optical and Geometric Distortions	1204	301
CSIQ dataset [36]	Natural scene image	Single distortion	720	30
NN dataset [37]	Newspaper image	Single distortion	4104	171
Total images			6028	502

range of image complexity are used for evaluation. The employed datasets also contain images with controlled single distortion and realistic combined distortion that normally occur in the MAVS applications, by which the influence of realistic optical and geometric distortions on the proposed feature selection method is studied. The details of the datasets are listed in Table 4-3 .

The first dataset is the 'cover images' set of the Stanford MVS dataset [194], including CD, DVD, and book covers. This dataset provides different types of cover images captured from heterogeneous low and high-end camera phones. These images contain complicated combined optical and geometric distortions that reflect realistic situations: rigid objects, widely varying lighting conditions, perspective distortion, foreground and background clutter, JPEG compression. The ground-truth clean

Table 4-4 Summary of applied single distortion

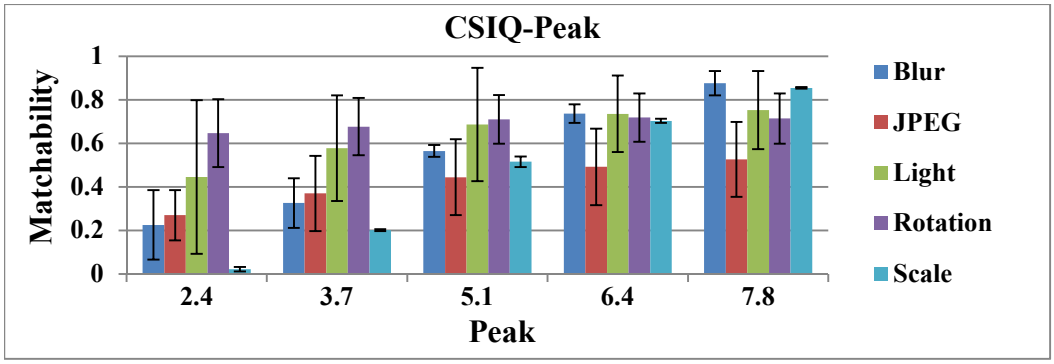
Distortion Type	Parameters	Simulation method	Inducement
Out-of-Focus Blurring	Gaussian standard deviation σ from low blurring to high blurring: B3=0.8, B5=1.1, B7=1.4, B9=1.7	The out-of-focus blurring is simulated by using a Gaussian Filter [38].	amateur capture and uncontrolled auto-focus
JPEG Compression	Quality Factor Q: Q2=2, Q5= 5, Q10=10, Q15=15, Q20=20, Q30=30	JPEG Encoder from Independent JPG Group [39].	Image save and transmission
Global Illumination Change	pixel intensity shift: L5=5, L30=30, L60=60, L80=80, L110=110, L128=128	Original pixel plus pixel intensity shift	Environmental lighting change
Rotation	Rotation angle: R15=15°, R30=30°, R45=45°, R60=60°, R75=75°, R90= 90°	counter clockwise image rotation	Random camera shot
Scaling	Scale Factor: S1/2=1/2, S1/4=1/4	image pyramid reduction	Random camera shot

reference images are also provided in this dataset. The second dataset is CISQ dataset [216], which contains natural scene images with a variety of content. The third dataset is National Newspaper (NN) dataset [169] with various images appearing in a range of published newspapers. The CSIQ and NN datasets both provide clean images. Several single distortions that normally occur during the capture of MAVS applications, including out-of-focus blurring, JPEG compression, global illumination change, rotation, and scaling, are applied to these clean images to generate controlled distortion images for study. The types and parameters of applied distortions are summarized in Table 4-4. The image pair list is established across these three datasets while each image pair contains distorted image and corresponding clean image.

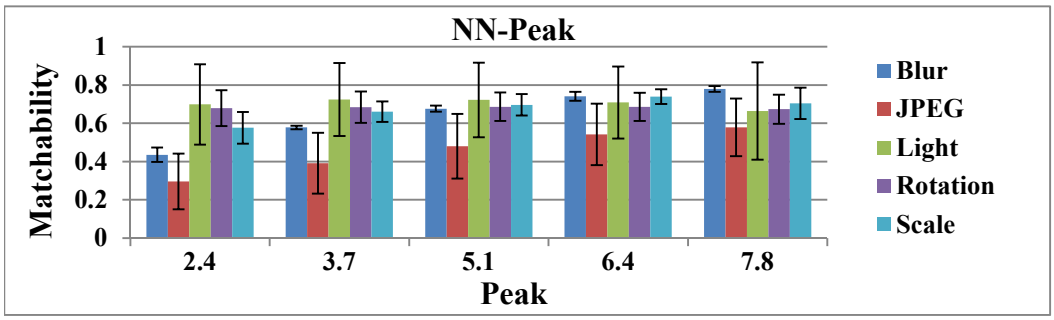
4.4.8 Learning the 'Matchability' using the proposed relevance metrics under varying single distortion type

Using the methodology introduced in Section 4.4.2, the 'matchability' of proposed methods is studied under varying single distortion in this section. The peak value of the SIFT detector for feature selection [153], [167] is also presented for comparison. Figure 4.11 to Figure 4.14 show the 'matchability' of using relevance metrics $\{\theta_{peak}, \theta_{LPE}, \theta_{DE}, \theta_{AC1}, \theta_{AC2}\}$ under varying distortions in CSIQ and NN datasets, respectively. Different relevance metrics have different response to varying distortion. The error bar shows the variance of 'matchability' under different degrees of distortions. See Table 4-4 for detailed distortion parameters.

From Figure 4.11, the 'matchability' of local features associated with θ_{peak} shows a linear increase with the growth of θ_{peak} both in CSIQ and NN dataset, except for rotation. Rotation only has little effect on the 'matchability' of θ_{peak} (i.e.



(a)

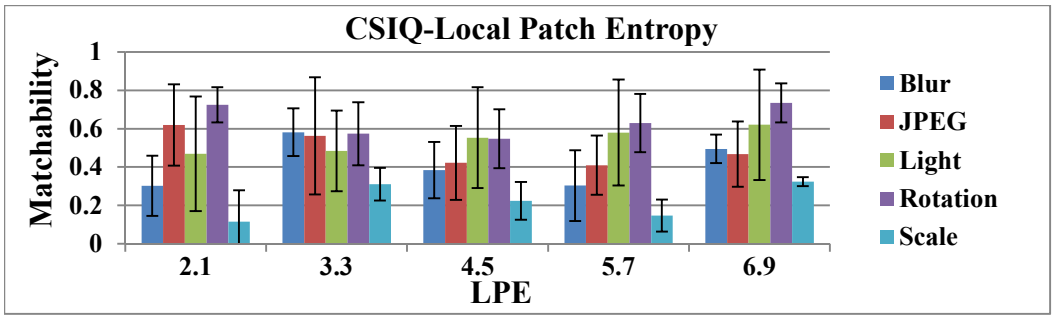


(b)

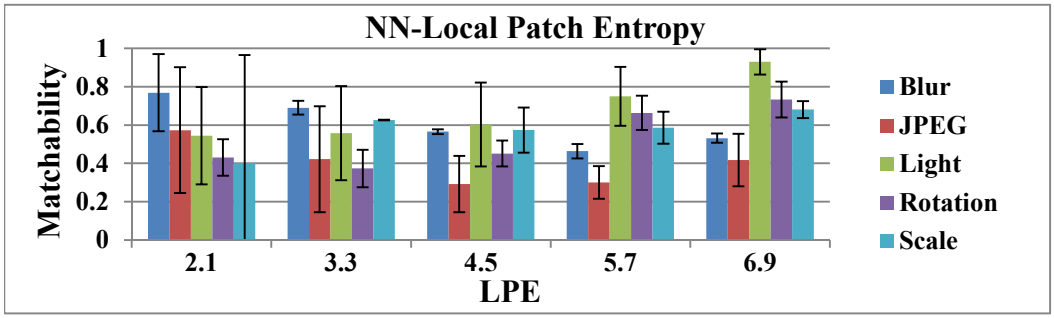
Figure 4.11 'Matchability' of Local features using selection metric $\{\theta_{peak}\}$ [34], [35] for feature selection in (a) CSIQ and (b) NN dataset under varying distortions.

the 'matchability' did not change a lot with the rotation). The distortion mainly influences the local features with low θ_{peak} while the local features with high θ_{peak} are more robust to varying distortions. The majority of local features with low θ_{peak} disappear under a high degree of distortion. The linear growth characteristic of 'matchability' of θ_{peak} indicates that θ_{peak} is a decent metric for feature selection under distortion.

From Figure 4.12, the single distortion shows a linear influence on the 'matchability' of θ_{LPE}. The stronger the distortion was, the lower the 'matchability' for the features. The features with medium θ_{LPE} were least robust to distortion. This is because such features are extracted from moderate textural image patches. Some features with low θ_{LPE} come from a shape edge, which makes these features distinctive for matching. In other words, the 'matchability' of local features



(a)



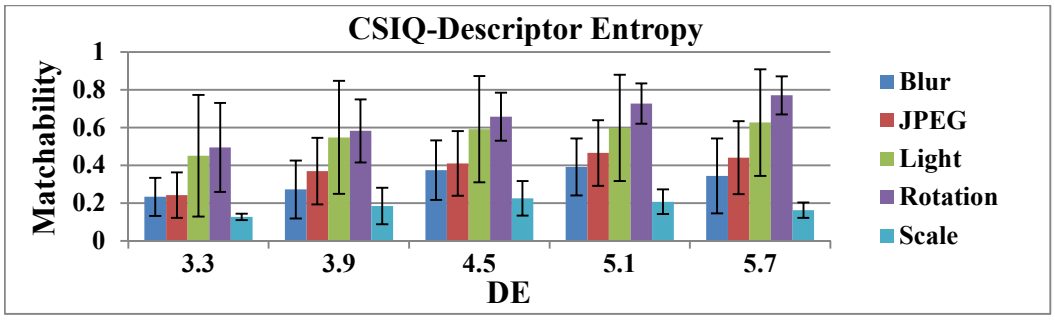
(b)

Figure 4.12 'Matchability' of Local features using proposed selection metric $\{\theta_{LPE}\}$ for feature selection in (a) CSIQ and (b) NN dataset under varying distortions.

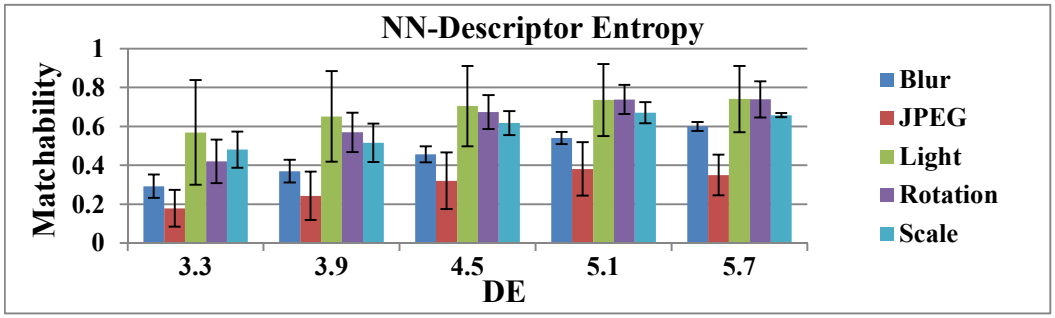
associated with θ_{LPE} shows a certain unique distribution. Such distributions can be used for efficient feature selection as well.

From Figure 4.13, it can be seen that with the increase of the strength of distortion, the 'matchability' of local features associated with θ_{DE} shows a linearly increasing trend with the increase of θ_{DE}, both for the CSIQ and NN datasets. This indicates that the higher the entropy of a feature descriptor, the more robust the feature is. A majority of features with low θ_{DE} disappeared under high distortion. The θ_{DE} is a good candidate to select features, which are significant for feature matching meanwhile robust against varying distortion;

From Figure 4.14, the 'matchability' of local features associated with $\theta_{AC1}, \theta_{AC2}$ shows different patterns under varying distortions. Similar to the situation of θ_{LPE}, a certain unique pattern can be found for feature selection by using



(a)

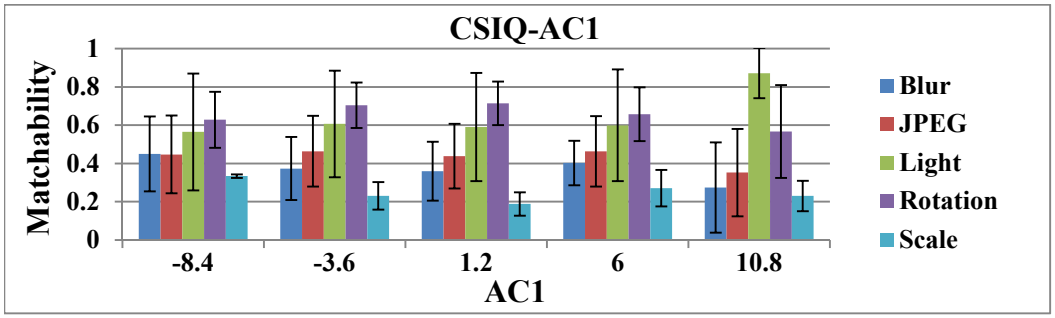


(b)

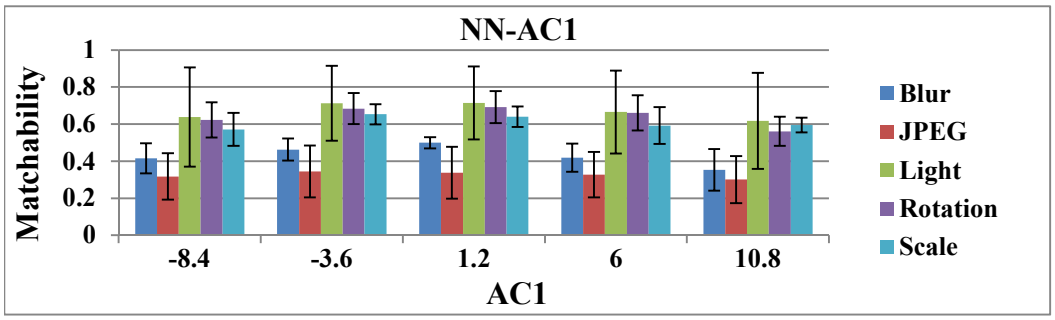
Figure 4.13 'Matchability' of Local features using proposed selection metric $\{\theta_{DE}\}$ for feature selection in (a) CSIQ and (b) NN dataset under varying distortions.

θ_{AC1} *and* θ_{AC2}.

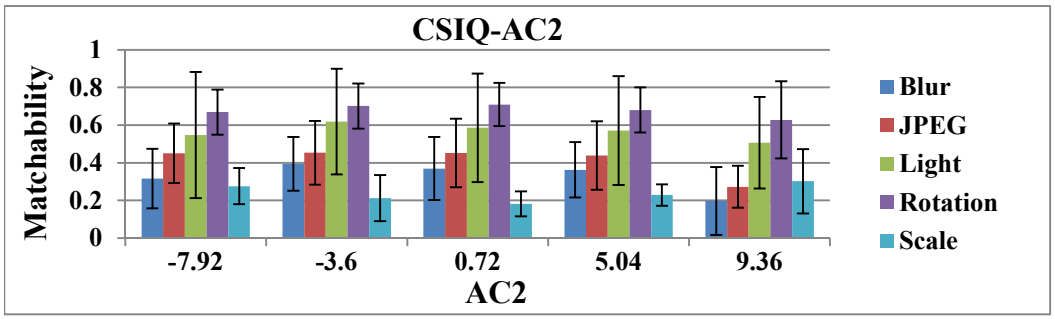
Overall, the single varying distortions investigated in this work have a linear impact on the 'matchability' of the proposed relevance metrics. The 'matchability' linearly decreases with the increase in distortion. The 'matchability' using θ_{peak} and θ_{DE} is more independent of image type compared to $\theta_{LPE}, \theta_{AC1}$ *and* θ_{AC2} as they both show consistent variance for different image types in the CSIQ and NN datasets. Therefore, a trained 'matchability' of θ_{peak} and θ_{DE} from one dataset may be universally used for another dataset.



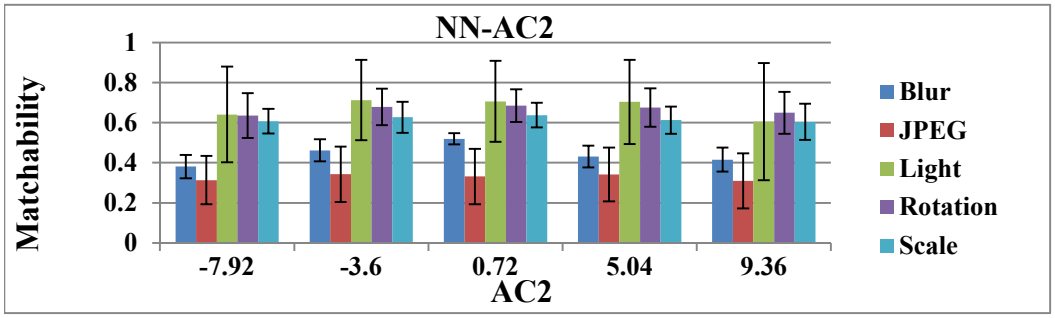
(a)



(b)



(c)



(d)

Figure 4.14 'Matchability' of Local features using proposed selection metric $\{\theta_{AC1}\}$ for feature selection in (a) CSIQ and (b) NN dataset and $\{\theta_{AC2}\}$ for feature selection in (c) CSIQ and (d) NN under varying distortions.

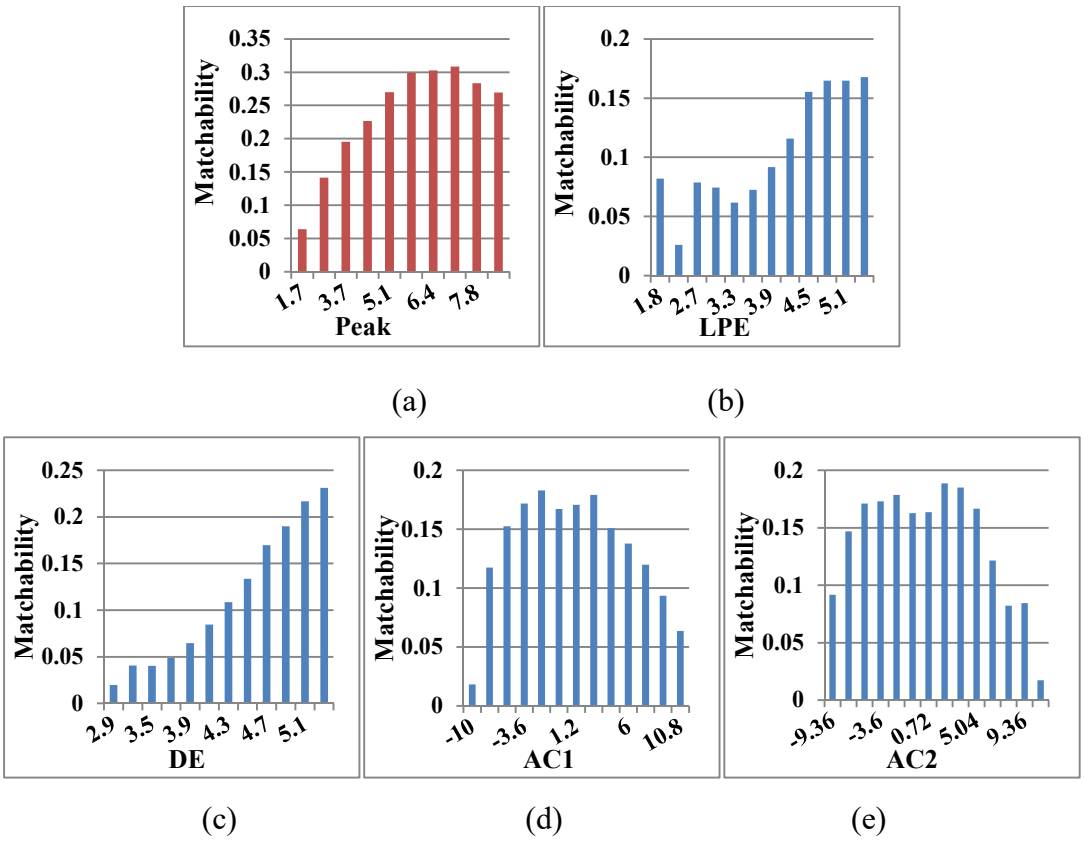


Figure 4.15 'Matchability' of Local features using the proposed method $\{\theta_{peak}, \theta_{LPE}, \theta_{DE}, \theta_{AC1}, \theta_{AC2}\}$ for feature selection in the MVS dataset under realistic combined distortions.

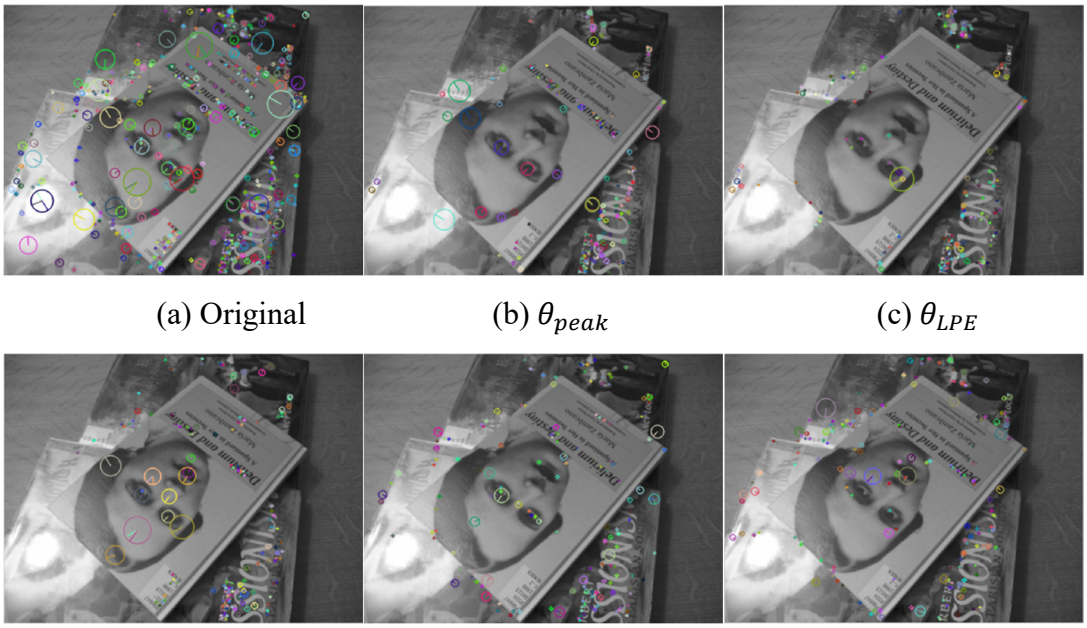
4.4.9 Learning the 'matchability' for feature selection under complex combined distortions in realistic

To verify that the proposed method can be used robustly against complex distortions in practice, the MVS dataset is employed to learn the 'matchability' of local features associated with $\{\theta_{peak}, \theta_{LPE}, \theta_{DE}, \theta_{AC1}, \theta_{AC2}\}$ under realistic combined distortion. Figure 4.15 shows the result of learned 'matchability'. It is evident that the 'matchability' still shows a linear growth trend when using θ_{peak} and θ_{DE}. The θ_{LPE}, θ_{AC1}, θ_{AC2} exhibit distinctive distributions which are also effective for filtering the local features. Overall, the 'matchability' of the proposed relevance metric θ_{DE} shows a linear increase with the growth of θ_{DE}. This result confirms the assumption that the higher the entropy embedded in a feature, the more important the feature is

for correct feature matching. Although the 'matchabilities' of the local patch entropy θ_{LPE} and DCT coefficients, θ_{AC1} and θ_{AC2}, exhibit nonlinear variation, these unique patterns can be utilized for feature selection as well.

