ABSTRACT

Title of dissertation: PRESERVING TRUSTWORTHINESS AND CONFIDENTIALITY FOR ONLINE MULTIMEDIA

Wenjun Lu, Doctor of Philosophy, 2011

Dissertation directed by: Professor Min Wu
Department of Electrical and Computer Engineering

Technology advancements in areas of mobile computing, social networks, and cloud computing have rapidly changed the way we communicate and interact. The wide adoption of media-oriented mobile devices such as smartphones and tablets enables people to capture information in various media formats, and offers them a rich platform for media consumption. The proliferation of online services and social networks makes it possible to store personal multimedia collection online and share them with family and friends anytime anywhere. Considering the increasing impact of digital multimedia and the trend of cloud computing, this dissertation explores the problem of how to evaluate trustworthiness and preserve confidentiality of online multimedia data.

The dissertation consists of two parts. The first part examines the problem of evaluating trustworthiness of multimedia data distributed online. Given the digital nature of multimedia data, editing and tampering of the multimedia content becomes very easy. Therefore, it is important to analyze and reveal the processing

history of a multimedia document in order to evaluate its trustworthiness. We propose a new forensic technique called "Forensic Hash", which draws synergy between two related research areas of image hashing and non-reference multimedia forensics. A forensic hash is a compact signature capturing important information from the original multimedia document to assist forensic analysis and reveal processing history of a multimedia document under question. Our proposed technique is shown to have the advantage of being compact and offering efficient and accurate analysis to forensic questions that cannot be easily answered by convention forensic techniques. The answers that we obtain from the forensic hash provide valuable information on the trustworthiness of online multimedia data.

The second part of this dissertation addresses the confidentiality issue of multimedia data stored with online services. The emerging cloud computing paradigm makes it attractive to store private multimedia data online for easy access and sharing. However, the potential of cloud services cannot be fully reached unless the issue of how to preserve confidentiality of sensitive data stored in the cloud is addressed. In this dissertation, we explore techniques that enable confidentiality-preserving search of encrypted multimedia, which can play a critical role in secure online multimedia services. Techniques from image processing, information retrieval, and cryptography are jointly and strategically applied to allow efficient rank-ordered search over encrypted multimedia database and at the same time preserve data confidentiality against malicious intruders and service providers. We demonstrate high efficiency and accuracy of the proposed techniques and provide a quantitative comparative study with conventional techniques based on heavy-weight cryptography primitives.

PRESERVING TRUSTWORTHINESS AND CONFIDENTIALITY FOR ONLINE MULTIMEDIA

by

Wenjun Lu

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2011

Advisory Committee:
Professor Min Wu, Chair/Advisor
Professor K. J. Ray Liu
Professor Rama Chellappa
Professor Carol Espy-Wilson
Professor David Jacobs

To my parents and Shanshan.

ACKNOWLEDGEMENTS

I would like to express my greatest gratitude to my advisor, Prof. Min Wu, for her unwavering support and constant encouragement during my graduate study. Her vision and critical insights have led and guided me to explore this interesting area of multimedia forensics and security. Her patience and constructive advice have escorted me through many difficult times, from which, I learned to always look at problems at a higher level and from different angles. Her enthusiasm and endless pursuit of excellence inspired me to work hard, strive for the best, and maintain the highest standard in research and professional careers. I would never achieve this milestone without her guidance and support, and all the lessons and traits that I learned from her will be invaluable for my future endeavor.

I thank Prof. Ray Liu and Prof. Rama Chellappa, from whom I took several excellent courses on signal processing and pattern recognition, which have brought great benefits to my research. Their knowledge and comments on my research have offered me great help. I am also very grateful to Prof. Espy-Wilson and Prof. Jacobs for serving on my dissertation committee and providing valuable comments and suggestions on my thesis.

I would also like to thank my colleagues and office-mates at University of Maryland. I thank Dr. Ashwin Swaminathan and Dr. Avinash Varna for providing guidance and comments on my research. I thank my office-mates Wei-Hong Chuang, Ravi Garg, and Hui Su, with whom I have had and always enjoyed many discussions on various aspects of research and life. I thank all the friends that I have come to know in Maryland. Life will never be the same without them.