After feature detection, each feature is assigned a posterior probability for correct matching based on the relevance metrics and corresponding trained 'matchability'. The features are then ranked from high probability to be matched to low probability. After sorting, the local features extracted from the captured scene can be easily filtered on the basis of ranked features using a threshold. The threshold can be a certain 'matchability' or a certain feature number according to different application requirements. If a MAVS application requires high accuracy, a low 'matchability' threshold can be used to choose more features to perform retrieval. Alternatively, if fast processing and transmission is desired, a feature number threshold can be employed to choose a certain number of features which have high 'matchability'.

An example of using the proposed relevance metric $\{\theta_{peak}, \theta_{LPE}, \theta_{DE}, \theta_{AC1}, \theta_{AC2}\}$ to select 210 local features from a CD cover with complex distortions is shown in Figure 4.16. The selected local features show slightly different distributions when using different methods. The features filtered using θ_{peak} are mainly located in the part of face and the text under the face. The features remaining after filtering by θ_{LPE} are mainly located in the part of face and the features extracted from the text are filtered out. The features remaining after filtering by θ_{DE} are mainly located on the face and some text features remain as well. The features filtered by θ_{AC1} and θ_{AC2} are mainly located along the edges. It is also noted that the features filtered by $\{\theta_{DE}, \theta_{AC1}, \theta_{AC2}\}$ are more clustered compared to $\{\theta_{peak}, \theta_{LPE}\}$. Such clustering can be a



(a) Original (b) θ_{peak} (c) θ_{LPE}

(d) θ_{DE} (e) θ_{AC1} (f) θ_{AC2}

Figure 4.16 Example images of using the proposed method $\{\theta_{peak}, \theta_{LPE}, \theta_{DE}, \theta_{AC1}, \theta_{AC2}\}$ for feature under realistic combined distortions.

benefit for geometric verification and results in better retrieval accuracy. This will be discussed further in the next section of retrieval experiment.

4.4.10 Retrieval Experimental result of using proposed feature selection

Recall the review in Section 2.10, the feature selection performance is reported using precision and recall of local features which solely measure the robustness of retrieving true positive local features (i.e. image pairwise matching accuracy) in the literature [162], [167], [168] while the position of the retrieved correct content in the retrieval list is not considered. However, the position of the correct content in the retrieved list is significant for MAVS applications because MAVS directly triggers the display of the first returned content. Therefore, in contrast to CDVS evaluation, which uses precision and recall to report retrieval results of local features, precision @ 1 is employed as in [224] to evaluate the retrieval accuracy of applying the proposed feature-relevance-based selection methods under different bitrates (i.e.

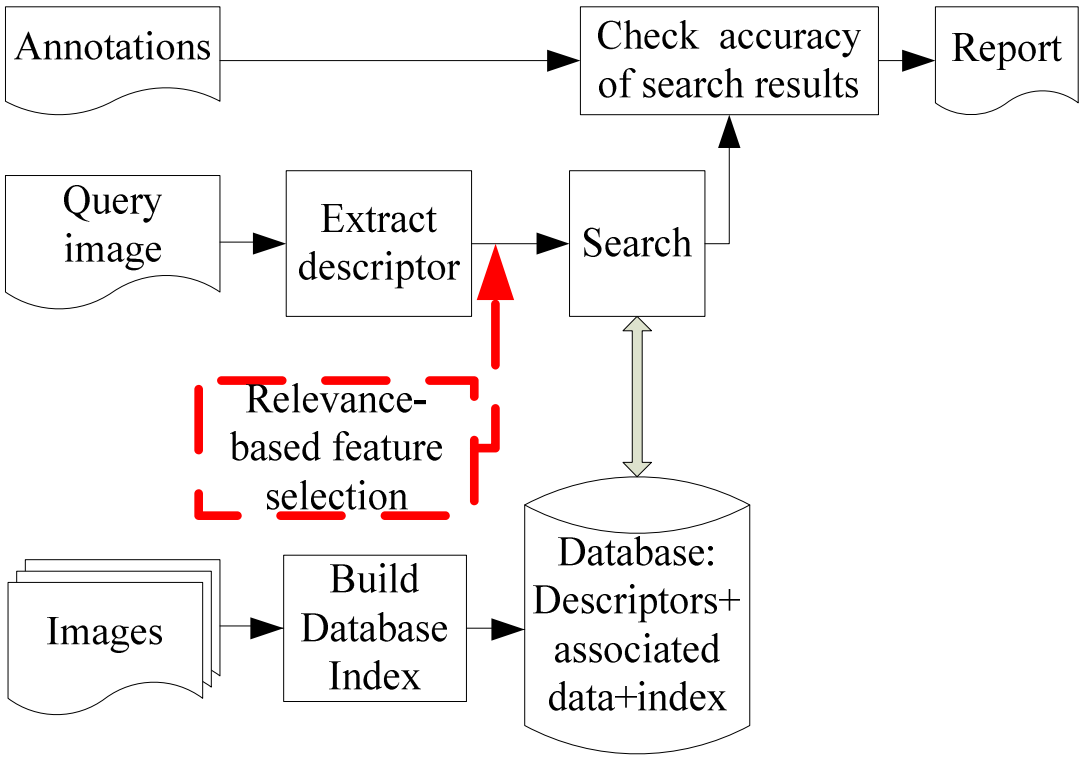


Figure 4.17 The retrieval experimental architecture. This architecture adopts the CDVS retrieval evaluation architecture with a modification of adding the proposed feature selection method. The red dotted block indicates the modification.

transmitting varying number of features). It is noted that precision @ 1 measures the retrieval performance of whether the predefined corresponding image is returned in the first place or not in the MAVS application. The retrieval experimental architecture adopts the CDVS retrieval experimental architecture with a modification of adding the proposed feature selection approach. The diagram of the retrieval architecture is shown in Figure 4.17.

The experimental procedure is as follows:

A. For each query image in the dataset:

(a) Detect and extract the local features;

(b) Select the specified number of features using the proposed feature selection methods. This forms the query feature set with the remaining features filtered out.

B. For the reference images in the dataset:

(a) Detect and extract the features for each reference image;

(b) Combine the detected features of each reference image to set up the training feature set;

(c) Perform K-Dimensional (KD) tree [225] *structure training to obtain the reference feature search space.*

C. For each query feature set:

(a) Perform the nearest neighbour search using KNN (k=1) for each query feature in the trained reference feature search space;

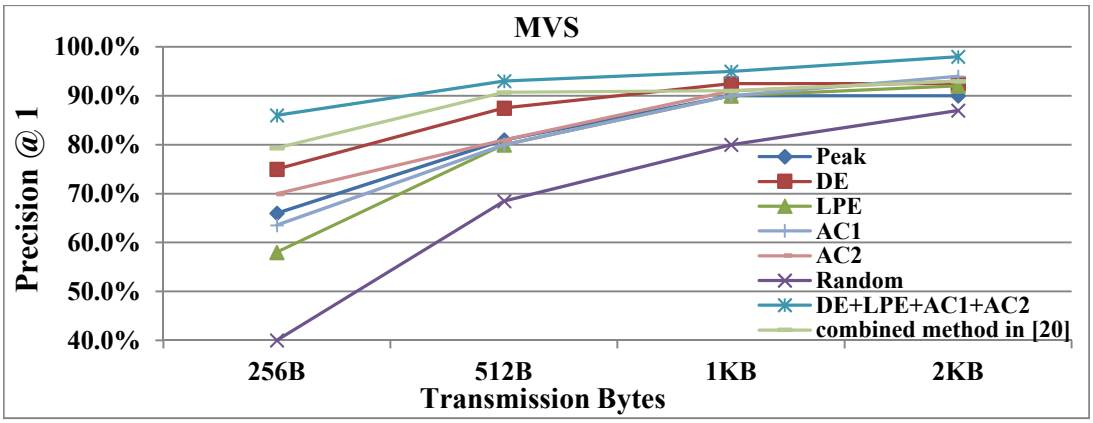
(b) Obtain the first N (N=3) (increasing N did not bring out significantly better retrieval results) reference images with maximum feature matching pairs (instead of using a threshold of distance ratio test in [167]*, the maximum number of feature matching pairs is used to determine the rank 1 candidate);*

(c) Perform cross-check KNN (k=1) search within each chosen reference image to further filter the features;

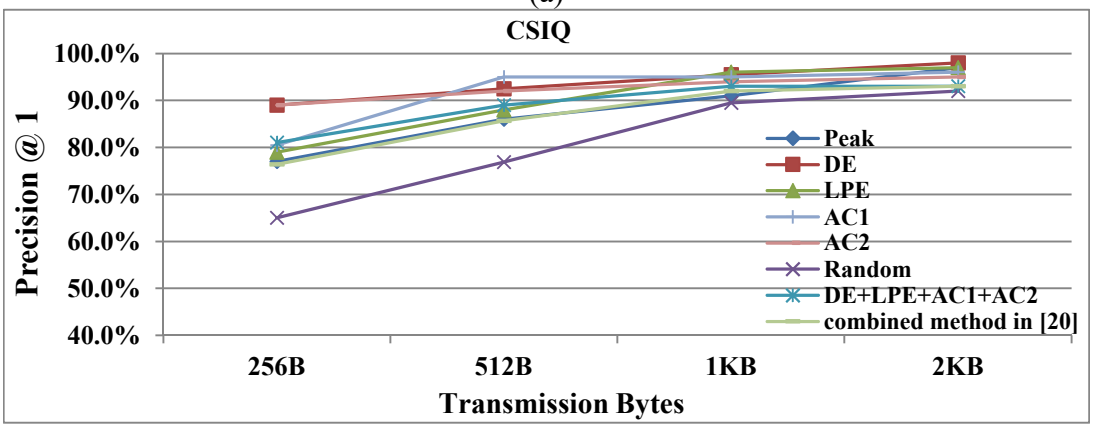
(d) Apply geometric verification (RANSAC) to find the final true positive feature matching pairs.

(e) Locate the reference image on the basis of the highest number of true positive feature matching pairs;

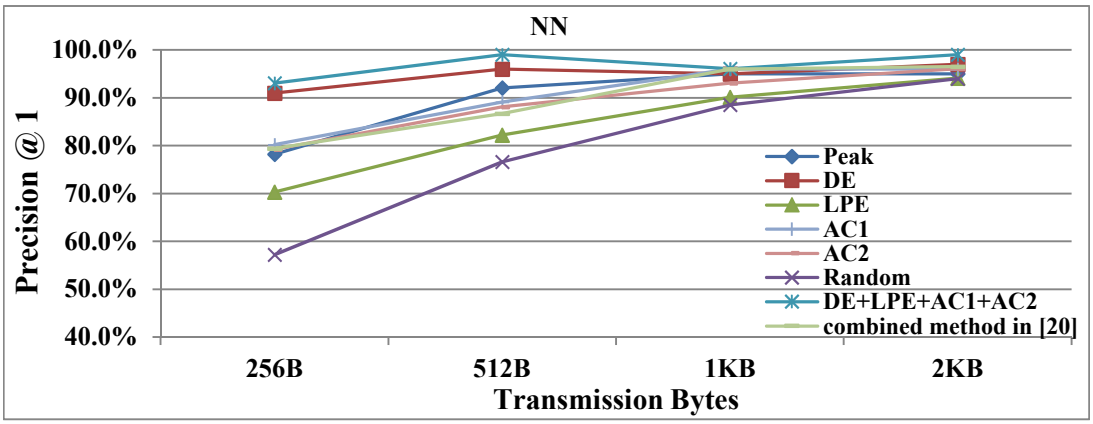
(f) The matching accuracy is evaluated based on the precision @ 1 to judge the retrieval performance [224], [226] *under different bitrate as reviewed in Section 2.10.3.*



(a)



(b)



(c)

Figure 4.18 The retrieval performance of proposed feature-relevance-based feature selection methods compared with peak-based, combination (i.e. peak+central bias+orientation+scale) and random feature selection method under varying low bitrate in MVS, CSIQ and NN datasets as shown in (a),(b), (c), respectively..

4.4.11 Comparison experimental results for using proposed feature selection methods

For comparison, the retrieval experimental results of using the proposed feature-relevance-based selection methods based on $\{\theta_{LPE}, \theta_{DE}, \theta_{AC1}, \theta_{AC2}\}$ and the hybrid method of these four metrics θ_{LDAC}, the θ_{peak} based feature selection and combination method defined as θ_{pcos} (i.e. using peak, central bias, orientation and scale together) and random feature selection in [167] for SIFT feature are presented in Figure 4.18. The reason of choosing θ_{peak} based feature selection and the combination method for comparison is that these two methods achieved the best true positive rates in [167]. The random feature selection generates a random keypoint index list to choose features. Four different feature number conditions are considered in the experiment 279, 210, 114 and 50 which correspond to 2KB, 1KB, 512B and 256B compressed feature transmission sizes. The first three bit rates are standardized in the MPEG-7 CDVS [152]. The fourth bit rate is also considered in the scenario of a very poor communication condition or processing condition where a very fast transmission is desired (e.g. processing a stream of video frames to repeatedly look for a matching reference image).

The retrieval results of using different feature selection methods in MVS, CSIQ and NN datasets are shown in Figure 4.18: 1) from Figure 4.18-(a), it is evident that θ_{LDAC} outperforms other selection methods in MVS dataset. The θ_{LDAC} achieves a 6.7% retrieval performance gain at 256B compared to the θ_{pcos}. θ_{DE} achieves a 6% and 9% retrieval performance gain at 512B and 256B compared to θ_{peak}, respectively. Considering the total test database containing 6028 query images, these are significant differences. The θ_{LDAC} improves precision @ 1 with 11%

increment at 256B compared to θ_{DE} in MVS dataset while the θ_{pcos} achieves 13% improvement at 256B compared to θ_{peak}. The θ_{AC1} and θ_{AC2} can also efficiently select the important features in the MVS dataset. The θ_{AC1} and θ_{AC2} achieves comparable retrieval result at 1KB and 2KB while θ_{AC2} achieves 4% performance gain at 256B compared to peak-based method. The θ_{LPE} is comparable to the θ_{peak} method and the worst performance degradation is 8% when using 256B; 2) as shown in Figure 4.18-(b), the θ_{DE} shows better retrieval accuracy than θ_{LDAC} while θ_{pcos} and θ_{peak} achieve comparable retrieval accuracy in CSIQ dataset. The θ_{DE} and θ_{AC2} achieve at least 8% retrieval performance gain at 256B compared to other selection methods. θ_{LPE} shows better retrieval accuracy than θ_{peak} in the CSIQ dataset; 3) θ_{LDAC} and θ_{DE} achieve the best retrieval results in the NN dataset as shown in Figure 4.18-(c) with 14% and 13% performance gain compared to θ_{pcos} and θ_{peak}, respectively. θ_{LDAC} achieves an average 2.5% performance gain compared to θ_{DE} at 256B and 512B. The θ_{pcos} only outperforms θ_{LPE} while θ_{AC1} and θ_{AC2} achieve around 2% better retrieval accuracy than θ_{pcos} in the NN dataset. 4) As expected, the random selection method (i.e. randomly choosing a certain number of features without any criteria) degrades the matching accuracy compared to the other methods. For 2KB transmission (i.e. 279 features), the random method still achieves around 90% precision @ 1 because it selects on average more than 85% of the features generated by the SIFT algorithm (the total number of detected SIFT features is determined by the complexity of an image). During the experiment, we discovered that the hybrid feature selection method using the assumption of naïve Bayesian can bring out slightly higher precision @ 1 results in MVS and NN datasets but has less effect on CSIQ dataset, especially at high bitrate as shown in Figure 4.18-(b). But the number

of matched local feature pairs had been increased slightly which indicated that more true positive local features had been selected at the expense of computation. This observation is consistent with the result of [167], [168]. However, extra computation is required by the hybrid methods as four selection metrics are needed to be computed for each feature, which will cause extra system delay when the feature number is large.

4.4.12 Generality and applicability of proposed feature selection methods

To study the generality and applicability of the proposed methods using single feature metric for fast feature selection, another three feature detectors for which the

Table 4-5 The precision @ 1 results of MSER detectors under different bitrate using different feature selection methods

MSER				
	256B	**512B**	**1KB**	**2KB**
DE	0.304	0.304	0.340	0.312
PE	0.292	0.352	0.348	0.344
AC1	0.284	0.324	0.336	0.312
AC2	0.284	0.320	0.332	0.360

Table 4-6 The precision @ 1 results of ORB detectors under different bitrate using different feature selection methods

ORB				
	256B	**512B**	**1KB**	**2KB**
DE	0.626	0.703	0.649	0.580
PE	0.458	0.451	0.558	0.482
AC1	0.428	0.496	0.573	0.458
AC2	0.428	0.527	0.474	0.474

Table 4-7 The precision @ 1 results of SURF detectors under different bitrate using different feature selection methods

SURF				
	256B	**512B**	**1KB**	**2KB**
DE	0.794	0.840	0.826	0.743
PE	0.559	0.591	0.623	0.674
AC1	0.674	0.683	0.725	0.669
AC2	0.711	0.725	0.725	0.651

Table 4-8 The precision @ 1 results of MSER, ORB and SURF detectors without selection

MSER	**ORB**	**SURF**
0.28	0.42	0.72

Note: all the detected features from query image are used for matching

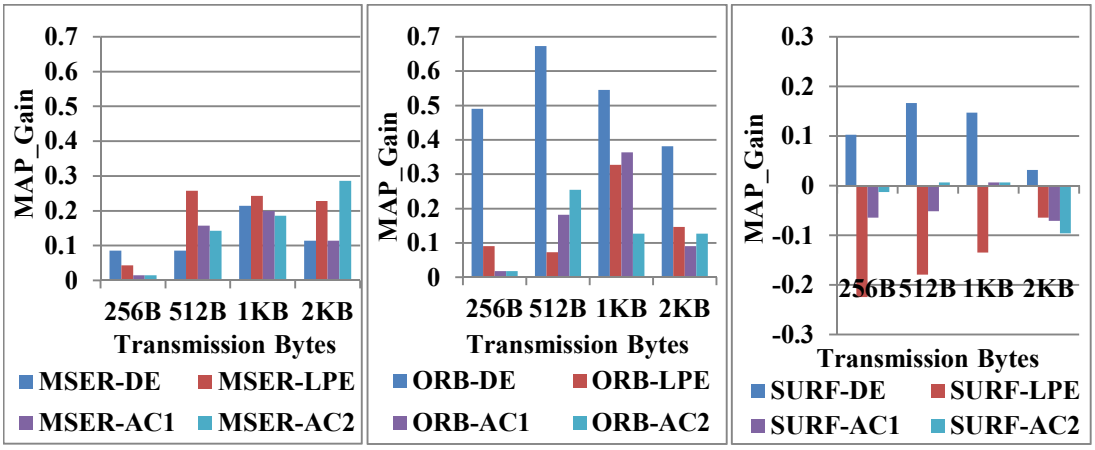


Figure 4.19 The MAP gain results for different feature detectors of using different selection methods compared to the method without selection.

θ_{peak} value is unavailable are employed. These are MSER [81], ORB [82] and SURF [83]. The precision @ 1 results of different detectors under different bit rate are shown in Table 4-5~Table 4-8.

As different detectors result in different precision @ 1, to show the effect of the proposed selection methods, the MAP gain (difference between the precision @ 1 results of using selection methods and precision @ 1 results without selection) are presented in. The positive values in Figure 4.19 indicates the precision @ 1 is improved by employing selection methods compared to the precision @ 1 result without selection method while negative values indicates the degradation of precision @ 1. The equation for calculating the MAP gain is defined as:

$$MAP\_Gain = \frac{MAP_{Selection} - MAP_{noselection}}{MAP_{noselection}} \quad (4.12)$$

where $MAP_{Selection}$ is the precision @ 1 result using feature selection methods, $MAP_{noselection}$ is the precision @ 1 result without feature selection. The legend denotes the used detector and selection method as 'Detector-Selection', for example, using MSER as detector and DE as selection method are referred as MSER-DE.

The precision @ 1 results without feature selection methods for MSER, ORB,

and SURF are 28%, 42%, 72% as shown in Table 4-8, respectively which are consistent with the results in [213], [227]. The MSER and ORB did not achieve good precision @ 1 due to complex distortions in the experimental dataset. In addition, The MSER and ORB are not scale-invariant compared to SURF and SIFT. However, we are more interested in how the proposed method can improve the precision @ 1 result. Figure 4.19 shows that the proposed feature selection methods improve the precision @ 1 for all features. The maximum gains are 28.5% for MSER using θ_{AC2} at 2KB and 67.2% for ORB using θ_{DE} at 512B. For SURF, only the θ_{DE} method achieves a maximum of 16.7% gain at 512 KB while the other selection methods lead to a negative gain (as much as 22% degradation at 256B for θ_{LPE}). The main reason for the improvement of precision @ 1 is that the false positive features are filtered out which is beneficial to the cross check matching and geometric verification. Hence, to maximize the precision @ 1 under distorted query images for MAVS applications, it is suggested that the selection method should be chosen based on the image feature and transmission bit rate being used in the matching system.

4.4.13 Conclusion

Novel methods for feature selection are proposed by which a subset of robust detected features in terms of their ability to correctly match a captured image to a reference image can be selected and transmitted at low bitrate to accurately retrieve an augmented multimedia content. The proposed metrics take advantage of the discriminative information embedded in the entropy of the local image patch, entropy of the descriptor and DCT coefficients for feature selection. When compared to start-of-the-art peak based feature selection, the proposed methods based on descriptor entropy and DCT coefficients achieve superior retrieval accuracy on a dataset with complex realistic distortions, particularly at low bit rates. The proposed

methods also improve the accuracy of MSER, ORB and SURF detectors which not only demonstrates the generality and applicability of the proposed methods but also indicates that feature selection should be still applied to the distorted query images to ensure high retrieval accuracy even if all the features can be transmitted to server under high transmission bandwidths.

4.5 Summary

In this chapter, aiming to ensure a high quality of experience in the MAVS applications, two fast and accurate low bit rate MAVS solutions are presented. One low bit rate solution based on the low frequency response of SIFT feature achieved high matching accuracy against a wide range of typical image distortions including scaling, rotation, additive noise, image blurring and illumination while the transmission data rates are comparable to existing compressed domain image features. The proposed system only needs the low frequency DCT components to be transmitted. It achieved more than 97% precision @ 1 at the transmission bitrate of 4.8KB on average in the experimental dataset. The system latency has been significantly reduced by using customised quantization table in the JPEG coder to reduce the processing delay in the client compared to performing feature detection and extraction on the client side. Another low bit rate solution makes use of novel relevance-based feature selection technology to select essential SIFT features as few and robust features as possible such that the matching accuracy is invariant to distortions caused by camera capture whilst minimising the bit rate required for their transmission. Novel relevance-based feature selection methods are proposed, based on the entropy of the image content in the keypoint domain, the entropy of the extracted features in the descriptor domain and the Discrete Cosine Transformation (DCT) coefficients in the compressed domain. The selection methods proposed in the

descriptor domain and compressed domain achieve better matching accuracy under low bit rate transmission than start-of-the-art peak based feature selection used within the MPEG-7 Compact Descriptor for Visual Search approach while the method proposed in the keypoint domain achieves comparable performance. The proposed system is robust to complex real world capturing distortion including varying lighting conditions, perspective distortion, foreground and background clutter. The proposed relevance-based feature selection methods can not only achieves low bit rate transmission but also result in a higher matching accuracy than using all features when applied to distorted images. The proposed relevance-based feature selection methods achieved more than 80% precision @ 1 when only transmitting 50 features (i.e. 256B) on the expense of extra processing delay to perform feature detection, extraction and selection on the client side. In the next chapter, focused on system delay, the influence of waiting time on the MAVS system is studied from the aspect of ensuring the quality of experience. The QoE performance of aforementioned methods is discussed as well.

5 QOE ESTIMATION BASED ON WAITING TIME AND MATCHING ACCURACY FOR MAVS APPLICATIONS

5.1 Introduction

Different state-of-the-art local feature algorithms have been evaluated in terms of the matching accuracy and processing time in the context of MAVS applications in Chapter 3. Aiming for low bitrate transmission and highly accurate retrieval, two low bitrate MAVS systems are developed in Chapter 4. The previous two chapters mainly focused on the investigation of achieving high accuracy under realistic distortions and reducing transmission delay and processing delay to maximize the QoE perceived by users. For the practical use of MAVS applications, the waiting time, as another key influencing factor for ensuring the user perceived quality, is studied in this chapter. A cut-off of waiting time that is acceptable by users is required in practice to help system designers to deploy different technologies /set different parameters in a MAVS application because different deployments can result in different waiting time and matching accuracy that will lead to different QoE perceived by users. Therefore, aiming to help system designers to make decisions in the MAVS application design, a QoE model based on the waiting time and matching accuracy is investigated in this chapter. In addition, the proposed QoE model is used to evaluate the QoE for proposed fast, low bitrate and accurate MAVS system. The study starts from a subjective evaluation experiment to investigate the effects of several impact factors on the waiting time, including the influence of changed interaction between users and media content, different media types and different progress bars. The results are then compared to traditional QoE studies in terms of the waiting time of web service for page load, video streaming, communication

connection and social network authentication. The results derived from the subjective test are employed to predict the QoE for the start-of-the-art feature selection in MPEG-7 CDVS and the proposed MAVS system in Chapter 4 based on waiting time and matching accuracy as judged by retrieval experiments on a realistic image dataset with real-world distortions caused by image capture.

The remainder of the chapter is organised as follows: Section 5.2 presents the subjective test, including the experimental methodology and MOS results. Section 5.3 presents the QoE estimation for the feature selection in MPEG-7 CDVS and the proposed MAVS systems. Summaries and conclusions are drawn in Section 5.4.

5.2 Subjective test to study the influence of waiting time for QoE estimation

5.2.1 Overview and Novelty

With the explosion of MAVS applications deployed in smart devices, users are experiencing new augmented quality of experiences brought by such emerging multimedia applications. Compared to traditional retrieval-by-click multimedia applications, such as a web service where users use the mouse to click the predefined link to explore different web content, MAVS applications involve more intuitive interaction with users, which is referred to as retrieval-by-capture. Recall the review in Section 2.11, unlike operating mouse in the traditional multimedia services, in a MAVS application, users capture the Region of Interest (ROIs) (i.e. image or video clip) of a real world scene or printed picture by a camera and then the captured content is analyzed to generate query information. Generally, the capture and query process induces a user to experience a certain waiting time (WT).

The waiting time is critical to the user perceived Quality of Experience (QoE) and different technical characteristics influence the perception of a user with respect

to WT from the aspects of application, context, network and users [228] as reviewed in Section 2.4.4 and Section 2.5:

1). the impact of waiting time on QoE has been proven to be application-dependent [228], [77], [75], [170]. For example, the comparison results in [75], [77], [228] suggested that the same initial delay of 8s led to different MOS values for different multimedia services: a) 4 for video streaming; b) 3.3 for 3G connection establishment; c) 2.5 for social network log on. In addition, the WT of video buffering at the beginning of the video playback causes a different user perception compared to the video buffering during the playback. The WT during the service (i.e. due to video stalling) resulted in the significant degradation of MOS values compared to the same WT for an initial delay from the video buffering before the video consumption [77];

2). the WT in terms of QoE is context-related. The tolerance of WT for loading video differs to the WT for loading a web page [75] or the WT for a mobile augmented application of a wine shop assistant system [170];

3). the WT is closely related to the network in terms of QoE. The bandwidth was not linearly related to WT due to the complexity and interaction of the network protocol but indirectly has the impact on the user perceived WT [73]. Different file download sizes resulted in different MOS values even if the WT for downloading the file was the same, due to the different expectations of the user. Moreover, the expectation of users for the accessed network influences user perception and such expectation can be affected by solely changing the label of the connection type [229];

4). the tolerance of waiting time in terms of QoE is influenced by users themselves. The results in [170] suggested that two groups of users who experienced the same levels of waiting time but in a different order resulted in the 5s and the 8S

acceptable level of waiting time, respectively. This phenomenon is caused by memory effect. Moreover, results from a user survey [170] indicated that users would tolerate more delay for more important shopping items and users would reject the application if the accuracy of results was less than 95%.

As an emerging networked multimedia application, the waiting time of MAVS applications investigated in this work is defined as from the start of launching the camera to the display of the retrieved content. As the key impact factor, the waiting time is mainly determined by several key procedures. Firstly, the most time-consuming part of the process is the analysis of the captured image to generate efficient query information (e.g. compressed image, selected DCT coefficients or selected local features). Here, we assume this is performed at the client side on the mobile device and the speed of this process highly depends on the employed algorithms and the computational capacity of the device. Moreover, the process of searching and matching relevant media content in the server or cloud is another time-consuming procedure depending on the retrieval method. In addition, the transmission capacity of the wireless network may be limited. As the accuracy of feature algorithms and matching methods increase, so too does the computational complexity, memory resources and transmission data sizes that are required in a MAVS system [230]. As a result, the processing time and transmission time both increase. Among these procedures, quality influencing factors from the technology domain, human domain or contextual domain interact with each other to influence the users' perceived QoE in terms of waiting time according to the review in the Section 2.4.2 of the Laghari QoE model. In addition, users may have different prior-knowledge about these factors, which may influence the user's perception and expectation as well.

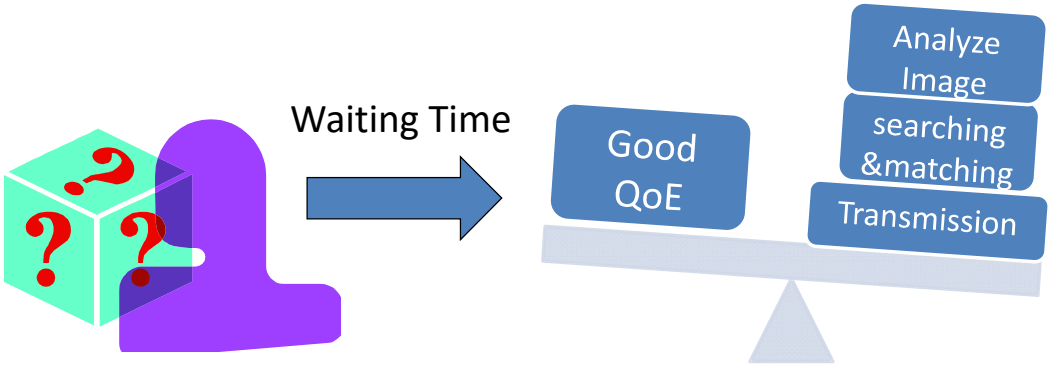


Figure 5.1 The QoE influence factors in terms of ensuring good QoE to users

Users are unlikely to accept too long a waiting time whilst too short a response time may cause the user's suspicion of the validity of the result [170]. Therefore, a key research question to answer is then: **"What is the maximum waiting time that the users would still accept and rate the service good in the MAVS applications?"** as illustrated in Figure 5.1. The trade-offs of deploying various feature extraction algorithms, transmission data size and matching algorithms are required to be carefully designed within this tolerance to ensure that the QoE perceived by users is maximized. Thus, a guideline for the waiting time is needed to help application developers to consider these trade-offs within the scope of QoE when designing and deploying MAVS applications.

In this chapter, the WT is studied from several aspects in a MAVS system: 1) a subjective user study is conducted to study the influence of linking two different media content types (video and web page) to a printed image captured by a mobile camera; 2) the impact of using different progress indicators on the QoE; 3) a comparison of subjective test results for the perception of WT for the MAVS application and other previously investigated web-based applications. Particularly, the following questions are answered:

1. *Are the user's satisfaction and acceptability of MAVS applications the same*

as the conventional retrieval-by-click applications?

2. *Does the type of linked content influence the user's satisfaction and acceptability?*
3. *Are the user's satisfaction and acceptability affected by the progress indicators within the application?*

The methodology of the subjective experiment to study the influence of waiting time on QoE is presented in the next section.

5.2.2 Subjective experimental methodology

The investigated MAVS application operates the camera in video mode (i.e. streaming images). Users 'scan' over a printed picture on a page without clicking a button to find the corresponding matching image within a database of unique images. Similar to emerging applications in newspapers and magazines as introduced in Chapter 1 [169], a matched image triggers the presentation of the 'linked' content, such as a web page or a video (both explored in this work) that was chosen by the author when creating the printed article.

For measuring the QoE in terms of waiting time, a proper QoE metric is required to find the relationship between WT and user perception. Recall the review in Section 2.4.1, the QoE as defined by the International Telecommunications Union (ITU) is "the overall acceptability of an application or service, as perceived subjectively by the end-user." [231]. Another QoE definition from the European Network on Quality of Experience in Multimedia Systems and Services (QUALINET) is "the degree of delight or annoyance of the user of an application or service. It results from the fulfilment of his or her expectations with respect to the utility and / or enjoyment of the application or service in the light of the user's personality and current state" [232]. The former definition emphasizes

"acceptability" that is perceived by the user as one metric whilst the latter definition takes the users' "satisfaction" as a key metric and explicitly states that the results are from the fulfilment of users' expectations. Furthermore, the effect of "satisfaction", as one subjective QoE measure of rating of overall satisfaction with content, is illustrated in [58]. Thus, these two metrics are utilized to evaluate the users' perceived QoE in terms of the WT for the MAVS application in this work. As one of the determinants of QoE, it is of importance to study both the acceptable threshold of the WT and the degree of satisfaction affected by varying the WT.

Proper methods are required to measure the aforementioned two metrics. Recall the review in Section 2.4.3, the widespread methodologies to measure the QoE are qualitative and quantitative methods to perform a survey to acquire the users' opinion by questionnaires or rating scales [58]. The Mean Opinion Scores (MOS) [233] test is employed to measure the user's "satisfaction" using a 5-point opinion scale. A binary scale "yes/no" is used to measure the users' decision in terms of "acceptability". Normally, users accept the quality of the system and say "yes" when the waiting time of an application under testing is over a threshold. Such a threshold indicates the quality level of the user's tolerance [234], [235]. In fact, these two subjective measures enable two kinds of measurement. MOS is a fine-grained judgment while "yes/no" is a coarse-grained judgment. These two measures are both important and offer guidelines to the service operators and providers for deploying different technical components in a MAVS system.

5.2.3 Experimental platform and procedure

The subjective experimental platform and detailed experimental procedure is presented in this section. To simulate the real-world usage environment, an Android application was developed on a 7 inch Samsung Galaxy Tab to perform the whole

procedure of capturing the image using a mobile camera, indicating the progress of image analysis and then loading relevant content. The experimental platform and process is shown in Figure 5.2.

To precisely control the WT and ensure the returned multimedia content are corresponding to the captured image (i.e. to isolate the influence of incorrect matching), the image feature and matching algorithms are disabled in the test platform. Instead, participants are informed that when they use the mobile phone

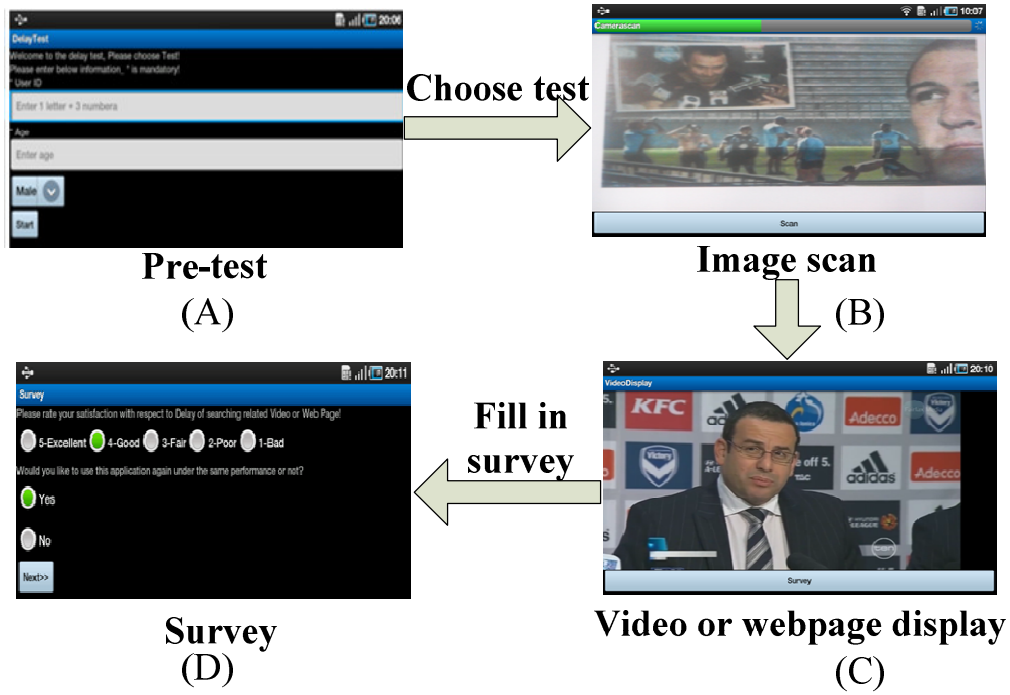


Figure 5.2 Screenshot of the experimental procedures.

camera to capture the image, the image is being processed to find a match. Meanwhile, the camera preview and the progress indicator are displayed and after a simulated waiting time the linked content (video or webpage) is displayed. The linked content, which are corresponding to the print media image presented to the participants, are predefined in the retrieval database so that the correct linking content are triggered to display to the participants.

The resulting combination of factors composed of 9 delay × 2 types of linked

Table 5-1 The experimental parameters of WT for linking video/web page to print media with different indicators

	Indicators	Linked content	Waiting time
Group 1	Progress Bar	Video	0,0.25,0.5,1, 2,4,8,10,12
		Web	
Group 2	Indeterminate spinning wheel	Video	
		Web	

content × 2 types of progress indicator as shown in Table 5-1. The participants are divided into two groups. Each group experiences the different order of delay with different indicators. For each group, both tests of linking printed images to the video and the web pages are conducted. The whole subjective experiment is composed of three phases: 1) brief instruction before the active test; 2) active test and 3) post interview. In the brief instruction, the experimental purpose and the MOS method is introduced to participants as well as the interface and the manipulation of the employed application. The activities are monitored by the researcher during the active test to ensure the participants correctly follow the experimental instruction. After the active test, a post debriefing interview in terms of waiting time is also conducted and recorded. The interview questions were:

1. "What is the maximum delay you have experienced during the test?"

2. "If the linked content can be guaranteed to be correct, but it may require longer waiting time, will you accept this?"

3. "What kind of progress indicator do you prefer?"

As the key procedure of the subjective experiment, the active test is designed as follows:

1. After launching the application, a welcome page will show up and ask the participant to complete a brief demography including the user ID, age and gender as shown in Figure 5.2-(A).

2. The participants then can choose the test of linking the print media to a video or a webpage.

3. After choosing the test, the camera is activated in the video mode and camera preview of the current capture image is displayed to the participant as shown in Figure 5.2-(B).

4. Then, the participant is required to hold the camera steady to scan a printed image to trigger the playback of a corresponding video or loading of a web page by pressing a “scan” button as shown in Figure 5.2-(C). A progress bar or spinning wheel will indicate the progress to the participant. In total, the participant is required to scan 9 different images in each test and experience varying WT, in random order, as shown in Table 5-1.

5. Finally, the participant is required to fill in an electronic survey to give their rating and opinion as shown in Figure 5.2-(D). This involves rating their degree of satisfaction experienced for each WT based on a 5-point absolute category rating scale [236] and also answering if this WT is acceptable or not.

5.2.4 Experimental images and retrieved video/webpage

A total of 36 printed images from a range of published newspapers in National Newspaper (NN) dataset [169] are printed and taken as the images to be scanned, where 18 images correspond to 18 related videos and the other half images correspond to 18 related webpages. During the subjective test, when a participant scans an image, only one corresponding video or a web page will be loaded instantaneously to provide the perception of a fast Wi-Fi connection. This aims to study if the user’s perception will be influenced by the type of linked content. Different content classes are used in the experiment to achieve a broad range of content diversity, with topics including: sports (5), Australian national news (15), world news (5), education (3), entertainment (5) and business (3), where the number of examples is indicated in brackets.

5.2.5 Experimental progress indicators

To study the influence of the different progress indicators, two indicators are used in

(A) progress bar

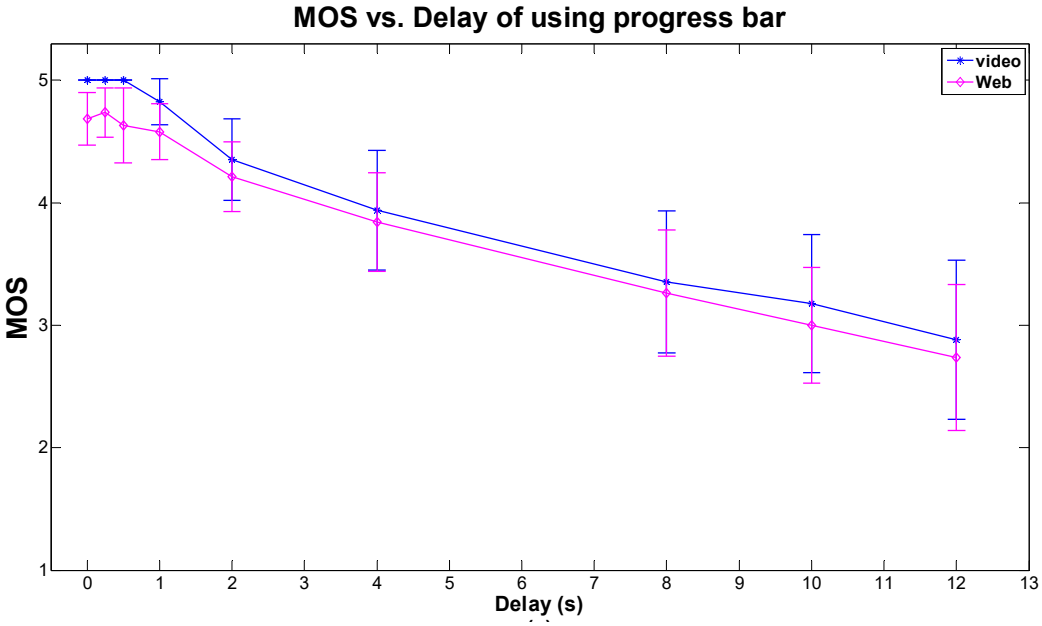
(B) spinning wheel with indeterminate time to completion

Figure 5.3 Investigated progress indicators.

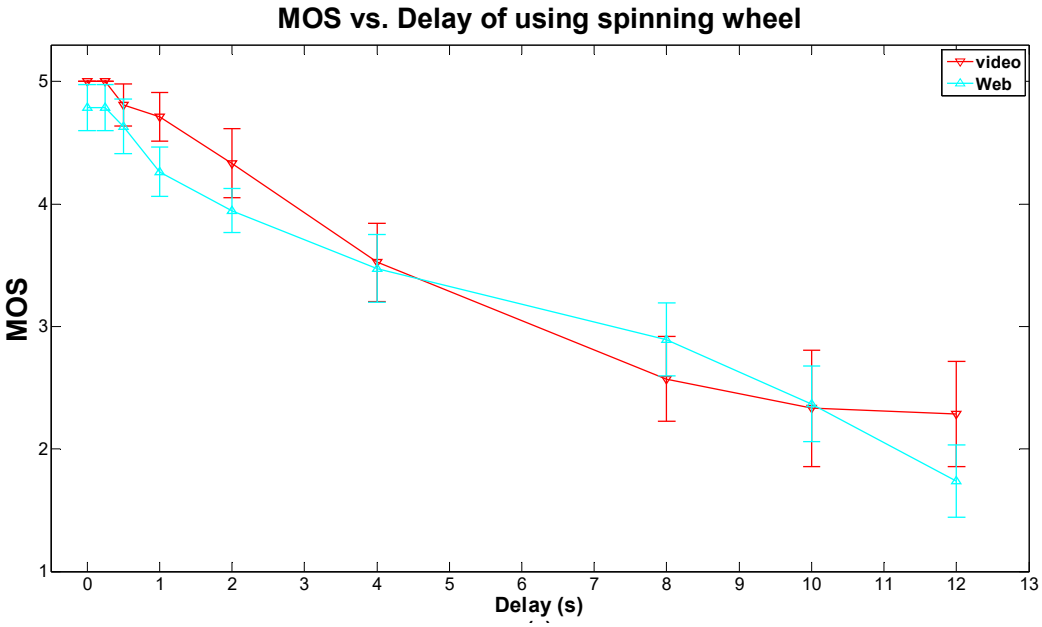
this study as shown in Figure 5.3: 1) progress bar; 2) spinning wheel with indeterminate time to completion. The progress bar shows the percentage of the processing progress while the spinning wheel just indicates the processing. The participants are divided into two groups. One group will experience the progress of image processing with the progress bar while the other group will experience the progress of using the spinning wheel. Under these two conditions, the participants will experience the same delay but in random order.

5.2.6 Experimental participants profile

A total of 51 participants were invited to take part in the subjective experiment. They are randomly divided into two groups. One group conducted the experiment using a progress bar while another group experienced a spinning wheel. Each group has at least 25 participants. All of the participants are students and staff from the University of Wollongong with the age ranging from 21 to 38. These participants have different professional status and discipline backgrounds, including informatics, engineering, medicine and commerce. The majority of participants had no prior experience of



(a)



(b)

Figure 5.4 Users' satisfaction evaluation of varying waiting time with different indicator and different linked content by using MOS. Error bars indicate 95% confidence intervals.