Finally, I give my heartfelt gratitude to my parents, who gave me unconditional love and support, and to Shanshan, whose love and companionship have blessed me ever since we met. Words cannot express my gratitude here and this dissertation is dedicated to them.

Table of Contents

List of Tables

List of Figures

CHAPTER 1

Introduction

1.1 Trustworthiness and Confidentiality of Online Multimedia

Recent years have witnessed important technology advancements and trends that have brought stronger momentum and ubiquitousness to online multimedia data such as digital images and videos. First of all, the flourish of mobile devices, especially smartphones and portable cameras, brings the media generation capability to the general public. With the greatly improved camera quality and increasingly larger device storage, people can take images/videos anytime anywhere to save important moments of their lives and record interesting events around them. This leads to the explosion on the number of online multimedia data, as evidenced by the online media sharing sites such as YouTube and Flickr.

Fuelled by the explosion of user-generated multimedia content, media sharing and consumption becomes an indispensable component of today's online experience. Various media sharing sites and social networks, such as YouTube, Flickr, Facebook, Twitter, offer a convenient platform for multimedia data to reach the widest range of audience ever possible. In addition to user-generated content, professional news agencies and media companies are also putting great effort in utilizing the online media sharing services to help them reach a greater amount of audience in a fast and effective way. Multimedia is having a far wider social impact than ever before.

The wide availability of mobile devices and media sharing services have helped diversify the way how multimedia data are consumed. In addition to direct point-to-point transmission, such as web-based browsing, multimedia can also be transmitted through peer-to-peer networks and wireless networks such as 3G and 4G networks. This diversified communication channel along with the digital nature of multimedia data allows an adversary to easily modify the multimedia content and convey completely different information to the end user. This brings serious questions on how to evaluate the trustworthiness of online multimedia data that we look at everyday.

The enabling technology behind various online services is cloud computing, which aims at providing online services for many kinds of digital data. In such a setting, the data will be stored online, and all the management and computation tasks will be performed by the server. Cloud computing has the advantage of reliable storage and providing easy access anytime anywhere. Given the rapid growth of personal multimedia collections, storing and managing multimedia data online is an attractive option. However, online systems are vulnerable to attacks and intrusions.

Therefore, protecting the confidentiality of sensitive data stored online is a critical research issue that has to be addressed in order for the cloud computing to reach its full potential.

1.2 Main Contributions and Dissertation Organization

Motivated by the above mentioned technology trends and the challenges involved therein, we explore two research problems in this dissertation. In the first problem, we examine the problem of how to evaluate trustworthiness of online multimedia data. We propose and develop new forensic tools to detect and estimate several important operations that a multimedia document may have undergone. In contrast to conventional authentication that provides primarily a binary answer of being trustworthy or not, our proposed techniques can reveal more information on the processing history. This capability can help people make better utilization of online multimedia data. The second problem that we explored is on confidentiality-preserving content-based search of online multimedia. We propose efficient techniques by combining areas of image processing, information retrieval, and cryptography. Comparison with conventional cryptography-based approaches is also provided to justify the good trade-off between efficiency and security offered by our proposed techniques. The outline of this dissertation is illustrated in Fig. 1.1. Below, we highlight the key contributions of this dissertation research.