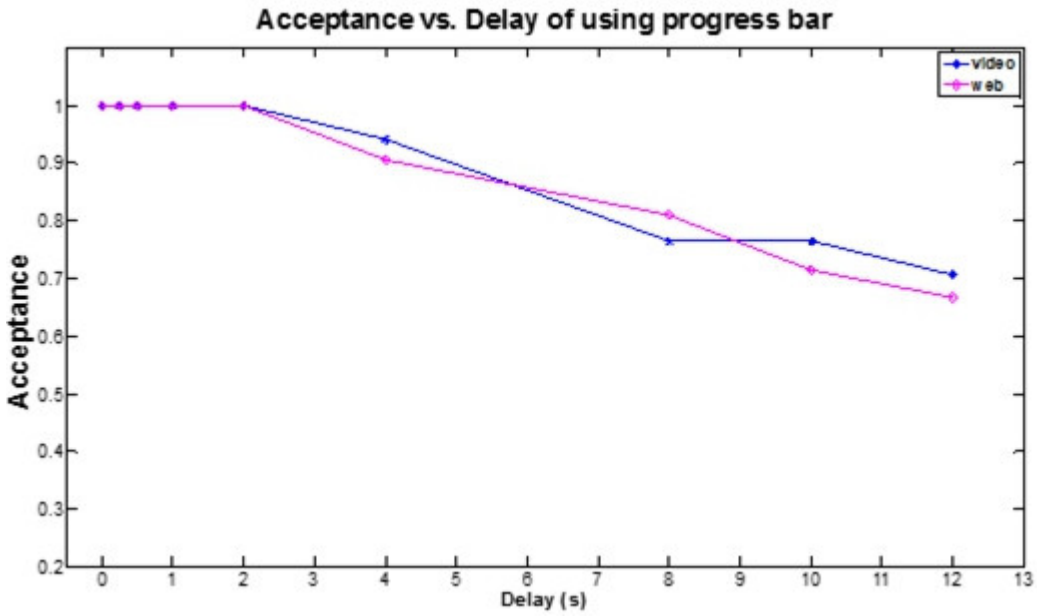
using MAVS applications and different understanding about the underlying technology of MAVS applications. Therefore, their expectation of application performance in terms of WT was considered to be varying.

As the user interface has been changed from clicking a mouse to operate a mobile camera, it is also expected that the QoE associated with WT may differ from traditional retrieval-by-click applications [75], [77]. The results of the subjective experiment are analyzed in the next section regarding the questions stated in Section 5.2.1. Firstly, the results of satisfaction and acceptance associated with WT are presented and then compared to the previous results for the conventional retrieval-by-click applications. Secondly, the influence of different progress indicators is discussed to check if the user's perception has been affected or not. Thirdly, the user's rating diversity is discussed.

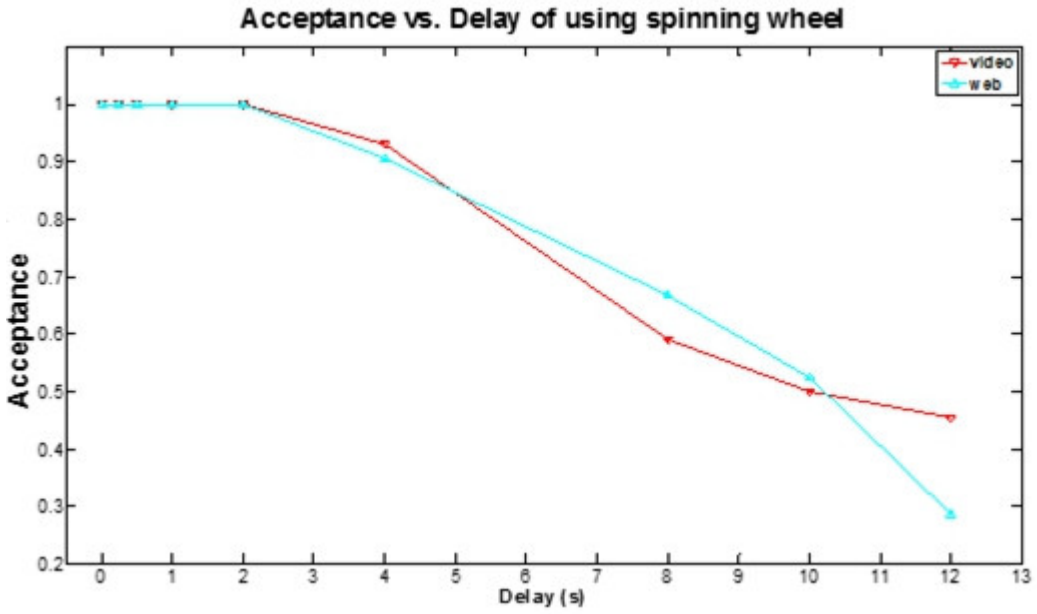
5.2.7 The influence of different multimedia type and different processing indicators for users' satisfaction and acceptance

Figure 5.4 shows the MOS results of varying WT for different progress indicators and different linked content with 95% confidence intervals. The MOS results indicate that the linked content type (i.e. video or webpage) does not have a significant influence on a participant's satisfaction for all delays (as judged by the 95% confidence levels), with the most similar results obtained for a WT less than 1 s. In particular, the maximum difference in MOS between the two linked content types is 0.368 at a WT of 0.5s when using the progress bar as shown in Figure 5.4-(a) while the maximum difference is 0.451 at a WT of 1s when using the spinning wheel as shown in Figure 5.4-(b). For a WT of more than 1s, while the satisfaction of the participants declines for each type of the linked content, the reduction in MOS is more rapid for the spinning wheel in Figure 5.4-(b) compared to the progress bar in Figure 5.4-(a).

Figure 5.5 shows the users' acceptance of varying WT for the different



(a)



(b)

Figure 5.5 Users' acceptance evaluation of varying waiting time with different indicator and different linked content by using "yes (1) /no (0)".

progress indicators and the different linked content in the test platform. The acceptability of participants is 100% when the WT is less than 2s. The WT of 2s is a turning point after which the participants start to show less tolerance to the WT. It is

also interesting that the acceptability of participants is influenced by the progress indicators. The participants gave more tolerance to the delay when they were indicated by a progress bar than the indeterminate spinning wheel. The acceptability declines faster in Figure 5.5-(b) compared to Figure 5.5-(a). Meanwhile, similar to the MOS results, the different linked content does not show a strong impact on the acceptability. These findings suggest that the linked content is not a major influence on the user's satisfaction. One possible reason is that the users may have different expectations about the loading time of the video and the web pages, although in this experiment the content was pre-loaded in the application so that loading times were not a significant factor. Further, the linked content was not the result of a traditional mobile visual search, where users would have a strong expectation on the correctness of the returned content. Rather, content was pre-defined to complement and augment the printed picture.

A significant finding to emerge from this study is that the progress indicator has an influence on the perception of users in terms of WT. One possible reason is that the progress bar, compared to the indeterminate spinning wheel, provides more feedback to users to clearly indicate the progress of the processing and provides confidence that the application is retrieving a result effectively.

5.2.8 The influence of different user interaction from click to capture

Figure 5.6 compares the results for the MAVS application here with results from the previous studies of traditional retrieval-by-click applications [75]. Although the content here is a video and a webpages, these differ to previous results for similar content from the Youtube video loading and the web page loading. The results are closer to the UMTS connection and the social network authentication. One possible

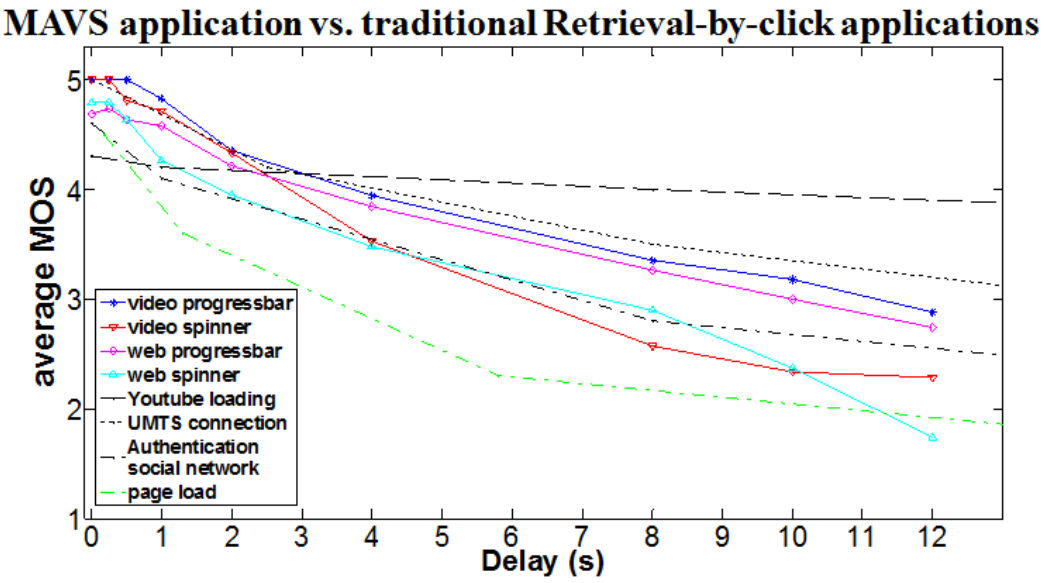


Figure 5.6 Comparison of the quality of experience in terms of waiting time in MAVS applications and traditional click-based multimedia applications.

reason is that such applications have a common character of establishing a connection. And also, the user's perception will be influenced by the difference between clicking the mouse and capturing the image by a camera.

A nonlinear least square regression method is used to fit the MOS results using different models and the results are shown in Table 5-2 along with the resulting errors measured using three methods: SSE, COD, and RMSE. It is of note that the waiting time below 0.5s should be considered especially due to the imperceptible difference [237]. This also can be found from Figure 5.4-(a) and Figure 5.4-(b) that the users' ratings are nearly the same for the waiting time below 0.5s. Therefore, the users' ratings below 0.5s are not considered when performing the curve fitting. For linking the content using a progress bar indicator and linking to a web page using spinning wheel indicator, the best fitting function is a logarithmic function. This conforms to the WQL hypothesis: "The relationship between Waiting time and its QoE evaluation on a linear ACR scale is Logarithmic" [73] and the

opinion model for web-browsing applications in ITU-T Recommendation G.1030 Annex A [238]. The only exception is linking the video using the spinning wheel indicator where the best fitting function is an exponential function. However, it also can be found that there is no significant difference between the logarithmic function and the exponential function. There are other nonlinear factors which may have an influence on users' perceived waiting time, such as the browser rendering of different web pages, users' expectation for different augmented content when scanning images using a mobile phone camera. The exact effects of these nonlinear factors and their impact on user perceived quality of experience requires further research.

Table 5-2 Mapping function between waiting time (x) and MOS (y) for different scenarios along with error measures: Sum of Squared Error (SSE); Coefficient of Determination (CoD); Root Mean Square Error (RMSE).

		Mapping function		SSE	CoD(R^2)	RMSE
Progress Bar	Linking Video	Polynomial	$y = -0.175x + 4.872$	0.1736	0.9579	0.1863
		Exponential	$y = 2.569e^{-0.163x}+2.602$	0.0279	0.9932	0.0836
		Logarithmic	$y=-1.118\ln(x+1.648)+5.864$	0.0141	0.9966	0.0594
	Linking Web	Polynomial	$y = -0.164x + 4.631$	0.0543	0.9845	0.1042
		Exponential	$y = 3.077e^{-0.088x}+1.695$	0.0125	0.9964	0.0559
		Logarithmic	$y=-1.824\ln(x+5.843)+8.022$	0.0099	0.9972	0.0499
Spinning Wheel	Linking Video	Polynomial	$y = -0.24x + 4.79$	0.4556	0.9399	0.3018
		Exponential	$y = 3.510e^{-0.159x}+1.671$	0.0435	0.9943	0.1042
		Logarithmic	$y=-1.809\ln(x+2.659)+6.992$	0.079	0.9896	0.1405
	Linking Web	Polynomial	$y = -0.238x + 4.654$	0.2278	0.9771	0.1804
		Exponential	$y = 5.995e^{-0.053x}-1.271$	0.1918	0.9807	0.1788
		Logarithmic	$y=-3.327\ln(x+8.998)+12.07$	0.1779	0.9821	0.1722

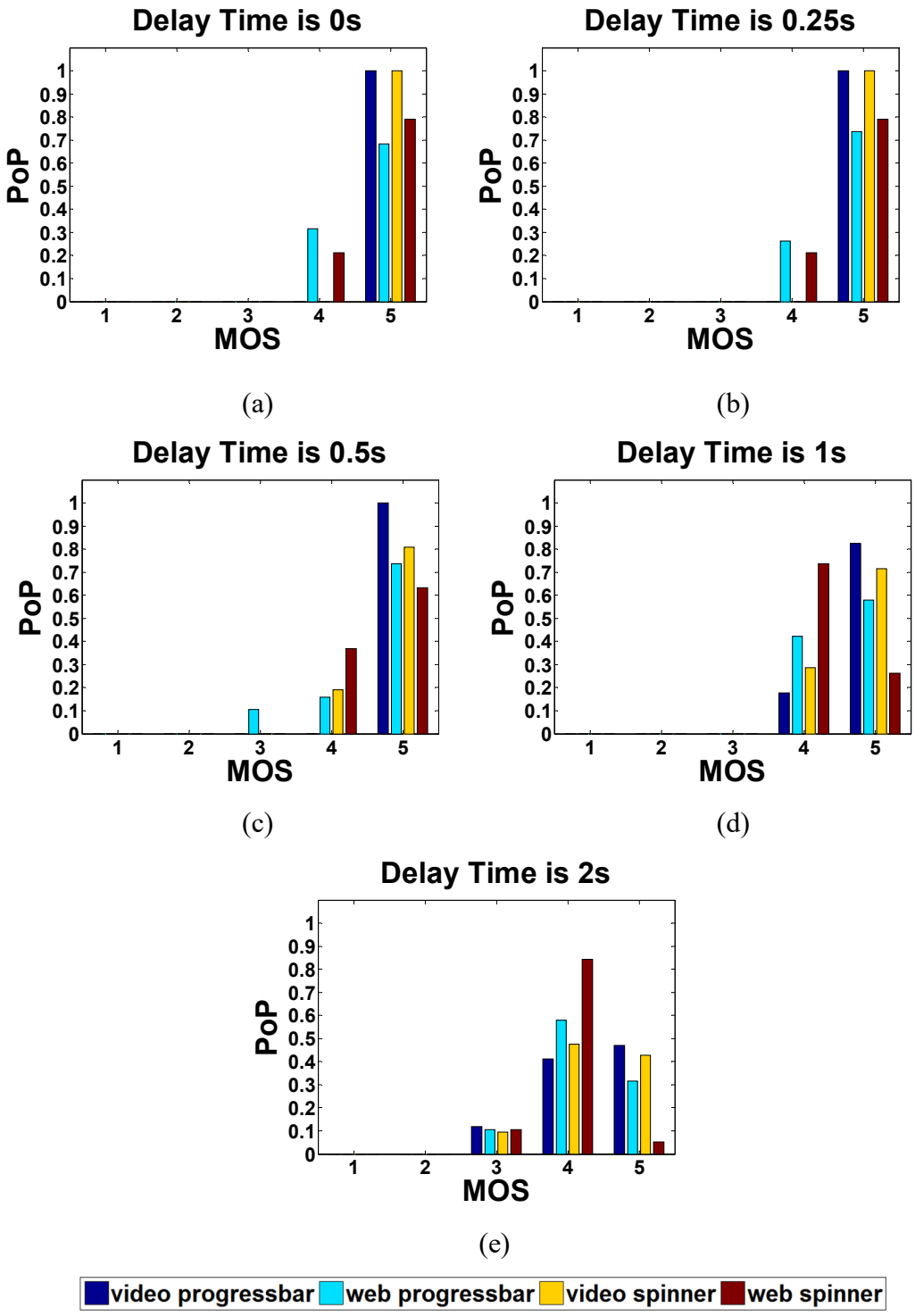


Figure 5.7 Users' rating diversity as percentage of participants (PoP) of rating results in different scenarios when the waiting time is less than 2s (i.e. acceptance is 100%)

5.2.9 User rating diversity and the influence of memory effect

The users' rating behavior is studied in this section. The percentage of participants

rating results is used to study the diversity of users' rating as shown in Figure 5.7 and Figure 5.8. When the waiting time is less than the 2s, a majority of participants gave a MOS value above 3. It is obvious that the user rating is similar when the WT is 0s and 0.25s, respectively, due to the imperceptible WT, but shows a different distribution with the increment of WT from Figure 5.7. It is of note that a small percentage of participants rate a MOS value of 4 for web page when the WT is 0s and 0.25s while the video scores 5 regardless of the indicators. In addition, the percentage of participants when using the progress bar is slightly more than when using the spinning wheel. It seems that some participants are more critical to web content delays. To a certain extent, it implies that the rating behaviors are influenced by the user's expectation of loading different multimedia content. Moreover, within the same WT, the rating behavior varies because of the difference between the type of linked content and the progress indicators as both shown in Figure 5.7 and Figure 5.8.

Another observation is that the user's perception is still affected by the memory effect even though a random order of WT was used in the experiment. For example, when the WT is 0.5s, some participants gave a MOS value of 3 while when the WT is 1s, no one gave 3 as shown in Figure 5.7-(c) and (d). After revisiting the experimental data, it is found that the participants who gave a MOS value of 3 experienced the WT of 0s and 0.25 before 0.5s and such situations only occurred when the progress bar was used. One reason for this may be due to the participants being more sensitive to the progress bar than the spinning wheel. It is also surprising that even when the WT time is increased to 8s, 10s, and 12s, some participants still gave a MOS value of 5 for video and web when using the progress bar. According to the answer of the post interview, some participants said: "It is excellent that the

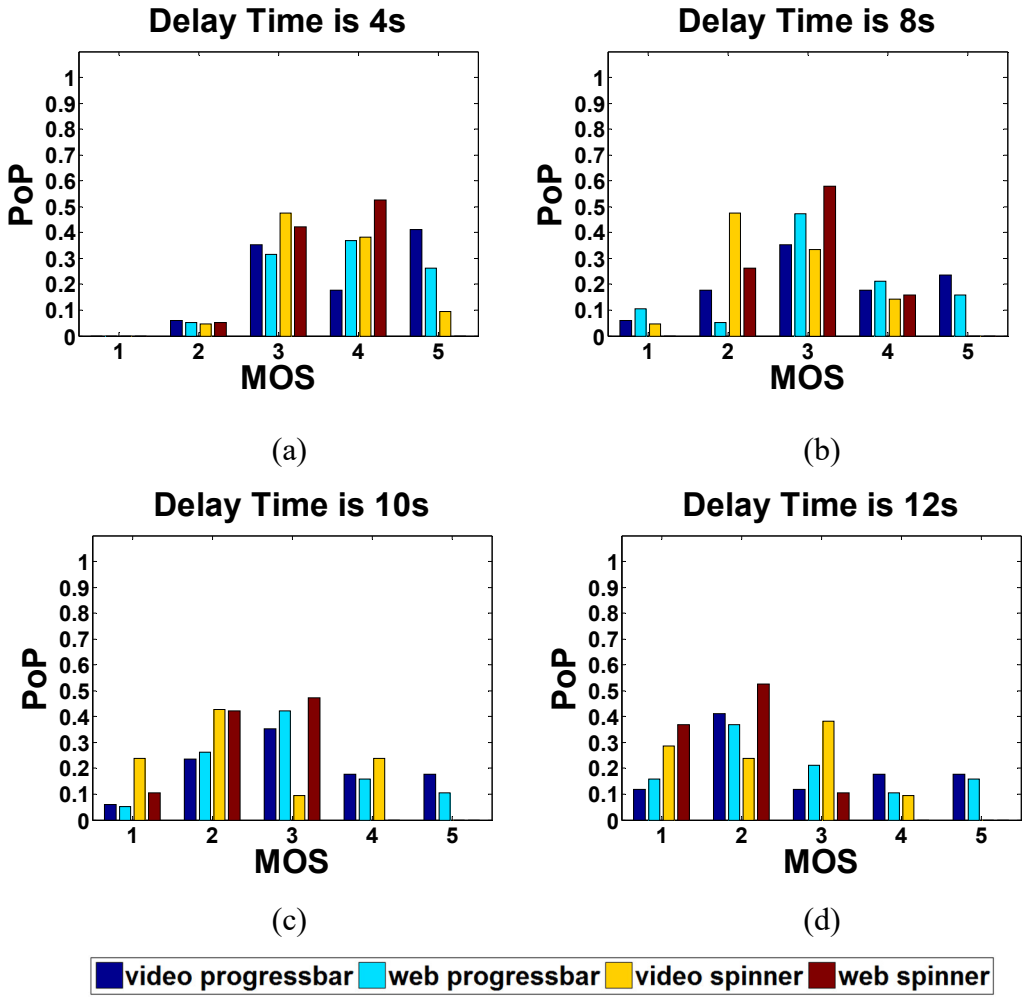


Figure 5.8 Users' rating diversity as percentage of participants (PoP) of rating results in different scenarios when the waiting time is larger than 2s (i.e. acceptance starts to drop down)

application can find the correct relevant content. I can tolerate even more delay only if the linked content is correct." While some other participants who used the application with the progress bar said: "The process is really fast. I think the maximum delay is only around 5s." It seems that there are another two elements that take effect on the user's perception: 1) the accuracy of the linked content; 2) the user's perception of time is distracted because of the behavior of capturing the image. These two elements don't exist in click-based applications, because the link is fixed beforehand and the click behavior is instantaneous. These are new issues for

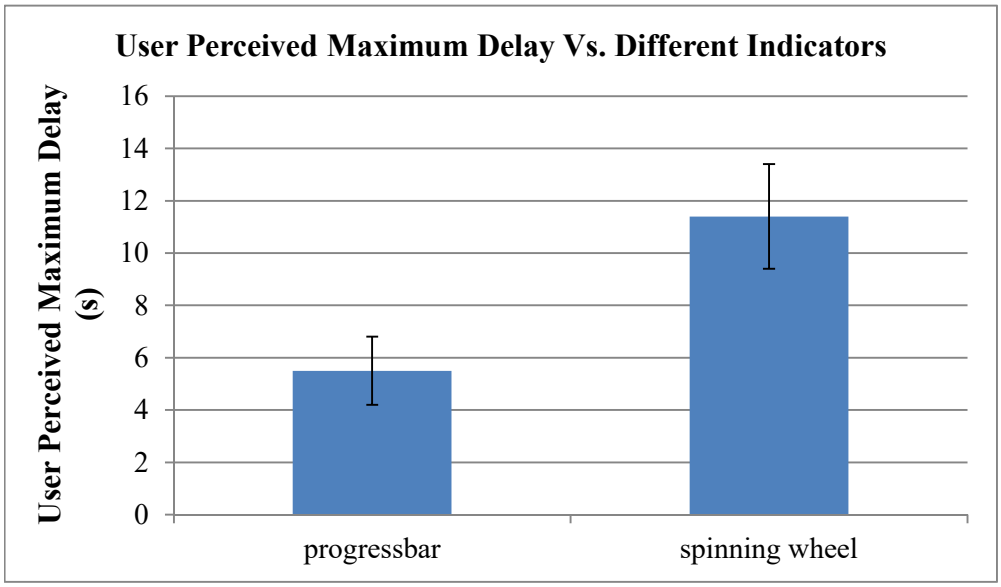


Figure 5.9 The diversity of user perceived maximum waiting time from the answer of "What is the maximum delay you have experienced during the test?"

emerging MAVS applications. It is apparent that the accuracy of the results will directly influence the user's expectation.

In addition, the result of the post interview question "What is the maximum delay you have experienced during the test?" is illustrated in Figure 5.9. It is interesting that the majority of participants think that the maximum delay is around 5s when the progress bar is used. On the contrary, the participants feel a longer delay of approximately 11s when the indeterminate spinning wheel is used. It is indicated that the different progress indicators do influence the users' perception of waiting time and thus influence the QoE perceived by the users.

5.2.10 Conclusion

The perceived QoE due to WT is studied by comparing the targeted MAVS applications and the conventional retrieval-by-click applications in a mobile test platform. The influencing factors of the linked content type and progress indicators are discussed. The content type does not have a significant impact on user's

perception of WT while the different progress indicators do have a significant influence. Based on examining the variability of user ratings and post interview questions, the user's perception in terms of WT is not only influenced by the user's expectation but also influenced by other context-based elements, such as the accuracy of the linked content and the user interface between the application and the users.

5.3 QoE prediction for proposed MAVS system

5.3.1 Overview and novelty

The ultimate goal of this work is to ensure the quality of experience in the emerging MAVS applications which enhance a user's experience by linking printed media to digital content such as a video or webpage as investigated so far. The system diagram associated with the waiting time is shown in Figure 5.10 and the overall waiting time (WT) begins when a user first begins to 'scan' a printed picture with their mobile device camera and ends when they receive the related digital content. In Figure 5.10, the captured video stream is processed on a frame-by-frame basis, with times for each part of the process as indicated. Generally, the capture and query process induces a user to experience a certain waiting time, which is critical to the user perceived QoE. During this procedure, the waiting time is mainly determined by several factors.

Firstly, the most time-consuming part of the process is the analysis of the captured image to generate query information at the client side on the mobile device. The speed of this process highly depends on the employed algorithms and the computational capacity of the device, which induce the time $T_{process}$.

Secondly, the process of searching and matching the relevant media content

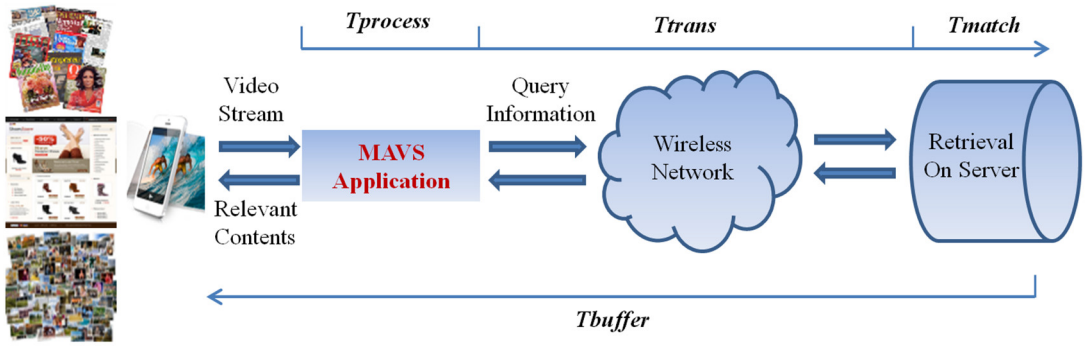


Figure 5.10 System diagram of MAVS applications associated with waiting time

in the server is another time-consuming procedure depending on the retrieval method, which induce the time T_{match}. As the accuracy of feature algorithms and matching methods increase, so too does the computational complexity, memory resources that are required. Thus, the T_{match} will increase correspondingly.

Thirdly, the transmission capacity of the wireless network may be limited. However, more transmission data sizes that are required [230] for more accurate retrieval result. As a result, the transmission time T_{trans} increases.

Also indicated is the buffer time at the client, T_{buffer}, before received digital content (e.g. video) begins to be displayed to a user.

It is of importance to consider the WT and matching accuracy (i.e. the retrieved content is correctly corresponding to the captured images) at the same time to maximize the QoE perceived by users. The WT should be as short as possible [173]. As found in the previous study, too long a waiting time can cause a user to become anxious and move the camera to an inappropriate position and capture an irrelevant content that will not match one of the predefined images and then exacerbates the problem of finding the correct correspondence. It is a dilemma to achieve high matching accuracy meanwhile keeping waiting time as fast as possible in the targeted MAVS applications. Therefore, QoE estimation is needed to help

application developers to consider the tradeoffs of different MAVS system architecture, different algorithms, and varying transmission data size to ensure that the QoE perceived by users is maximized.

In this work, QoE estimations for low bitrate and accurate MAVS systems proposed in Chapter 4 is analyzed based on the waiting time and matching accuracy as judged by retrieval experiments on a realistic image dataset with real-world distortions caused by the image capture. A Comparison with the start-of-the-art feature selection in MPEG-7 CDVS, which aims to develop high matching accuracy and low transmission solution for visual search applications, is presented from the aspects of ensuring good QoE to users. The prediction methodology is presented in the next section.

5.3.2 QoE prediction model based on Bernoulli trials

A QoE estimation method considering the matching accuracy and waiting time is presented based on the assumption of Bernoulli process. The matching accuracy and waiting time are both influenced by the size of data to be processed and transmitted. For each captured frame, after the processing to generate the query information, a certain amount of data is transmitted to the server to perform the search and matching. Each frame has a probability to be correctly matched or not according to the transmitted data and the employed search and matching algorithms. As performing the search and matching frame-by-frame, the matching result of each frame is considered to follow an identical and independent distribution. Therefore, considering the problem of finding a match (i.e. success or not) for each frame as Bernoulli trials [239], it is assumed that the probability that the first occurrence of successful match requires M number of frames, each with success probability P (i.e.

precision @ 1) depend on the employed algorithms. Then, the probability of finding a match after processing M frames is:

$$\mathrm{P}_{\mathrm{correct}} = 1 - (1 - \mathrm{P})^{\mathrm{M}} \tag{5.1}$$

Therefore, the frame M which makes $\mathrm{P}_{\mathrm{correct}} = 1$(i.e. 100% match) can be calculated by the precision @ 1 of employed algorithms in a MAVS system. The whole waiting time WT associated with the matching accuracy in the MAVS system can be defined as:

$$WT = M * \left(T_{process} + T_{trans} + T_{match} \right) + T_{buffer} \tag{5.2}$$

For each frame, the processing time $T_{process}$ is device-dependent and mainly determined by the computational capacity of the client device and the content of captured frame. The feature transmission time T_{trans} is inversely proportional to transmission bitrate C_{link} for certain transmission load L_{bit}. The matching time T_{match} is mainly determined by the computational capacity of the server and considered insignificant compare to other times assuming a powerful server with multicore CPU and GPU acceleration in this work (i.e. T_{match} is equal to 0) [240]. The buffer time T_{buffer} is configurable according to different players and conditions [241]. To isolate the effect of buffer time, a buffer time of 0.5s is considered for streaming related video content as previous research has indicated that such waiting time provides a satisfactory QoE [77]. The frame rate M, link capacity C_{link} and the transmission load L_{bit} are considered as key influencing factors of QoE in terms of the waiting time and matching accuracy in this work. In practice, a cutoff threshold of waiting time WT is needed when no match can be found after a certain waiting time (i.e. multiple consecutive frames from the camera 'scan' do no match any image in the database). In this case, the process should be stopped and a feedback is given to

the users. (5.2) can be redefined as:

$$WT = \begin{cases} M * \left(T_{process} + \frac{L_{bit}}{C_{link}} \right) + 0.5, a\ match\ is\ found \\ 2,\ \ no\ match\ is\ found \end{cases} \quad (5.3)$$

2s is chosen as a cutoff threshold of waiting time to ensure satisfactory QoE according to the evaluation in Section 5.2. C_{link} ranges from 50kbps to 4800kbps as considering in an typical 3G/4G wireless network [242]. Considering a MAVS system which links video content and uses a progress bar as progress indicator, a QoE function (5.4) derived from the subject experimental results of Table 5-2 is employed to map the waiting time WT to QoE:

$$QoE(t) = -1.118\, ln(t + 1.648) + 5.864 \quad (5.4)$$

Substituting t with (5.3) in (5.4) when assuming a match can be found, the predicted QoE related to the waiting time and matching accuracy (i.e. M frame to achieve 100% match) can be defined as:

$$\mathrm{QoE}(\mathrm{M}, \mathrm{L_{bit}}, \mathrm{C_{link}}) = -1.118 * \ln\left(\mathrm{M} * \left(\mathrm{T_{process}} + \frac{\mathrm{L_{bit}}}{\mathrm{C_{link}}} \right) + 2.148\right) + 5.864 \quad (5.5)$$

It is noted that M, $\mathrm{T_{process}}$ and $\mathrm{L_{bit}}$ vary in different MAVS system depend on the employed algorithms, devices and transmission bitrate.

In the following subsections, a test platform with quad-core 1.6GHz CPU and 2G RAM is used to simulate the computational capacity of a current state-of-the-art smart phone [243] to estimate $T_{process}$ varies with different algorithms. Three different algorithms are evaluated in the context of a MAVS system associated with the waiting time and matching accuracy from the point view of QoE. These three algorithms are peak-based feature selection in MPEG-7 CDVS, two proposed low bit

rate MAVS algorithms in Chapter 4 based on low frequency DCT coefficients and relevance-based feature selection, respectively. To obtain the matching accuracy and waiting time, the printed media images from the MVS dataset [194] are used to perform image retrieval. The dataset contains more than 1200 camera-phone captured different types of print images including CD covers, DVD covers and book covers. These images are denoted as query images, which contain images with widely varying lighting conditions, perspective distortion, foreground and background clutter. The ground-truth reference images are also available and used for training. These ground-truth images are denoted as reference images. The matching accuracy of different algorithms is reported using precision @ 1 meanwhile the waiting time is recorded when performing the retrieval experiment. And then, the recorded matching accuracy and waiting time are used with aforementioned QoE prediction model to estimate the QoE.

5.3.3 QoE prediction result for the peak-based feature selection in MPEG-7 CDVS

The output parameters including the Difference-of-Gaussian (DOG) response θ_{peak}, scale θ_{scale}, orientation $\theta_{orientation}$, location $\theta_{distance}$ and their combination $\theta_{combination}$ are employed in the MPEG-7 CDVS for feature selection using the probability mass function of these relevance metrics learned from correctly matched features pairs in the dataset. The θ_{peak} and $\theta_{combination}$ is superior for identifying the most relevant features compared to other parameters of the output of SIFT detector, including $\theta_{orientation}, \theta_{scale}, \theta_{distance}$ [167], [168]. Thus, the peak-based feature selection using θ_{peak} and $\theta_{combination}$ is chosen to investigate in this work. The retrieval experiment utilizes the same experimental architecture and procedure as

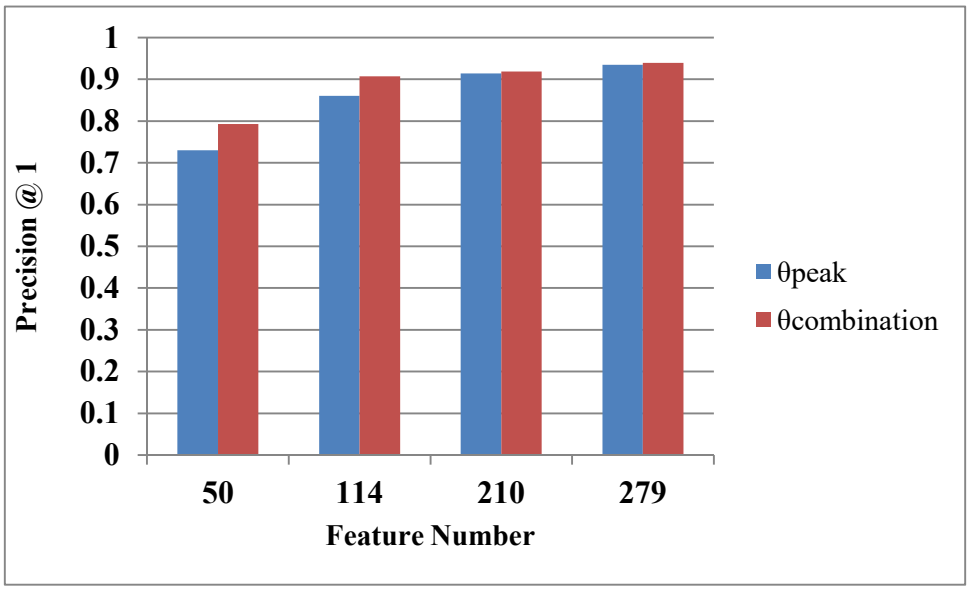


Figure 5.11 The retrieval results of using peak-based feature selection in MPEG-7 CDVS under varying feature number (i.e. 279, 210, 114 and 50 feature numbers correspond to 2KB, 1KB, 512B and 256B compressed feature transmission sizes).

described in Section 4.4.10. Four different feature number conditions are considered in the experiment 279, 210, 114 and 50 which correspond to 2KB, 1KB, 512B and 256B compressed feature transmission sizes. The first three bit rates are standardized in the MPEG-7 CDVS [152]. The fourth bit rate is also considered in the scenario of a very poor communication condition or processing condition where a very fast transmission is desired (e.g. processing a stream of video frames to repeatedly look for a matching reference image). Therefore, the frame number M which leads to 100% match using the assumption of Bernoulli trials in (5.1) can be calculated by the

Table 5-3 The number of frame M which makes $P_{correct} = 1$ using (5.1) when using the peak-based feature selection in MPEG-7 CDVS

Feature Number	50	114	210	279
L_{bit}	256B	512B	1KB	2KB
M of θ_{peak}	10	7	6	5
M of $\theta_{combination}$	8	6	5	5

retrieval precision @ 1 of the peak-based feature selection as shown in Figure 5.11. The result of frame amount M is shown in Table 5-3.

When using the peak-based feature selection, the process time $T_{process}$ consists of T_{ex} and T_{sele}, which are the feature detection and extraction time and the feature selection time, respectively. The feature extraction time T_{ex} of the selection methods using $\theta_{peak}$and $\theta_{combination}$ are the same as the feature extraction time T_{ex} is device-dependent and mainly determined by the computational capacity of the client device and the content of captured frame. The maximum variation of T_{ex} across the whole dataset in the test platform is only 4ms. Thus, the average $T_{ex} = 0.138ms$ is used both for θ_{peak} and $\theta_{combination}$ selection methods. The feature selection time T_{sele} is related to feature number N and mainly depends on the

Table 5-4 The predicted QoE when using the peak-based feature selection in MPEG-7 at varying feature number and transmission bitrate according to (5.5)

	Selection methods	Feature number	Transmission bitrate (kbps)					
			50	200	500	1000	1500	2000
Predicted QoE	θ_{peak}	50	4.31	4.39	4.41	4.42	4.42	4.42
		114	4.39	4.52	4.55	4.56	4.56	4.56
		210	4.32	4.54	4.59	4.60	4.61	4.61
		279	4.19	4.53	4.62	4.65	4.66	4.66
	$\theta_{combination}$	50	4.41	4.48	4.50	4.51	4.51	4.51
		114	4.45	4.57	4.60	4.61	4.61	4.61
		210	4.40	4.60	4.64	4.66	4.66	4.66
		279	4.18	4.53	4.61	4.64	4.65	4.66

selection algorithms and the computational capacity of the client device. The T_{sele} of the selection methods using $\theta_{peak}$and $\theta_{combination}$ both vary with different feature number but the maximum variation across the whole dataset is only 2ms and 1.8ms, respectively. Therefore, the average selection time of $avgT_{sele}^{peak} = 10ms$ and $avgT_{sele}^{combination} = 13ms$ are used. The feature transmission time T_{trans} is inversely proportional to transmission bitrate C_{link} for certain transmission load $L_{bit}(N)$ (i.e. 256B, 512B, 1KB, 2KB corresponds to 50, 114, 210, 279 features, respectively). The predicted QoE results by using (5.5) and Table 5-3 under varying L_{bit} and C_{link} are shown in Table 5-4. The results of transmission bitrates above 2000kbps were truncated as the predicted QoE results show a flat trend beyond these transmission bitrates. The results suggest that the peak-based feature selection in MPEG-7 CDVS can achieve good QoE results (i.e. above 4 as shown in Table 5-4) in the context of a MAVS system at all transmission conditions. The θ_{peak} and $\theta_{combination}$ achieved the similar performance in terms of the predicted QoE because despite the $\theta_{combination}$ achieved better precision @ 1 than θ_{peak}, a little more feature selection time was required by $\theta_{combination}$ method. To illustrate the key results, the predicted QoE of θ_{peak} at varying transmission bitrate and load (i.e. varying feature number) is shown in Figure 5.12. The transmission bitrate C_{link} and the transmission load $L_{bit}(N)$ only have a minor effect on the predicted QoE at the low transmission bitrate for the transmitting feature number 210/279 because of the increased transmission load. It is obvious that it is better to transmit fewer feature number at the low transmission condition because the increased transmission delay becomes more significant than increased processing delay due to more frames. At the high transmission condition, the predicted QoE is slightly decreased with the reduction of feature number as the matching accuracy is decreased. It indicates that it is better to transmit more features

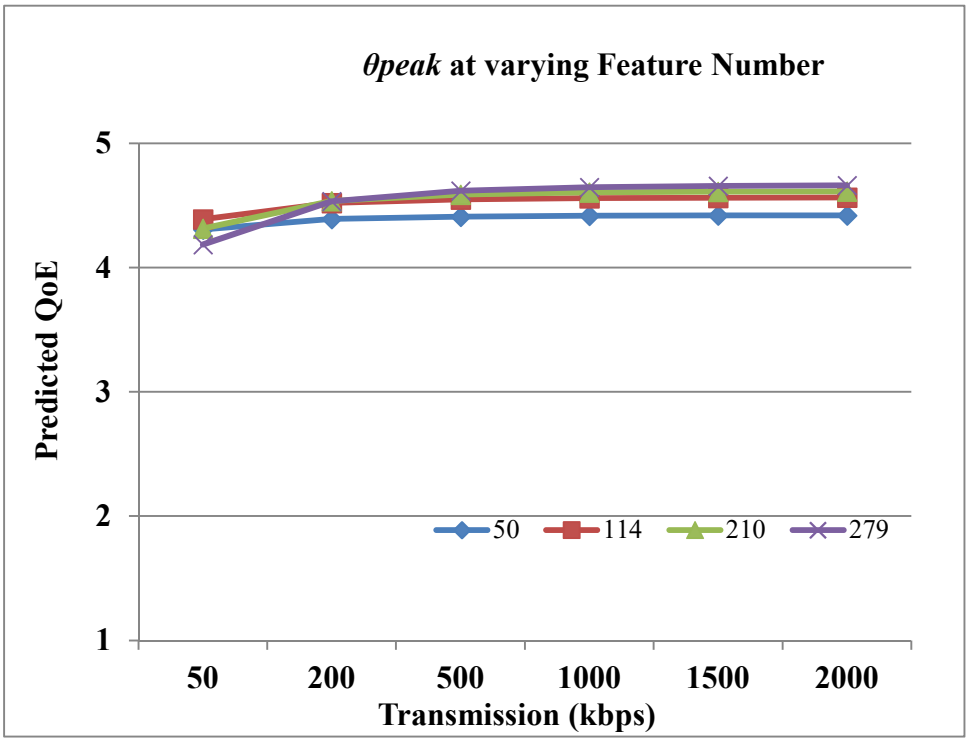


Figure 5.12 The predicted QoE for Peak-based feature selection method of using θ_{peak} under 50kbps~2000kbps bitrate and varying feature number

under high transmission condition. It is not only beneficial to ensuring matching accuracy but also makes the system work under low frame condition which can reduce the computation consumption. Besides, the low frame rate system can reduce the probability that users move the camera to an inappropriate position and capture an irrelevant content.

5.3.4 QoE prediction result for the MAVS system using the relevance-based feature selection

The relevance-based feature selection methods of using local patch entropy θ_{LPE} in the keypoint domain, descriptor entropy θ_{DE} in the descriptor domain, DCT coefficients θ_{AC1} and θ_{AC2} in the compressed domain and their combination θ_{LDAC} are proposed in Section 4.3. The selection methods of θ_{DE} and θ_{LDAC} are chosen to study from the point view of QoE as they achieved better matching accuracy than the other

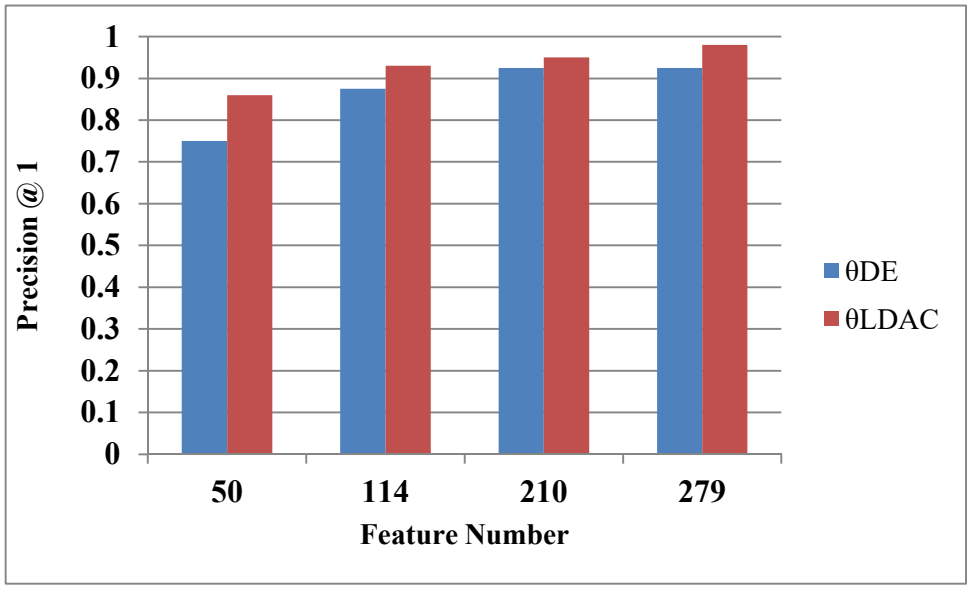


Figure 5.13 The retrieval results of using relevance-based feature selection θ_{DE} and θ_{LDAC} under varying feature number (i.e. 279, 210, 114 and 50 feature numbers correspond to 2KB, 1KB, 512B and 256B compressed feature transmission sizes).

methods. The same dataset and experimental architecture as used in previous section are used to get the precision @ 1 of θ_{DE} and θ_{LDAC} as shown in Figure 5.13. And then, the frame number M of using θ_{DE} and $\theta_{LDAC}$is calculated by (5.1) as shown in Table 5-5.