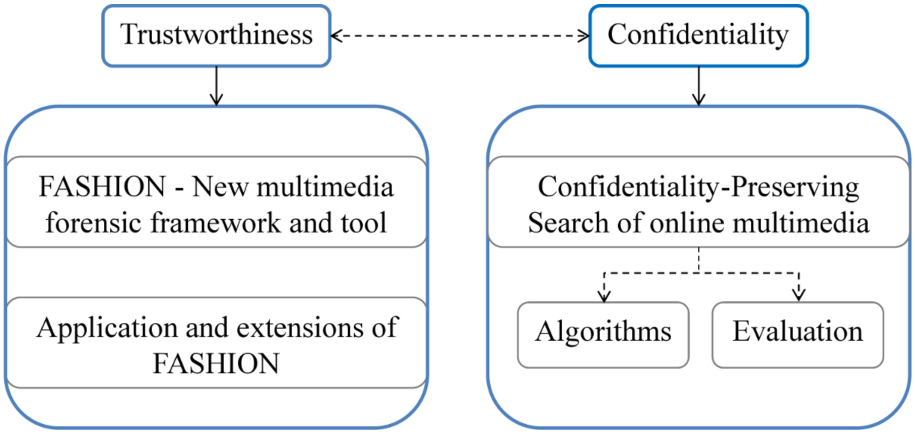

Trustworthiness
Confidentiality
FASHION - New multimedia forensic framework and tool
Application and extensions of FASHION
Confidentiality-Preserving Search of online multimedia
Algorithms
Evaluation

Figure 1.1: Dissertation outline

1.2.1 Forensic Hash for Multimedia Information Assurance

We propose a new multimedia forensic framework by utilizing side information called forensic hash. A forensic hash is a compact signature capturing important information from the original multimedia data for later forensic analysis. This new framework draws synergy from two related research areas, namely, image hashing and no-reference multimedia forensics. The forensic hash is nearly as compact as traditional image hash but can be used to answer a broader scope of forensic questions than a simple binary authenticity answer. Compared with no-reference forensics, the proposed forensic hash offers more efficient and accurate forensic analysis, and can answer questions that are difficult to answer in a blind scenario.

To avoid the dilemma of one-scheme-fit-all, the forensic hash takes a modular design such that different modules tackle different forensic questions separately and at the same time they complement each other and work together synergistically. This modular design brings the advantage of extensibility to forensic hash.

We proposed two novel constructions of forensic hash and demonstrate that they can provide robust and accurate estimation of geometric transform such as rotation and scaling on modified images. Furthermore, we also demonstrate the forensic hash's capability for locating tampering and estimating advanced editing operations such as seam carving, which is an adaptive image resizing technique.

Finally, we extend the spirit of forensic hash to the task of image quality assessment. By utilizing compact side information, we propose novel techniques for reduced-reference quality assessment on images that have undergone retargeting, which is the first endeavor on this problem to the best of our knowledge. The proposed quality metrics show positive correlation with human subjective ratings; furthermore, the proposed quality assessment algorithm can provide a detailed distortion map to assist human observers to make a personalized decision rather than accepting a single quality score as in conventional image quality assessment work.

1.2.2 Confidentiality-Preserving Search of Online Multimedia

To the best of our knowledge, our work is the first endeavor in the community to explore techniques for confidentiality-preserving content-based search of multimedia. This problem has unique challenges as compared to many existing secure computation works, in terms of large volume of multimedia data, requiring rank-ordered retrieval, demanding efficient computation and minimum user-involvement. We address this problem from a joint signal processing and cryptography point of view and propose efficient techniques with good security-efficiency trade-off.

The key techniques proposed in this work is distance-preserving randomization. By strategically utilizing techniques from image processing, information retrieval, and cryptography, we propose efficient randomization algorithms with good distance-preserving property. In addition to randomizing visual features for similarity comparison, we also explore randomization techniques for state-of-the-art search indexes. Experimental results demonstrate high search accuracy and efficiency of the proposed techniques.

We also carry out a quantitative study on the amount of randomness and confidentiality protection offered by our proposed techniques. A comprehensive comparison with cryptographic approaches based on homomorphic encryption is carried out, which demonstrates the pros and cons of various alternatives for the problem of confidentiality-preserving multimedia search. Such a comparative study provides valuable insight in designing other confidentiality-preserving computation techniques for various online applications that involve digital multimedia.