Similar to the peak-based feature selection method, the process time $T_{process}$ of the relevance-based feature selection can also be divided to the feature detection and extraction T_{ex} and the feature selection time T_{sele}. As using the same feature

Table 5-5 The number of frame M which makes $P_{correct} = 1$ using (5.1) when using the relevance-based feature selection θ_{DE} and θ_{LDAC}

Feature Number	50	114	210	279
L_{bit}	256B	512B	1KB	2KB
M of θ_{DE}	9	6	5	5
M of θ_{LDAC}	7	5	5	4

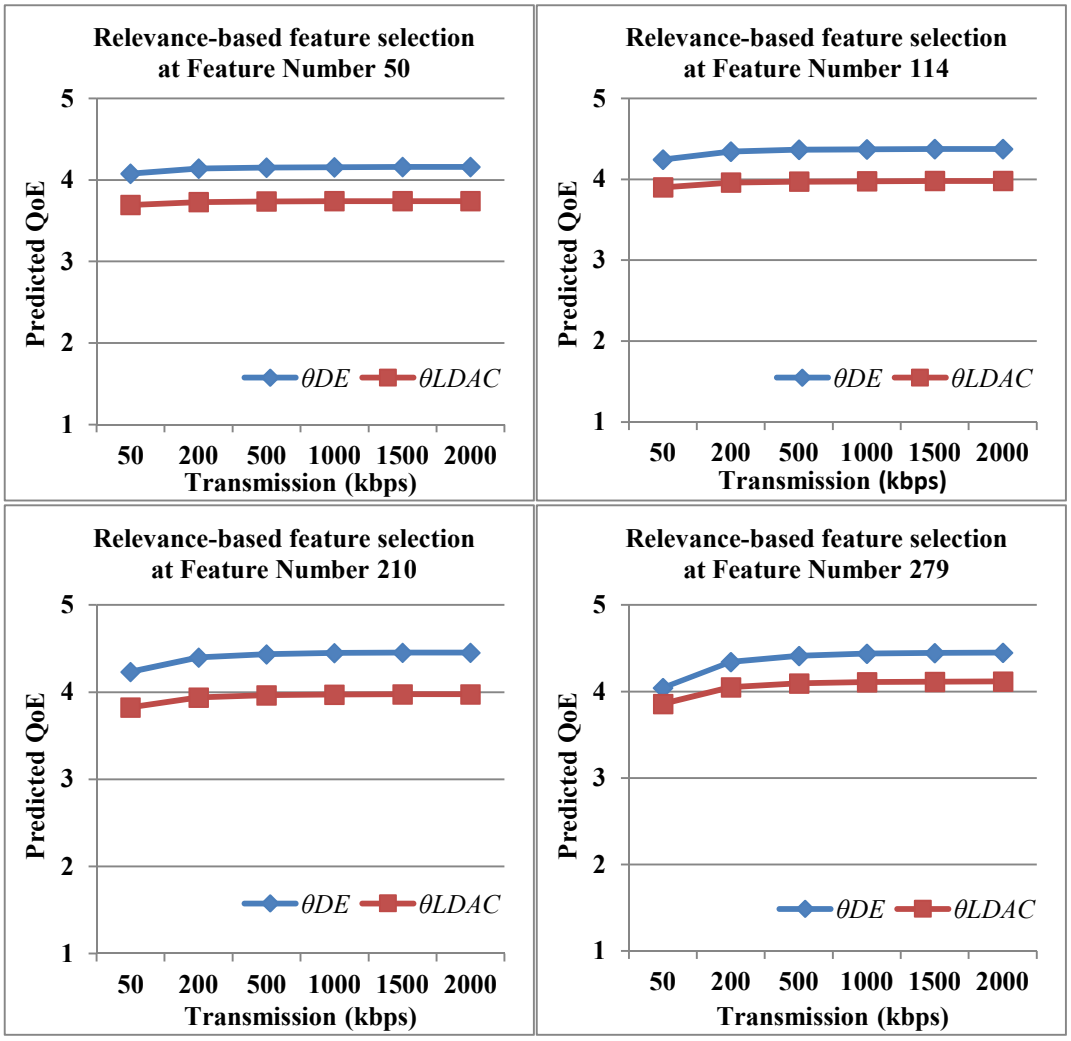


Figure 5.14 The predicted QoE of relevance-based feature selection under 50kbps~2000kbps bitrate and varying feature number.

detector and descriptor extractor and the same dataset, the feature extraction time T_{ex} of using θ_{DE} and θ_{LDAC} are nearly the same as θ_{peak} and $\theta_{combination}$ with only less than 1.4ms variation. Thus, the same average $T_{ex} = 0.138ms$ is used both for the θ_{DE} and θ_{LDAC} selection methods. The feature selection time T_{sele} is different because more time is required to calculate the relevance metrics θ_{DE} and θ_{LDAC} for each feature, which results in longer T_{sele}. The average selection time of $avgT_{sele}^{DE} = 133ms$ and $avgT_{sele}^{LDAC} = 510ms$ are used for QoE estimation. The predicted QoE results present the similar trend along with the variation of transmission bit rate and

the feature number as the peak-based feature selection in Figure 5.14. The relevance-based feature selection method of using θ_{DE} shows the slightly better QoE than θ_{LDAC}. Although the θ_{LDAC} achieved better precision @ 1 than the θ_{DE} and was required fewer frame number, the predicted QoE is significantly influenced by the longer selection time. It is noted that the θ_{LDAC} might achieve better QoE if the selection time could be reduced, for example, a more powerful client device was employed.

5.3.5 QoE prediction result for MAVS system using low frequency DCT coefficients

The MAVS system using the low frequency DCT coefficients as proposed in Section 4.2 employed a different system architecture compared to the previous two systems using feature selection technologies. The feature detection and extraction is shifted to the server side and no feature selection technology is employed. In sequence, the feature detection and extraction can achieve real-time performance in a powerful server equipped with multicore CPU and GPU acceleration [244]–[246]. Therefore, the process time $T_{process}$ is mainly determined by the time of extracting and compressing the low frequency DCT coefficients T_c. To obtain the T_c and precision @ 1 of the MAVS system using the low frequency DCT coefficients, the experimental architecture proposed in Section 4.3 is employed to perform the retrieval experiment on the MVS dataset with varying number of low frequency DCT

Table 5-6 The number of frame M which makes $P_{correct} = 1$ using (5.1) when using varying number of low frequency DCT coefficients.

DCT Coefficients	DC	DC2AC	DC3AC	DC8AC
Frame Number M	13	11	6	5

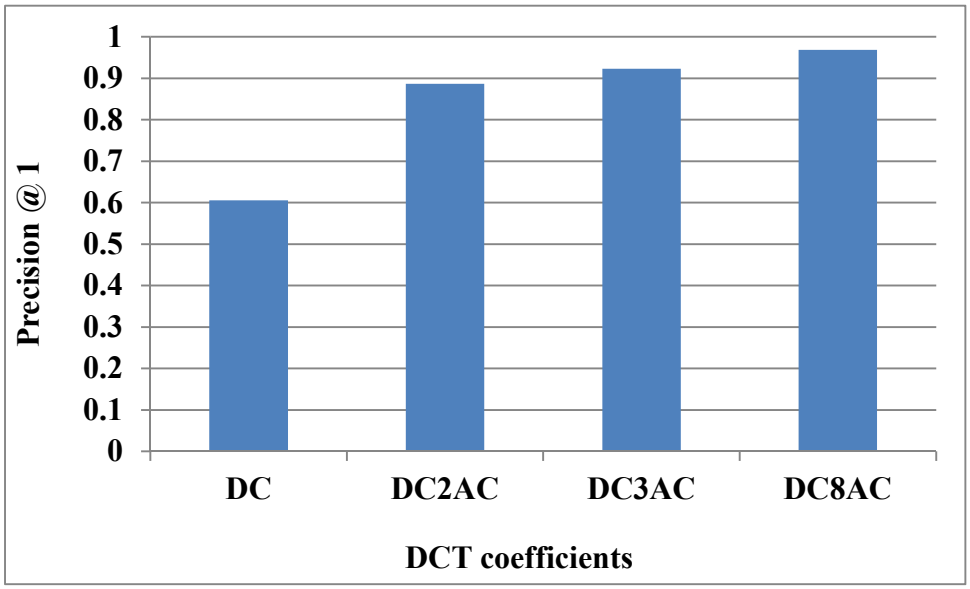


Figure 5.15 The retrieval results of using varying number of low frequency DCT coefficients.

coefficients (i.e. DC coefficients only, DC coefficients + first 2 AC coefficients, DC coefficients + first 3 AC coefficients and DC coefficients + first 8 AC coefficients). The average T_c across the whole MVS dataset is 6ms with 0.45ms variation, which is used for the QoE estimation. The precision @ 1 of using varying number of low frequency DCT coefficient is shown in Figure 5.15. The corresponding frame number according to the results in Figure 5.15 is calculated by (5.1) as shown in Table 5-6. The predicted QoE results using (5.5) decline logarithmically with the decrease of transmission bit rate as shown in Figure 5.16. The predicted QoE results are higher than 4 when transmission bit rate is greater than 200kbps. Although the DC3AC and DC8AC provided the better precision @ 1 than the DC and DC2AC, they required more bandwidth for transmission which significantly reduces the QoE at low transmission network condition of 50kbps. By comparing the results in Figure 5.12, Figure 5.14, and Figure 5.16, the MAVS system of sending the DC coefficients achieved the best predicted QoE when the transmission bitrate was above 500kbps

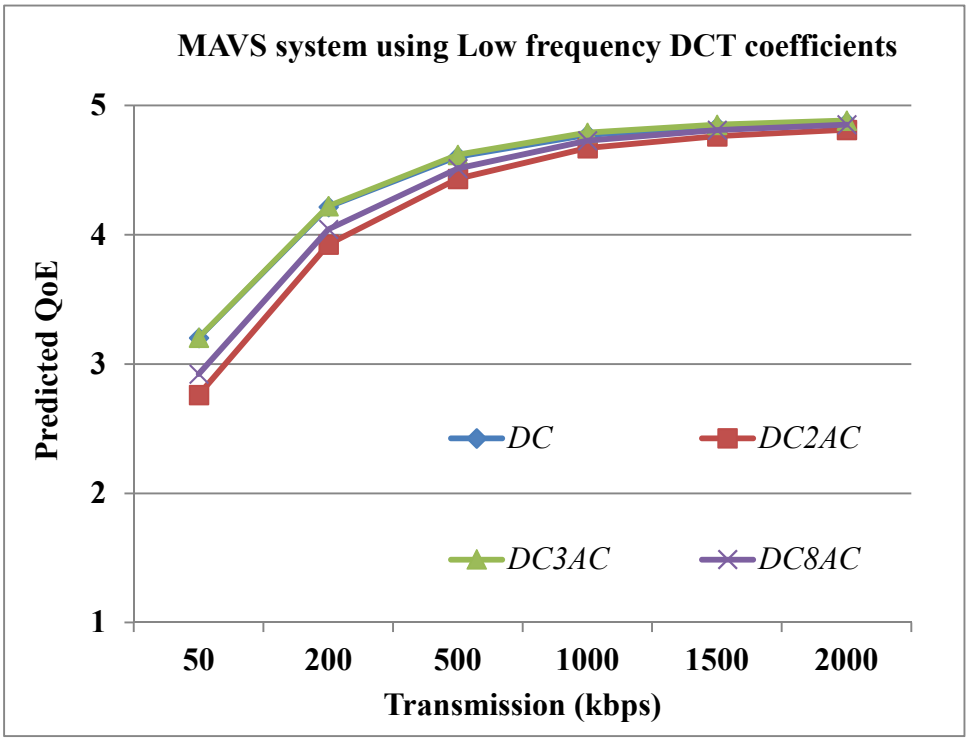


Figure 5.16 The predicted QoE of using low frequency DCT coefficients under 50kbps~2000kbps bitrate

while the MAVS system of using θ_{DE} and θ_{peak} achieved better QoE at the low transmission bitrate below 200kbps under the assumption of Bernoulli trials on the MVS dataset.

5.4 Conclusion

The influence of two key impact factors known as waiting time and matching accuracy is investigated from the point view of QoE in this chapter. Waiting time, as a directly perceptible key influencing factor to users, was studied by conducting a subjective test on a Samsung Galaxy Tab with a specific developed application simulated the whole procedure of targeted MAVS applications. 51 participants were invited to attend the test to operate the mobile phone camera in video mode and then 'scan' over a printed picture on a page to find the corresponding matching image within a database of unique images to trigger the augmented content (i.e. a video and

a webpage). The subjective test results suggested that the QoE perceived by the users had a logarithmic function of waiting time and the linked content had little influence on the resulting QoE while the different progress bars had an effect on users' perception in terms of waiting time and consequently aroused different QoE. Then, a QoE model is proposed based on the waiting time and matching accuracy to evaluate the QoE perceived by users. This model can help system designers to make decisions to choose different technologies/parameters. Moreover, the proposed QoE model is employed to evaluate the QoE of the proposed MAVS applications using different algorithms. Three different MAVS systems of employing peak-based feature selection, relevance-based feature selection and low frequency DCT coefficients were studied from the aspects of waiting time and matching accuracy in terms of QoE by using the results of subjective experiment and a retrieval experiment on a realistic image dataset with real-world distortions caused by image capture. The QoE estimation results suggest that the peak-based feature selection in MPEG-7 CDVS, the proposed relevance-based feature selection and the low frequency DCT coefficients methods can provide good QoE to users.

6 CONCLUSIONS AND FUTURE WORK

6.1 Conclusions

Mobile augmented visual search applications have been investigated in this thesis. The MAVS applications are motivated by the prosperous development of digital products and services which are continuously evolving and have dramatically changed the way people interact with multimedia content. The MAVS applications are based on the images or a short video clip captured by user and then automatically matching a relevant image in a predefined database to trigger corresponding augmented multimedia content to users. Focused on maximizing the QoE in such type of applications, several investigations are conducted from the aspects of overcoming real world distortions (e.g. partial occlusions, lighting conditions, motion blurring), limitation of mobile devices' capacity, constraint of network bandwidth and minimising the time for users to begin receiving linked digital content. On the basis of the investigations, effective ways to manage key influencing factors related to QoE, such as matching accuracy and waiting time, are developed to ensure the user QoE is maximized by using a MAVS application that has high accuracy, low bitrate transmission requirements and low latency. The proposed MAVS application can be used to link different types of images, such as book covers, CD covers, DVD covers, museum painting images, newspaper images and natural scene images.

The following main conclusions from this thesis are drawn:

- To find the most efficient feature and the influence of codec distortion to address research questions 1 and 2 of Section 2.4.4, the performance of different local feature algorithms is investigated under various image compression rates from the aspects of matching accuracy and processing

time as well as trade-off between transmitting entire images and transmitting image features in Section 3.2 and 3.3. Flexible feature selection or combinations of distinctive feature algorithms can be employed in the server or cloud regardless of computational constraint in the client side to improve the retrieval performance.

- To study the influence of real world distortions and address research question 2 of Section 2.4.4, the evaluation of joint effect of illumination changes and image blurring in the context of the MAVS application is performed in Section 3.4 to Section 3.7. The performance of various feature algorithms is investigated under various joint distortions with two different cameras from the aspect of matching accuracy. Illumination changes have more influences on matching accuracy compared to image blurring for the studied combinations of local feature algorithms under tested image datasets. Different cameras also affect the performance of different combinations of local feature algorithms.

- To find an efficient way to process the most significant information to achieve low bit rate transmission and low computation and address research question 3 of Section 2.4.4, a new MAVS system proposed in Section 4.3 is based on SIFT features derived from DC coefficients in the 2D block-based DCT domain to enable a low complexity, fast and accurate implementation on mobile devices whilst requiring low bit rate transmission and using a powerful remote server to accelerate the most time-consuming processing. The method achieves more than 97% matching accuracy while reducing the transmission bandwidth requirement by more than 97%, whilst reducing client side processing time by

approximately 50% compared to an existing low-bit rate feature matching system. The DC image can be reduced to half scale to further reduce transmission bandwidth under poor transmission situations without obvious loss of matching accuracy. Alternatively, the JPEG encoder can use a customized quantization table to discard the AC coefficients. There is no need to modify current image codecs in mobile devices.

- To select the most important feature and minimize the feature computation and feature transmission to further solve research question 3, novel methods for relevance-based feature selection are proposed in Section 4.4 on the basis of selecting a subset of robust detected features to correctly match a captured image to a reference image and transmit at low bitrate to retrieve an augmented multimedia content from a remote server accurately. The discriminative information embedded in the entropy of local image patch, entropy of descriptor and DCT coefficients are found to be efficient and sufficient for feature selection. The proposed methods achieve superior image retrieval performance on a dataset with complex realistic distortions, particularly at low bit rates.

- To find the influence of waiting time in terms of QoE and address research question 4 of Section 2.4.4, the perceived QoE due to waiting time is studied by comparing MAVS applications and conventional retrieval-by-click applications within the mobile image matching system in Section 5.2. The QoE influencing factor of linking content type does not have a significant impact on user's perception of waiting time while the different progress indicators do have a significant influence. The user's perception of waiting time for perceived QoE is not only influenced by the user's

expectation but also influenced by other context-based elements, such as the accuracy of the linked content.

- To estimate the users perceived QoE in the proposed MAVS systems, a QoE estimation method based on waiting time and matching accuracy is studied on the assumption of Bernoulli trails by performing retrieval experiments on a realistic image dataset with real-world distortions caused by image capture. The predicted QoE proves that the proposed MAVS systems can provide good QoE to users under varying transmission conditions. However, how to guarantee the QoE in the targeted MAVS system in real-time requires further study.

.

6.2 Future work

With the evolution of technology and the emergence of new mobile devices, the MAVS applications are developing fast. New challenges are coming out to ensure the QoE in the new generation MAVS applications. The possible future work based on this thesis includes:

- Investigation into compressing the image further whilst maintaining the matching accuracy by combining template matching and ROI features of some advanced encoder, such as JPEG2000 and HDPhoto.

- Investigation of incorporating the other image matching technologies to improve the matching accuracy in the proposed MAVS systems, e.g. color histogram comparison.

- Investigation into more comprehensive joint distortions, such as shadow, newspaper wrinkle, gloss paper reflection to identify the most significant

negative effects and how to overcome these effects in real time mobile applications by using image enhancement and normalization technologies.

- Investigation of more state-of-art local feature algorithms and extend current work from DCT domain to other transforms, for example, wavelet transform or lapped transform [247].
- Extending the feature selection work to embed feature selection in the feature detection and extraction stage to accelerate the processing speed.
- Investigation of the possibility of employing a deep learning network to accurately recognize the captured images or video in real time to avoid time consuming feature detection, extraction, matching and geometric verification.
- Investigation of global image signatures for efficient retrieval in a MAVS system when encountering large scale database or on-device-aided search by maintaining a very compact local index database.
- Extending the key influencing factors of QoE study in the context of evolving MAVS applications, for example linking multimedia content or 3D content on wearable devices.
- Extending the QoE estimation model from Bernoulli Trail to other models, for example, Markov process.

REFERENCES

[1] J. Johnson, “The Master Key: L. Frank Baum envisions augmented reality glasses in 1901,” *Mote Beam*, 2013.

[2] I. E. Sutherland, “A head-mounted three dimensional display,” in *Proceedings of the December 9-11, 1968, fall joint computer conference, part I*, 1968, pp. 757–764.

[3] L. B. Rosenberg, “Virtual fixtures as tools to enhance operator performance in telepresence environments,” in *Optical Tools for Manufacturing and Advanced Automation*, 1993, pp. 10–21.

[4] S. Feiner, B. Macintyre, and D. Seligmann, “Knowledge-based Augmented Reality,” *Commun ACM*, vol. 36, no. 7, pp. 53–62, Jul. 1993.

[5] F. J. Delgado, M. F. Abernathy, J. White, and W. H. Lowrey, “Real-time 3D flight guidance with terrain for the X-38,” in *AeroSense’99*, 1999, pp. 149–156.

[6] H. Kato and M. Billinghurst, “Marker tracking and hmd calibration for a video-based augmented reality conferencing system,” in *Augmented Reality, 1999.(IWAR’99) Proceedings. 2nd IEEE and ACM International Workshop on*, 1999, pp. 85–94.

[7] D. Wagner and D. Schmalstieg, “First steps towards handheld augmented reality,” in *Seventh IEEE International Symposium on Wearable Computers, 2003. Proceedings*, 2003, pp. 127–135.

[8] “NMC Horizon Report 2011 Higher Ed Edition | The New Media Consortium.” [Online]. Available: http://www.nmc.org/publications/horizon-report-2011-higher-ed-edition. [Accessed: 30-Oct-2014].

[9] J. Luo, D. Joshi, J. Yu, and A. Gallagher, “Geotagging in multimedia and computer vision—a survey,” *Multimed. Tools Appl.*, vol. 51, no. 1, pp. 187–211, 2011.

[10] “Photos Around - Android Apps on Google Play.” [Online]. Available: https://play.google.com/store/apps/details?id=com.mandreasson.photosaround. [Accessed: 31-Oct-2014].

[11] “HOTELS.COM – SEARCH FOR HOTELS - Wikitude augmented reality sdk.” [Online]. Available: http://www.wikitude.com/showcase/hotels-com-search-for-hotels/. [Accessed: 31-Oct-2014].

[12] “andar - AndAR - Android Augmented Reality - Google Project Hosting.” [Online]. Available: https://code.google.com/p/andar/. [Accessed: 30-Oct-2014].

[13] Google goggles,http://www.google.com/mobile/goggles/#text, .

[14] “viewa,” *viewa*. [Online]. Available: http://viewa.net/. [Accessed: 14-May-2014].

[15] “Home | Augmented Reality | Interactive Print,” *Layar*. [Online]. Available: https://www.layar.com/. [Accessed: 14-May-2014].

[16] R. O’Grady, “Fairfax Launches AirLink,” *Business 2*. .

[17] “ITU-T SG12: Performance, QoS and QoE,” *ITU*. [Online]. Available: http://www.itu.int/en/ITU-T/studygroups/2013-2016/12/Pages/default.aspx. [Accessed: 21-Aug-2015].

[18] “QUALINET – European Network on Quality of Experience in Multimedia Systems and Services | Multimedia Technology Group.” .

[19] “Compact Descriptors for Visual Search.” [Online]. Available: http://mpeg.chiariglione.org/standards/mpeg-7/compact-descriptors-visual-search. [Accessed: 21-Aug-2015].

[20] K. A. Parulski and M. Rabbani, “The continuing evolution of digital cameras and digital photography systems,” in *The 2000 IEEE International Symposium on Circuits and Systems, 2000. Proceedings. ISCAS 2000 Geneva*, 2000, vol. 5, pp. 101–104 vol.5.

[21] S. Kawamura, “Capturing images with digital still cameras,” *IEEE Micro*, vol. 18, no. 6, pp. 14–19, Nov. 1998.

[22] W. Yu, “An embedded camera lens distortion correction method for mobile computing applications,” *IEEE Trans. Consum. Electron.*, vol. 49, no. 4, pp. 894–901, Nov. 2003.

[23] C. Simon, Williem, J. Choe, I. D. Yun, and I. K. Park, “Correcting Photometric Distortion of Document Images on a Smartphone,” in *2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2014, pp. 199–200.

[24] “More than megapixels - what really counts in a smartphone camera,” *NDTV Gadgets*. [Online]. Available: http://gadgets.ndtv.com/mobiles/features/more-than-megapixels-what-really-counts-in-a-smartphone-camera-495572. [Accessed: 03-Feb-2015].

[25] A. Chowdhury, R. Darveaux, J. Tome, R. Schoonejongen, M. Reifel, A. De Guzman, S. S. Park, Y. W. Kim, and H. W. Kim, “Challenges of megapixel camera module assembly and test,” in *Electronic Components and Technology Conference, 2005. Proceedings. 55th*, 2005, pp. 1390–1401 Vol. 2.

[26] “Camera megapixels: Why more isn’t always better (Smartphones Unlocked),” *CNET*. [Online]. Available: http://www.cnet.com/au/news/camera-megapixels-why-more-isnt-always-better-smartphones-unlocked/. [Accessed: 03-Feb-2015].

[27] K. Lim, H. So, S. Kang, J. Kim, and S. Kim, “3 Megapixel Camera Signal Processor for Mobile Camera Applications,” in *13th IEEE International Conference on Electronics, Circuits and Systems, 2006. ICECS ’06*, 2006, pp. 886–889.

[28] R. C. Gonzalez, R. E. Woods, and S. L. Eddins, *Digital image processing using MATLAB*. Pearson Education India, 2004.

[29] A. Chaker, M. Kaaniche, and A. Benazza-Benyahia, “An improved image retrieval algorithm for JPEG2000 compressed images,” in *IEEE International Symposium on Signal Processing and Information Technology*, 2012.

[30] D. Edmundson and G. Schaefer, “Exploiting JPEG Compression for Image Retrieval,” in *2012 IEEE International Symposium on Multimedia (ISM)*, 2012, pp. 485–486.

[31] G. Schaefer, “Does compression affect image retrieval performance?,” *Int. J. Imaging Syst. Technol.*, vol. 18, no. 2–3, pp. 101–112, Jan. 2008.

[32] D. Edmundson and G. Schaefer, “An overview and evaluation of JPEG compressed domain retrieval techniques,” in *ELMAR, 2012 Proceedings*, 2012, pp. 75 –78.

[33] M. Hatzigiorgaki and A. N. Skodras, “Compressed domain image retrieval: a comparative study of similarity metrics,” in *Visual Communications and Image Processing*, 2003.

[34] “Independent JPEG Group.” [Online]. Available: http://www.ijg.org/. [Accessed: 06-Feb-2015].

[35] W. B. Pennebaker and J. L. Mitchell, *JPEG: Still image data compression standard*. Springer Science & Business Media, 1993.

[36] “The JPEG committee home page.” [Online]. Available: http://old.jpeg.org/index.html. [Accessed: 06-Feb-2015].

[37] S. Singh, V. Kumar, and H. K. Verma, “Reduction of blocking artifacts in JPEG compressed images,” *Digit. Signal Process.*, vol. 17, no. 1, pp. 225–243, 2007.

[38] “HD Photo Feature Spec 1.0,” *Microsoft Download Center*. [Online]. Available: http://www.microsoft.com/en-au/download/details.aspx?id=1915. [Accessed: 06-Feb-2015].

[39] “T.832 : Information technology - JPEG XR image coding system - Image coding specification.” [Online]. Available: http://www.itu.int/rec/T-REC-T.832. [Accessed: 06-Feb-2015].

[40] S. Srinivasan, C. Tu, S. L. Regunathan, and G. J. Sullivan, “HD Photo: a new image coding technology for digital photography,” in *Optical Engineering+ Applications*, 2007, p. 66960A–66960A.

[41] “JPEG and JPEG2k Artifacts.” [Online]. Available: http://www.stat.columbia.edu/~jakulin/jpeg/artifacts.htm. [Accessed: 08-Feb-2015].

[42] M. W. Marcellin, M. J. Gormish, A. Bilgin, and M. P. Boliek, “An overview of JPEG-2000,” in *Data Compression Conference, 2000. Proceedings. DCC 2000*, 2000, pp. 523–541.

[43] M. D. Adams, *The JPEG-2000 still image compression standard*. 2001.

[44] A. Skodras, C. Christopoulos, and T. Ebrahimi, “The JPEG 2000 still image compression standard,” *IEEE Signal Process. Mag.*, vol. 18, no. 5, pp. 36–58, Sep. 2001.

[45] “Definition of Quality of Experience (QoE).” [Online]. Available: http://ties.itu.int/ftp/public/itu-t/fgiptv/readonly/Previous_Meetings/20070122_MountainView/il/T05-FG.IPTV-IL-0050-E.htm. [Accessed: 06-Nov-2014].

[46] Patrick Le Callet, Sebastian Möller and Andrew Perkis, eds., “Qualinet White Paper on Definitions of Quality of Experience (2012). European Network on Quality of Experience in Multimedia Systems and Services (COST Action IC 1003),” Lausanne, Switzerland, Version 1.2, March 2013.

[47] Recommendation E.800 (08/94), “E.800 : Terms and definitions related to quality of service and network performance including dependability.” [Online]. Available: http://www.itu.int/rec/T-REC-E.800-199408-S. [Accessed: 12-Nov-2014].

[48] ITU-T, “G.1080 : Quality of experience requirements for IPTV services.” [Online]. Available: http://www.itu.int/rec/T-REC-G.1080-200812-I. [Accessed: 06-Nov-2014].

[49] K. Kilkki, “Quality of Experience in Communications Ecosystem.,” *J UCS*, vol. 14, no. 5, pp. 615–624, 2008.

[50] S. Moller, K.-P. Engelbrecht, C. Kühnel, I. Wechsung, and B. Weiss, “A taxonomy of quality of service and Quality of Experience of multimodal human-machine interaction,” in *International Workshop on Quality of Multimedia Experience, 2009. QoMEx 2009*, 2009, pp. 7–12.

[51] D. Geerts, K. De Moor, I. Ketykó, A. Jacobs, J. Van den Bergh, W. Joseph, L. Martens, and L. De Marez, “Linking an integrated framework with appropriate methods for measuring QoE,” in *2010 Second International Workshop on Quality of Multimedia Experience (QoMEX)*, 2010, pp. 158–163.

[52] K. U. R. Laghari and K. Connelly, “Toward total quality of experience: A QoE model in a communication ecosystem,” *IEEE Commun. Mag.*, vol. 50, no. 4, pp. 58–65, Apr. 2012.

[53] “P.800 : Methods for subjective determination of transmission quality.” [Online]. Available: https://www.itu.int/rec/T-REC-P.800-199608-I/en. [Accessed: 11-Dec-2014].

[54] “P.910 : Subjective video quality assessment methods for multimedia applications.” [Online]. Available: http://www.itu.int/rec/T-REC-P.910-200804-I/en. [Accessed: 11-Dec-2014].

[55] “BT.500 : Methodology for the subjective assessment of the quality of television pictures.” [Online]. Available: http://www.itu.int/rec/R-REC-BT.500-13-201201-I/en. [Accessed: 11-Dec-2014].

[56] “P.913 : Methods for the subjective assessment of video quality, audio quality and audiovisual quality of Internet video and distribution quality television in any environment.” [Online]. Available: https://www.itu.int/rec/T-REC-P.913-201401-I/en. [Accessed: 11-Dec-2014].

[57] R. Schatz, T. Hoßfeld, L. Janowski, and S. Egger, “From Packets to People: Quality of Experience as a New Measurement Challenge,” in *Data Traffic Monitoring and Analysis*, E. Biersack, C. Callegari, and M. Matijasevic, Eds. Springer Berlin Heidelberg, 2013, pp. 219–263.

[58] P. Brooks and B. Hestnes, “User measures of quality of experience: why being objective and quantitative is important,” *Netw. IEEE*, vol. 24, no. 2, pp. 8 –13, Apr. 2010.

[59] Q. Huynh-Thu, M.-N. Garcia, F. Speranza, P. Corriveau, and A. Raake, “Study of rating scales for subjective quality assessment of high-definition video,” *Broadcast. IEEE Trans. On*, vol. 57, no. 1, pp. 1–14, 2011.

[60] “P.10 : New Appendix I - Definition of Quality of Experience (QoE).” [Online]. Available: http://www.itu.int/rec/T-REC-P.10-200701-S!Amd1/en. [Accessed: 06-Nov-2014].

[61] W. Song and D. W. Tjondronegoro, “Acceptability-Based QoE Models for Mobile Video,” *IEEE Trans. Multimed.*, vol. 16, no. 3, pp. 738–750, Apr. 2014.

[62] K. U. R. Laghari, I. Khan, N. Crespi, and others, “Quantitative and Qualitative Assessment of QoE for Multimedia Services in Wireless Environment,” in *Proceedings of 4th ACM Workshop on Mobile Video in conjunction with ACM Multimedia Systems Conference 2012*, 2012, pp. 7–12.

[63] M. Volk, J. Sterle, U. Sedlar, and A. Kos, “An approach to modeling and control of QoE in next generation networks [Next Generation Telco IT Architectures],” *Commun. Mag. IEEE*, vol. 48, no. 8, pp. 126–135, 2010.

[64] T. Wang, A. Pervez, and H. Zou, “VQM-based QoS/QoE mapping for streaming video,” in *Broadband Network and Multimedia Technology (IC-BNMT), 2010 3rd IEEE International Conference on*, 2010, pp. 807 –812.

[65] Zona Research, “The Economic Impacts of Unacceptable Web Site Download Speeds,” Research Report, http//www. zonaresearch. com, 1999.

[66] A. Bouch, A. Kuchinsky, and N. Bhatti, “Quality is in the Eye of the Beholder: Meeting Users’ Requirements for Internet Quality of Service,” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, New York, NY, USA, 2000, pp. 297–304.

[67] J. Nielsen, *Usability engineering*. Elsevier, 1994.

[68] T. Hosfeld, S. Biedermann, R. Schatz, A. Platzer, S. Egger, and M. Fiedler, “The memory effect and its implications on Web QoE modeling,” in *Teletraffic Congress (ITC), 2011 23rd International*, 2011, pp. 103 –110.

[69] S. C. Seow, *Designing and engineering time: the psychology of time perception in software*. Addison-Wesley Professional, 2008.

[70] N. Bhatti, A. Bouch, and A. Kuchinsky, “Integrating User-Perceived Quality into Web Server Design,” in *IN 9TH INTERNATIONAL WORLD WIDE WEB CONFERENCE*, 2000, pp. 1–16.

[71] van Moorsel, Aad, “Metrics for the Internet Age: Quality of Experience and Quality of Business,” 2001.

[72] E. Ibarrola, F. Liberal, I. Taboada, and R. Ortega, “Web QoE evaluation in multi-agent networks: validation of ITU-T G. 1030,” in *Autonomic and Autonomous Systems, 2009. ICAS’09. Fifth International Conference on*, 2009, pp. 289–294.

[73] S. Egger, P. Reichl, T. Hoßfeld, and R. Schatz, “Time is Bandwidth? Narrowing the Gap between Subjective Time Perception and Quality of Experience,” in *Proc. IEEE International Conference on Communications (ICC 2012)-Communication QoS, Reliability and Modeling Symposium, Ottawa, Canada (June 2012)*, 2012.

[74] S. Niida, S. Uemura, and H. Nakamura, “Mobile Services,” *IEEE Veh. Technol. Mag.*, vol. 5, no. 3, pp. 61 –67, Sep. 2010.

[75] S. Egger, T. Hossfeld, R. Schatz, and M. Fiedler, “Waiting times in quality of experience for web based services,” in *2012 Fourth International Workshop on Quality of Multimedia Experience (QoMEX)*, 2012, pp. 86 –96.

[76] P. Reichl, S. Egger, R. Schatz, and A. D’Alconzo, “The Logarithmic Nature of QoE and the Role of the Weber-Fechner Law in QoE Assessment,” 2010, pp. 1–5.

[77] T. Hossfeld, S. Egger, R. Schatz, M. Fiedler, K. Masuch, and C. Lorentzen, “Initial delay vs. interruptions: Between the devil and the deep blue sea,” in *2012 Fourth International Workshop on Quality of Multimedia Experience (QoMEX)*, 2012, pp. 1 –6.

[78] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. V. Gool, “A Comparison of Affine Region Detectors,” *Int J Comput Vis.*, vol. 65, no. 1–2, pp. 43–72, Nov. 2005.

[79] T. Tuytelaars and K. Mikolajczyk, “Local invariant feature detectors: a survey,” *Found. Trends® Comput. Graph. Vis.*, vol. 3, no. 3, pp. 177–280, 2008.

[80] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[81] J. Matas, O. Chum, M. Urban, and T. Pajdla, “Robust wide-baseline stereo from maximally stable extremal regions,” *Image Vis. Comput.*, vol. 22, no. 10, pp. 761–767, 2004.

[82] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, “ORB: an efficient alternative to SIFT or SURF,” in *Computer Vision (ICCV), 2011 IEEE International Conference on*, 2011, pp. 2564–2571.

[83] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, “Speeded-up robust features (SURF),” *Comput. Vis. Image Underst.*, vol. 110, no. 3, pp. 346–359, 2008.

[84] M. Agrawal, K. Konolige, and M. R. Blas, “Censure: Center surround extremas for realtime feature detection and matching,” in *Computer Vision–ECCV 2008*, Springer, 2008, pp. 102–115.

[85] C. Harris and M. Stephens, “A combined corner and edge detector.,” in *Alvey vision conference*, 1988, vol. 15, p. 50.

[86] J. Shi and C. Tomasi, “Good features to track,” in *, 1994 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1994. Proceedings CVPR ’94*, 1994, pp. 593–600.

[87] T. Lindeberg, “Feature detection with automatic scale selection,” *Int. J. Comput. Vis.*, vol. 30, no. 2, pp. 79–116, 1998.

[88] D. G. Lowe, “Object recognition from local scale-invariant features,” in *The Proceedings of the Seventh IEEE International Conference on Computer Vision, 1999*, 1999, vol. 2, pp. 1150–1157 vol.2.

[89] K. Mikolajczyk and C. Schmid, “Scale and affine invariant interest point detectors,” *Int. J. Comput. Vis.*, vol. 60, no. 1, pp. 63–86, 2004.

[90] T. Tuytelaars and L. Van Gool, “Matching Widely Separated Views Based on Affine Invariant Regions,” *Int J Comput Vis.*, vol. 59, no. 1, pp. 61–85, Aug. 2004.

[91] T. Kadir, A. Zisserman, and M. Brady, “An affine invariant salient region detector,” in *Computer Vision-ECCV 2004*, Springer, 2004, pp. 228–241.

[92] H. Bay, T. Tuytelaars, and L. Van Gool, “Surf: Speeded up robust features,” in *Computer Vision–ECCV 2006*, Springer, 2006, pp. 404–417.

[93] D. Nistér and H. Stewénius, “Linear time maximally stable extremal regions,” in *Computer Vision–ECCV 2008*, Springer, 2008, pp. 183–196.

[94] E. Rosten, R. Porter, and T. Drummond, “Faster and Better: A Machine Learning Approach to Corner Detection,” *IEEE Trans Pattern Anal Mach Intell*, vol. 32, no. 1, pp. 105–119, Jan. 2010.

[95] G. Takacs, V. Chandrasekhar, S. Tsai, D. Chen, R. Grzeszczuk, and B. Girod, “Rotation-invariant Fast Features for Large-scale Recognition and Real-time Tracking,” *Image Commun*, vol. 28, no. 4, pp. 334–344, Apr. 2013.

[96] S. Leutenegger, M. Chli, and R. Y. Siegwart, “BRISK: Binary robust invariant scalable keypoints,” in *Computer Vision (ICCV), 2011 IEEE International Conference on*, 2011, pp. 2548–2555.

[97] E. Rosten and T. Drummond, “Machine learning for high-speed corner detection,” in *Computer Vision–ECCV 2006*, Springer, 2006, pp. 430–443.

[98] K. Mikolajczyk and C. Schmid, “A performance evaluation of local descriptors,” *Pattern Anal. Mach. Intell. IEEE Trans. On*, vol. 27, no. 10, pp. 1615–1630, 2005.

[99] Y. Ke and R. Sukthankar, “PCA-SIFT: a more distinctive representation for local image descriptors,” in *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004*, 2004, vol. 2, pp. II–506–II–513 Vol.2.

[100] V. Chandrasekhar, G. Takacs, D. M. Chen, S. S. Tsai, Y. Reznik, R. Grzeszczuk, and B. Girod, “Compressed Histogram of Gradients: A Low-Bitrate Descriptor,” *Int. J. Comput. Vis.*, pp. 1–16, 2012.

[101] I. W. Selesnick, R. G. Baraniuk, and N. C. Kingsbury, “The dual-tree complex wavelet transform,” *IEEE Signal Process. Mag.*, vol. 22, no. 6, pp. 123–151, Nov. 2005.

[102] T. Song and H. Li, “Local Polar DCT Features for Image Description,” *IEEE Signal Process. Lett.*, vol. 20, no. 1, pp. 59–62, Jan. 2013.

[103] M. Calonder, V. Lepetit, C. Strecha, and P. Fua, “BRIEF: Binary Robust Independent Elementary Features,” in *Computer Vision – ECCV 2010*, K.

Daniilidis, P. Maragos, and N. Paragios, Eds. Springer Berlin Heidelberg, 2010, pp. 778–792.
[104] A. Alahi, R. Ortiz, and P. Vandergheynst, “FREAK: Fast Retina Keypoint,” in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, 2012, pp. 510–517.
[105] J. Heinly, E. Dunn, and J.-M. Frahm, “Comparative evaluation of binary features,” in *Computer Vision–ECCV 2012*, Springer, 2012, pp. 759–773.
[106] D. Zhang and G. Lu, “Evaluation of similarity measurement for image retrieval,” in *Proceedings of the 2003 International Conference on Neural Networks and Signal Processing, 2003*, 2003, vol. 2, pp. 928–931 Vol.2.
[107] V. Di Gesù and V. Starovoitov, “Distance-based functions for image comparison,” *Pattern Recognit. Lett.*, vol. 20, no. 2, pp. 207–214, Feb. 1999.
[108] L. Wang, Y. Zhang, and J. Feng, “On the Euclidean distance of images,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1334–1339, Aug. 2005.
[109] R. Larkins and M. Mayo, “Adaptive Feature Thresholding for off-line signature verification,” in *Image and Vision Computing New Zealand, 2008. IVCNZ 2008. 23rd International Conference*, 2008, pp. 1–6.
[110] H. Yazid and H. Arof, “Gradient based adaptive thresholding,” *J. Vis. Commun. Image Represent.*, vol. 24, no. 7, pp. 926–936, 2013.
[111] S. J. H. Pirzada, M. W. Baig, E. U. Haq, and H. Shin, “A new adaptive threshold technique for improved matching in SIFT,” in *2011 IEEE 54th International Midwest Symposium on Circuits and Systems (MWSCAS)*, 2011, pp. 1–4.
[112] L. Cayton, “Fast nearest neighbor retrieval for bregman divergences,” in *Proceedings of the 25th international conference on Machine learning*, 2008, pp. 112–119.
[113] K. L. Clarkson, “Fast algorithms for the all nearest neighbors problem,” in *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, 1983, pp. 226–232.
[114] S. Ramaswamy and K. Rose, “Adaptive cluster-distance bounding for nearest neighbor search in image databases,” in *Image Processing, 2007. ICIP 2007. IEEE International Conference on*, 2007, vol. 6, pp. VI–381.
[115] S. Ramaswamy and K. Rose, “Adaptive cluster distance bounding for high-dimensional indexing,” *Knowl. Data Eng. IEEE Trans. On*, vol. 23, no. 6, pp. 815–830, 2011.
[116] R. Weber and S. Blott, “An approximation based data structure for similarity search,” Citeseer, 1997.
[117] R. Weber, H.-J. Schek, and S. Blott, “A quantitative analysis and performance study for similarity-search methods in high-dimensional spaces,” in *VLDB*, 1998, vol. 98, pp. 194–205.
[118] M. Brown and D. G. Lowe, “Invariant Features from Interest Point Groups.,” in *BMVC*, 2002.

[119] “Common Interfaces of Descriptor Matchers.” [Online]. Available: http://docs.opencv.org/modules/features2d/doc/common_interfaces_of_descriptor_matchers.html.
[120] “OpenCV Wiki [Online],.” .
[121] S. Arya, D. M. Mount, N. S. Netanyahu, R. Silverman, and A. Y. Wu, “An optimal algorithm for approximate nearest neighbor searching fixed dimensions,” *J. ACM JACM*, vol. 45, no. 6, pp. 891–923, 1998.
[122] S. Arya and D. M. Mount, “Algorithms for fast vector quantization,” in *Data Compression Conference, 1993. DCC’93.*, 1993, pp. 381–390.
[123] M. Muja and D. G. Lowe, “Fast approximate nearest neighbors with automatic algorithm configuration,” in *In VISAPP International Conference on Computer Vision Theory and Applications*, 2009, pp. 331–340.
[124] M. Muja and D. G. Lowe, “Fast matching of binary features,” in *Computer and Robot Vision (CRV), 2012 Ninth Conference on*, 2012, pp. 404–410.
[125] M. Muja and D. G. Lowe, “Scalable Nearest Neighbor Algorithms for High Dimensional Data,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 11, pp. 2227–2240, Nov. 2014.
[126] M. A. Fischler and R. C. Bolles, “Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography,” *Commun. ACM*, vol. 24, no. 6, pp. 381–395, 1981.
[127] P. J. Rousseeuw, “Least median of squares regression,” *J. Am. Stat. Assoc.*, vol. 79, no. 388, pp. 871–880, 1984.
[128] D. L. Massart, L. Kaufman, P. J. Rousseeuw, and A. Leroy, “Least median of squares: a robust method for outlier and model error detection in regression and calibration,” *Anal. Chim. Acta*, vol. 187, pp. 171–179, 1986.
[129] A. J. Stromberg, “Computing the exact least median of squares estimate and stability diagnostics in multiple linear regression,” *SIAM J. Sci. Comput.*, vol. 14, no. 6, pp. 1289–1299, 1993.
[130] C. J. ter Braak and S. Juggins, “Weighted averaging partial least squares regression (WA-PLS): an improved method for reconstructing environmental variables from species assemblages,” *Hydrobiologia*, vol. 269, no. 1, pp. 485–502, 1993.
[131] O. Chum and J. Matas, “Matching with PROSAC-progressive sample consensus,” in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, 2005, vol. 1, pp. 220–226.
[132] R. Raguram, J.-M. Frahm, and M. Pollefeys, “A comparative analysis of RANSAC techniques leading to adaptive real-time random sample consensus,” in *Computer Vision–ECCV 2008*, Springer, 2008, pp. 500–513.
[133] P. H. S. Torr and C. Davidson, “IMPSAC: Synthesis of importance sampling and random sample consensus,” *Pattern Anal. Mach. Intell. IEEE Trans. On*, vol. 25, no. 3, pp. 354–364, 2003.

[134] P. Meer, D. Mintz, A. Rosenfeld, and D. Y. Kim, "Robust regression methods for computer vision: A review," *Int. J. Comput. Vis.*, vol. 6, no. 1, pp. 59–70, 1991.

[135] Y. Sun, S. Todorovic, and S. Goodison, "Local-Learning-Based Feature Selection for High-Dimensional Data Analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1610–1626, Sep. 2010.

[136] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1226–1238, Aug. 2005.

[137] W. Jiang, G. Er, Q. Dai, and J. Gu, "Similarity-based online feature selection in content-based image retrieval," *IEEE Trans. Image Process.*, vol. 15, no. 3, pp. 702–712, Mar. 2006.

[138] X. Liu, L. Wang, J. Zhang, J. Yin, and H. Liu, "Global and Local Structure Preservation for Feature Selection," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 6, pp. 1083–1095, Jun. 2014.

[139] J. Tang, S. Alelyani, and H. Liu, "Feature selection for classification: A review," *Data Classif. Algorithms Appl. Ed. Charu Aggarwal CRC Press Chapman HallCRC Data Min. Knowl. Discov. Ser.*, 2014.

[140]S. Alelyani, J. Tang, and H. Liu, "Feature Selection for Clustering: A Review.," *Data Clust. Algorithms Appl.*, vol. 29, 2013.

[141] H. Liu and H. Motoda, *Feature selection for knowledge discovery and data mining*. Springer Science & Business Media, 1998.

[142] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, 2003.

[143] C.-L. Huang and C.-J. Wang, "A GA-based feature selection and parameters optimizationfor support vector machines," *Expert Syst. Appl.*, vol. 31, no. 2, pp. 231–240, 2006.

[144] R. H. Pinheiro, G. D. Cavalcanti, R. F. Correa, and T. I. Ren, "A global-ranking local feature selection method for text categorization," *Expert Syst. Appl.*, vol. 39, no. 17, pp. 12851–12857, 2012.

[145] A. Jain and D. Zongker, "Feature selection: Evaluation, application, and small sample performance," *Pattern Anal. Mach. Intell. IEEE Trans. On*, vol. 19, no. 2, pp. 153–158, 1997.

[146] P. Pudil, J. Novovičová, and J. Kittler, "Floating search methods in feature selection," *Pattern Recognit. Lett.*, vol. 15, no. 11, pp. 1119–1125, 1994.

[147] L. I. Kuncheva and L. C. Jain, "Nearest neighbor classifier: simultaneous editing and feature selection," *Pattern Recognit. Lett.*, vol. 20, no. 11, pp. 1149–1156, 1999.

[148] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artif. Intell.*, vol. 97, no. 1, pp. 273–324, 1997.

[149] ISO/IEC JTC1/SC29/WG11/N12551, "CDVS, Description of Core Experiments on Compact descriptors for Visual Search." Feb-2012.

[150] ISO/IEC JTC1/SC29/WG11/M23929, “Reference results of key point reduction.” 99th MPEG Meeting, Sanjose, USA, 2012.
[151] ISO/IEC JTC1/SC29/WG11/N12550, “Test Model 1: Compact Descriptors for Visual Search.” Feb-2012.
[152] ISO/IEC/JTC1/SC29/WG11/W12929, “Test Model 3: Compact Descriptor for Visual Search.” Jul-2012.
[153]“Study Text of ISO/IEC CD 15938-13 Compact Descriptors for Visual Search.” [Online]. Available: http://mpeg.chiariglione.org/standards/mpeg-7/compact-descriptors-visual-search/study-text-isoiec-cd-15938-13-compact-descriptors. [Accessed: 16-May-2014].
[154] ISO/IEC JTC1/SC29/WG11/N14393, “Test Model 10: Compact Descriptors for Visual Search.” Apr-2014.
[155] H. Cornelius, M. Perďoch, J. Matas, and G. Loy, “Efficient Symmetry Detection Using Local Affine Frames,” in *Image Analysis*, B. K. Ersbøll and K. S. Pedersen, Eds. Springer Berlin Heidelberg, 2007, pp. 152–161.
[156] Y. Keller and Y. Shkolnisky, “An algebraic approach to symmetry detection,” in *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004*, 2004, vol. 3, pp. 186–189 Vol.3.
[157] C. Sun and D. Si, “Fast Reflectional Symmetry Detection Using Orientation Histograms,” *Real-Time Imaging*, vol. 5, no. 1, pp. 63–74, Feb. 1999.
[158] F. J. Estrada, P. Fua, V. Lepetit, and S. Susstrunk, “Appearance-based keypoint clustering,” in *IEEE Conference on Computer Vision and Pattern Recognition, 2009. CVPR 2009*, 2009, pp. 1279–1286.
[159] P. Turcot and D. G. Lowe, “Better matching with fewer features: The selection of useful features in large database recognition problems,” in *2009 IEEE 12th International Conference on Computer Vision Workshops (ICCV Workshops)*, 2009, pp. 2109–2116.
[160] S. S. Tsai, D. Chen, G. Takacs, V. Chandrasekhar, J. P. Singh, and B. Girod, “Location coding for mobile image retrieval,” in *Proceedings of the 5th International ICST Mobile Multimedia Communications Conference*, 2009, p. 8.
[161] G. Tolias, Y. Kalantidis, and Y. Avrithis, “SymCity: feature selection by symmetry for large scale image retrieval,” in *Proceedings of the 20th ACM international conference on Multimedia*, 2012, pp. 189–198.
[162] X. Xin, Z. Li, Z. Ma, and A. K. Katsaggelos, “Robust feature selection with self-matching score,” in *2013 20th IEEE International Conference on Image Processing (ICIP)*, 2013, pp. 4363–4366.
[163] Y. Guan, M. I. Jordan, and J. G. Dy, “A unified probabilistic model for global and local unsupervised feature selection,” in *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, 2011, pp. 1073–1080.
[164] R. Kohavi and G. H. John, “Wrappers for feature subset selection,” *Artif. Intell.*, vol. 97, no. 1–2, pp. 273–324, Dec. 1997.

[165] L. Talavera, “An Evaluation of Filter and Wrapper Methods for Feature Selection in Categorical Clustering,” in *Advances in Intelligent Data Analysis VI*, A. F. Famili, J. N. Kok, J. M. Peña, A. Siebes, and A. Feelders, Eds. Springer Berlin Heidelberg, 2005, pp. 440–451.

[166]S. Das, “Filters, Wrappers and a Boosting-Based Hybrid for Feature Selection,” in *Proceedings of the Eighteenth International Conference on Machine Learning*, San Francisco, CA, USA, 2001, pp. 74–81.

[167] G. Francini, S. Lepsøy, and M. Balestri, “Selection of local features for visual search,” *Signal Process. Image Commun.*, vol. 28, no. 4, pp. 311–322, Apr. 2013.

[168] K. Lee, S. Lee, S. Na, S. Je, and W.-G. Oh, “Extensive analysis of feature selection for compact descriptor,” in *2013 19th Korea-Japan Joint Workshop on Frontiers of Computer Vision, (FCV)*, 2013, pp. 53–57.

[169] S. Davis, E. Cheng, C. Ritz, and I. Burnett, “Ensuring Quality of Experience for markerless image recognition applied to print media content,” in *2012 Fourth International Workshop on Quality of Multimedia Experience (QoMEX)*, 2012, pp. 158 –163.

[170] S. Ganapathy, G. J. Anderson, and I. V. Kozintsev, “MAR shopping assistant usage: Delay, error, and utility,” in *Virtual Reality Conference (VR), 2011 IEEE*, 2011, pp. 207 –208.

[171] “24 Augmented Reality Retail Experiences - From Virtual Window Shopping to Strippable Catalogs (TOPLIST),” 31-Oct-2013. [Online]. Available: http://www.trendhunter.com/slideshow/augmented-reality-retail. [Accessed: 31-Oct-2013].

[172] B. Girod, V. Chandrasekhar, D. M. Chen, N.-M. Cheung, R. Grzeszczuk, Y. Reznik, G. Takacs, S. S. Tsai, and R. Vedantham, “Mobile Visual Search,” *IEEE Signal Process. Mag.*, vol. 28, no. 4, pp. 61 –76, Jul. 2011.

[173] Y. Cao, C. Ritz, and R. Raad, “How much longer to go? The influence of waiting time and progress indicators on quality of experience for mobile visual search applied to print media,” in *2013 Fifth International Workshop on Quality of Multimedia Experience (QoMEX)*, 2013, pp. 112–117.

[174] K.-O. Cheng, N.-F. Law, and W.-C. Siu, “Fast extraction of wavelet-based features from JPEG images for joint retrieval with JPEG2000 images,” *Pattern Recognit.*, vol. 43, no. 10, pp. 3314–3323, 2010.

[175] C.-W. Ngo, T.-C. Pong, and R. T. Chin, “Exploiting image indexing techniques in DCT domain,” *Pattern Recognit.*, vol. 34, no. 9, pp. 1841–1851, Sep. 2001.

[176] J. A. Lay and L. Guan, “Image retrieval based on energy histograms of the low frequency DCT coefficients,” in *, 1999 IEEE International Conference on Acoustics, Speech, and Signal Processing, 1999. Proceedings*, 1999, vol. 6, pp. 3009–3012 vol.6.

[177] F. Arnia, I. Iizuka, M. Fujiyoshi, and H. Kiya, “Fast Method for Joint Retrieval and Identification of JPEG Coded Images Based on DCT Sign,” in

IEEE International Conference on Image Processing, 2007. ICIP 2007, 2007, vol. 2, pp. II – 229–II – 232.

[178] M. N. Do and M. Vetterli, "Wavelet-based texture retrieval using generalized Gaussian density and Kullback-Leibler distance," *Image Process. IEEE Trans. On*, vol. 11, no. 2, pp. 146–158, 2002.

[179] V. Chandrasekhar, G. Takacs, D. Chen, S. S. Tsai, J. Singh, and B. Girod, "Transform coding of image feature descriptors," 2009, vol. 7257, pp. 725710–725710–9.

[180] V. Chandrasekhar, M. Makar, G. Takacs, D. Chen, S. S. Tsai, N. M. Cheung, R. Grzeszczuk, Y. Reznik, and B. Girod, "Survey of SIFT compression schemes," in *Proceedings of International Mobile Multimedia Workshop (IMMW), IEEE International Conference on Pattern Recognition (ICPR), Istanbul, Turkey*, 2010.

[181] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, 2006.

[182] T. C. Landgrebe, P. Paclik, R. P. Duin, and A. P. Bradley, "Precision-recall operating characteristic (P-ROC) curves in imprecise environments," in *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, 2006, vol. 4, pp. 123–127.

[183] K. Zuva and T. Zuva, "Evaluation of Information Retrieval Systems," *Int. J. Comput. Sci. Inf. Technol. IJCSIT*, vol. 4, pp. 35–43, 2012.

[184] B. Girod, V. Chandrasekhar, R. Grzeszczuk, and Y. A. Reznik, "Mobile visual search: Architectures, technologies, and the emerging MPEG standard," *Multimed. IEEE*, vol. 18, no. 3, pp. 86–94, 2011.

[185] H. Li, Y. Wang, T. Mei, J. Wang, and S. Li, "Interactive Multimodal Visual Search on Mobile Device," *IEEE Trans. Multimed.*, vol. 15, no. 3, pp. 594–607, 2013.

[186] A. H. Behzadan, H. M. Khoury, and V. R. Kamat, "Structure of an Extensible Augmented Reality Framework for Visualization of Simulated Construction Processes," in *Simulation Conference, 2006. WSC 06. Proceedings of the Winter*, 2006, pp. 2055–2062.

[187] L.-Y. Duan, F. Gao, J. Chen, J. Lin, and T. Huang, "Compact descriptors for mobile visual search and MPEG CDVS standardization," in *2013 IEEE International Symposium on Circuits and Systems (ISCAS)*, 2013, pp. 885–888.

[188] ISO/IEC JTC1/SC29/WG11/N12202, "Evaluation Framework for Compact Descriptors for Visual Search," Jul. 2011.

[189] M. W. Marcellin, *JPEG2000 Image Compression Fundamentals, Standards and Practice: Image Compression Fundamentals, Standards, and Practice*, vol. 1. Springer Science & Business Media, 2002.

[190] "OpenCV documentation." [Online]. Available: http://docs.opencv.org/.

[191] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[192] E. Mair, G. D. Hager, D. Burschka, M. Suppa, and G. Hirzinger, “Adaptive and Generic Corner Detection Based on the Accelerated Segment Test,” in *Computer Vision – ECCV 2010*, K. Daniilidis, P. Maragos, and N. Paragios, Eds. Springer Berlin Heidelberg, 2010, pp. 183–196.

[193] “Camera Calibration and 3D Reconstruction — OpenCV 2.4.11.0 documentation.” [Online]. Available: http://docs.opencv.org/modules/calib3d/doc/camera_calibration_and_3d_reconstruction.html?highlight=homo#int%20cvFindHomography(const%20CvMat*%20src_points,%20const%20CvMat*%20dst_points,%20CvMat*%20homography,%20int%20method,%20double%20ransacReprojThreshold,%20CvMat*%20mask). [Accessed: 17-Jul-2015].

[194] V. R. Chandrasekhar, D. M. Chen, S. S. Tsai, N.-M. Cheung, H. Chen, G. Takacs, Y. Reznik, R. Vedantham, R. Grzeszczuk, J. Bach, and B. Girod, “The Stanford Mobile Visual Search Data Set,” in *Proceedings of the Second Annual ACM Conference on Multimedia Systems*, New York, NY, USA, 2011, pp. 117–122.

[195] “Kakadu Software.” .

[196] “A Day in the Life of 3G,” *PCWorld*, 28-Jun-2009. [Online]. Available: http://www.pcworld.com/article/167391/3GTests.html. [Accessed: 21-Apr-2015].

[197] E. Eibenberger and E. Angelopoulou, “Beyond the neutral interface reflection assumption in illuminant color estimation,” in *2010 17th IEEE International Conference on Image Processing (ICIP)*, 2010, pp. 4689 –4692.

[198] X. Liu and A. El Gamal, “Simultaneous image formation and motion blur restoration via multiple capture,” in *Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP’01). 2001 IEEE International Conference on*, 2001, vol. 3, pp. 1841–1844.

[199] R. Johansson, A. Storm, C. Stephansen, S. Eikedal, T. Willassen, S. Skaug, T. Martinussen, D. Whittlesea, G. Ali, J. Ladd, X. Li, S. Johnson, V. Rajasekaran, Y. Lee, J. Bai, M. Flores, G. Davies, H. Samiy, A. Hanvey, and D. Perks, “A 1/13-inch 30fps VGA SoC CMOS image sensor with shared reset and transfer-gate pixel control,” presented at the Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2011 IEEE International, 2011, pp. 414 –415.

[200] K.-B. Cho, C. Lee, S. Eikedal, A. Baum, J. Jiang, C. Xu, X. Fan, and R. Kauffman, “A 1/2.5 inch 8.1Mpixel CMOS Image Sensor for Digital Cameras,” presented at the Solid-State Circuits Conference, 2007. ISSCC 2007. Digest of Technical Papers. IEEE International, 2007, pp. 508 –618.

[201] J. Rhee and Y. Joo, “A new wide dynamic range fixed point ADC for FPAs,” in *The 2002 45th Midwest Symposium on Circuits and Systems, 2002. MWSCAS-2002*, 2002, vol. 2, pp. II–243 – II–245 vol.2.

[202] H. Tong, M. Li, H. Zhang, and C. Zhang, “Blur detection for digital images using wavelet transform,” in *2004 IEEE International Conference on Multimedia and Expo, 2004. ICME ’04*, 2004, vol. 1, pp. 17 – 20 Vol.1.

[203] R. Liu, Z. Li, and J. Jia, “Image partial blur detection and classification,” in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, 2008, pp. 1–8.

[204] R. L. Lagendijk and J. Biemond, *Basic Methods for Image Restoration and Identification*. 1999.

[205] Lagendijk, *Handbook of Image and Video Processing 2nd edition*. Burlington, MA: Elsevier Academic Press, 2005.

[206] J. Yuan, G. Zhao, Y. Fu, Z. Li, A. K. Katsaggelos, and Y. Wu, “Discovering Thematic Objects in Image Collections and Videos,” *IEEE Trans. Image Process.*, vol. 21, no. 4, pp. 2207 –2219, Apr. 2012.

[207] S. Vaddadi, O. Hamsici, Y. Reznik, J. Hong, and C. Lee, “Keypoint clustering for robust image matching,” *Proc SPIE 7798 Appl. Digit. Image Process. XXXIII 77980K*, Sep. 2010.

[208] D. Comaniciu and P. Meer, “Mean shift: A robust approach toward feature space analysis,” *Pattern Anal. Mach. Intell. IEEE Trans. On*, vol. 24, no. 5, pp. 603–619, 2002.

[209] V. R. Chandrasekhar, D. M. Chen, S. S. Tsai, N.-M. Cheung, H. Chen, G. Takacs, Y. Reznik, R. Vedantham, R. Grzeszczuk, J. Bach, and B. Girod, “The stanford mobile visual search data set,” in *Proceedings of the second annual ACM conference on Multimedia systems*, New York, NY, USA, 2011, pp. 117–122.

[210] X. Xin, Z. Li, and A. K. Katsaggelos, “LAPLACIAN SIFT IN VISUAL SEARCH,” in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012.

[211] Z. Xiong and T. S. Huang, “Wavelet-based texture features can be extracted efficiently from compressed-domain for JPEG2000 coded images,” in *Image Processing. 2002. Proceedings. 2002 International Conference on*, 2002, vol. 1, pp. I–481.

[212] B. Baharudin, “Effective content-based image retrieval: Combination of quantized histogram texture features in the DCT domain,” in *Computer & Information Science (ICCIS), 2012 International Conference on*, 2012, vol. 1, pp. 425–430.

[213] Y. Cao, C. Ritz, and R. Raad, “Image compression and retrieval for Mobile Visual Search,” in *Communications and Information Technologies (ISCIT), 2012 International Symposium on*, 2012, pp. 1027–1032.

[214] S. A. Khayam, “The discrete cosine transform (dct): theory and application,” *Mich. State Univ.*, 2003.

[215] N. Ponomarenko, V. Lukin, A. Zelensky, K. Egiazarian, M. Carli, and F. Battisti, “Tid2008-a database for evaluation of full-reference visual quality assessment metrics,” *Adv. Mod. Radioelectron.*, vol. 10, no. 4, pp. 30–45, 2009.

[216] E. C. Larson and D. M. Chandler, “Most apparent distortion: full-reference image quality assessment and the role of strategy,” *J. Electron. Imaging*, vol. 19, no. 1, pp. 011006–011006, Jan. 2010.

[217] G. Schaefer and M. Stich, “UCID: an uncompressed color image database,” 2003, vol. 5307, pp. 472–480.

[218] Y. Jia, J. Wang, G. Zeng, H. Zha, and X.-S. Hua, “Optimizing kd-trees for scalable visual descriptor indexing,” in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, 2010, pp. 3392–3399.

[219] J. S. Beis and D. G. Lowe, “Shape indexing using approximate nearest-neighbour search in high-dimensional spaces,” in *Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on*, 1997, pp. 1000–1006.

[220] M. Aly, M. Munich, and P. Perona, “Distributed kd-trees for retrieval from very large image collections,” in *British Machine Vision Conference, Dundee, Scotland*, 2011.

[221] D. Nister and H. Stewenius, “Scalable recognition with a vocabulary tree,” in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, 2006, vol. 2, pp. 2161–2168.

[222] Y. Cao, C. Ritz, and R. Raad, “The Joint Effect of Image Blur and Illumination Distortions for Mobile Visual Search of Print Media,” in *2013 International Symposium on Communications and Information Technologies (ISCIT)*, 2013.

[223] T. Kadir and M. Brady, “Saliency, scale and image description,” *Int. J. Comput. Vis.*, vol. 45, no. 2, pp. 83–105, 2001.

[224] D. Chen and B. Girod, “Memory-Efficient Image Databases for Mobile Visual Search,” 2013.

[225] V. Ramasubramanian and K. K. Paliwal, “Fast K-dimensional tree algorithms for nearest neighbor search with application to vector quantization encoding,” *IEEE Trans. Signal Process.*, vol. 40, no. 3, pp. 518–531, Mar. 1992.

[226] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to information retrieval*, vol. 1. Cambridge University Press Cambridge, 2008.

[227] O. Miksik and K. Mikolajczyk, “Evaluation of local detectors and descriptors for fast feature matching,” in *2012 21st International Conference on Pattern Recognition (ICPR)*, 2012, pp. 2681–2684.

[228] D. Geerts, K. De Moor, I. Ketyko, A. Jacobs, J. Van den Bergh, W. Joseph, L. Martens, and L. De Marez, “Linking an integrated framework with appropriate methods for measuring QoE,” in *Quality of Multimedia Experience (QoMEX), 2010 Second International Workshop on*, 2010, pp. 158 –163.

[229] A. Sackl, K. Masuch, S. Egger, and R. Schatz, “Wireless vs. wireline shootout: Howuser expectations influence quality of experience,” in *2012 Fourth International Workshop on Quality of Multimedia Experience (QoMEX)*, 2012, pp. 148 –149.

[230] Y. Cao, C. Ritz, and R. Raad, “Image compression and retrieval for Mobile Visual Search,” in *2012 International Symposium on Communications and Information Technologies (ISCIT)*, 2012, pp. 1027–1032.

[231] ITU-T:, “Definition of Quality of Experience (QoE).” International Telecommunication Union, Liaison Statement, Ref.: TD 109rev2 (PLEN/12), Jan-2007.

[232]QUALINET:, “Qualinet White Paper on Definitions of Quality of Experience.” European Network on Quality of Experience in Multimedia Systems and Services, Jun-2012.

[233] ITU-T, “Methods for subjective determination of transmission quality.” ITU-T Recommendation P.800, Aug-1996.

[234] S. Jumisko-Pyykkö, V. K. Malamal Vadakital, and M. M. Hannuksela, “Acceptance Threshold: A Bidimensional Research Method for User-Oriented Quality Evaluation Studies,” *Int. J. Digit. Multimed. Broadcast.*, vol. 2008, pp. 1–20, 2008.

[235] R. Schatz, S. Egger, and A. Platzer, “Poor, Good Enough or Even Better? Bridging the Gap between Acceptability and QoE of Mobile Broadband Data Services,” in *Communications (ICC), 2011 IEEE International Conference on*, 2011, pp. 1–6.

[236] ITU-T, “Subjective video quality assessment methods for multimedia applications.” ITU-T Recommendation P.910, Apr-2008.

[237] S. Grondin, “Timing and time perception: A review of recent behavioral and neuroscience findings and theoretical directions,” *Atten. Percept. Psychophys.*, vol. 72, no. 3, pp. 561–582, Apr. 2010.

[238] ITU-T, “Estimating end-to-end performance in IP networks for data applications.” ITU-T Recommendation G.1030, Nov-2005.

[239] A. Papoulis and S. U. Pillai, *Probability, random variables, and stochastic processes*. Tata McGraw-Hill Education, 2002.

[240] G. Condello, P. Pasteris, D. Pau, and M. Sami, “An OpenCL-based feature matcher,” *Signal Process. Image Commun.*, vol. 28, no. 4, pp. 345–350, Apr. 2013.

[241] “FLVPlayback bufferTime.” [Online]. Available: http://help.adobe.com/en_US/AS2LCR/Flash_10.0/help.html?content=00002386.html.

[242] Mark Sullivan, “Infographic: How fast are America’s wireless networks?,” *TechHive*, 23-May-2013. [Online]. Available: http://www.techhive.com/article/2039568/infographic-how-fast-are-americas-wireless-networks-.html. [Accessed: 06-May-2014].

[243] “Samsung I9500 Galaxy S4.” [Online]. Available: http://www.gsmarena.com/samsung_i9500_galaxy_s4-5125.php.

[244] M. Lalonde, D. Byrns, L. Gagnon, N. Teasdale, and D. Laurendeau, “Real-time eye blink detection with GPU-based SIFT tracking,” in *Fourth Canadian*

Conference on Computer and Robot Vision, 2007. CRV '07, 2007, pp. 481–487.

[245]C. Jiang, Z. Geng, X. Wei, and C. Shen, "SIFT implementation based on GPU," 2013, vol. 8913, pp. 891304–891304–7.

[246] S. N. Sinha, J.-M. Frahm, M. Pollefeys, and Y. Genc, "GPU-based video feature tracking and matching," in *EDGE, Workshop on Edge Computing Using New Commodity Architectures*, 2006, vol. 278, p. 4321.

[247] T. D. Tran, J. Liang, and C. Tu, "Lapped transform via time-domain pre- and post-filtering," *IEEE Trans. Signal Process.*, vol. 51, no. 6, pp. 1557–1571, Jun. 2003.