1.2.3 Dissertation Organization

The rest of the dissertation is organized as follows. In Chapter 2, we propose the framework of multimedia forensics using forensic hash. The main idea of the forensic hash and its relation with areas such as image hashing and non-reference forensics are discussed. We describe in detail the proposed hash constructions, namely, an alignment component for geometric transform estimation and an integrity component for tampering localization. Experimental results on discriminability, robustness, estimation accuracy of the forensic hash are presented at the end of Chapter 2. By extending the capability and spirit of forensic hash, we discuss in Chapter 3 two applications for retargeted images that have undergone adaptive resizing: one is estimating seam carving operation, and the other is reduced-reference quality assessment on retargeted images. In Chapter 4, we study the problem of confidentiality-preserving search of online multimedia. Randomization techniques for both the visual features and search indexes will be discussed in detail. The comparative study between the proposed techniques and conventional cryptography techniques is provided in Chapter 5, which demonstrates the different trade-offs between security and efficiency of different techniques. Finally in Chapter 6, we conclude and present some interesting research issues for future exploration.

CHAPTER 2

Forensic Hash for Multimedia Information Assurance

2.1 Background and Related Work

Recent years have witnessed rapid growth of mobile devices capable of capturing high quality images and videos, and social media networks that provide various media sharing and streaming services. These new trends have generated a huge amount of personal multimedia content and significantly increased multimedia consumption over the Internet and on various devices. Photos, videos, and recordings have long been used in news media as a vivid evidential representation of important events. However, the digital nature of multimedia data and the advancement of multimedia processing technologies have made it easy to modify the digital content. Multimedia data can be intentionally altered to create a forgery and convey a different meaning, so seeing is no longer enough for believing! For example, objects can be removed from or inserted into an image, and multiple pieces of content may be combined into a new creation. As such, it is critical to evaluate the trustworthiness of multimedia information and reveal its complete processing history in order to achieve better decision and usage of online multimedia information.

There are two traditional techniques to evaluate image trustworthiness and authenticity, namely, robust image hashing [36, 53, 74, 105, 109] and blind multimedia forensics [24]. In the scenario of point-to-point image authentication, the sender attaches a short signature or hash with the image, and the receiver computes the hash from the received image and compares it to the attached hash according to some distance measure. A small distance indicates the received image is authentic, while a large distance implies the received image is a different image or has undergone significant tampering. Such a simple distance comparison using image hashing answers mainly the binary question of image authenticity, and it is challenging to achieve a good trade-off between being robust against global operations, locating local tampering, and keeping the hash length short. On the other hand, existing research in multimedia forensics mainly tries to determine the origin and detect potential tampering for digitally acquired images/videos without proactive aids such as hash attachment or embedded watermark [24]. Such forensic techniques typically explore unique signal traces left on the content by potential processing operations, but the lack of any side information about the original data makes many tasks computationally intensive, as exhaustive search in a large parameter space is often required and the achieved accuracy can be limited. Furthermore, since the intrinsic signal traces are publicly accessible and no secret key is involved, these non-intrusive forensics can be vulnerable to anti-forensic attacks.

Considering these advances and limitations of related techniques, an important research question is to explore by appending a short string as in the conventional hashing applications, whether we can use such additional information to augment the capabilities of both conventional hash and non-intrusive forensics to determine the processing history that the source data has undergone with improved accuracy and efficiency. Such capabilities of evaluating the integrity, provenance, and processing history can enable us to assess the trustworthiness of multimedia data at a much more flexible and fine level while avoiding the dilemma of one-size-fits-all designs. We refer to such a new forensic framework as "Forensic Hash for Information Assurance", or FASHION in short. Below we first review the related literature in the area of image hashing and blind multimedia forensics, and then summarize the main idea and contribution of our work.

Related work on robust image hashing Robust image hashing is an extension from traditional cryptography hash. A cryptography hash is used to evaluate document authenticity and is sensitive to a single bit difference, while the image hash is designed to be similar across visually similar images that may have undergone moderate content preserving operations from a same original image but sensitive against malicious content tampering. The distance between two image hashes is compared with a threshold to determine whether the received image is authentic. Image hashing algorithms typically involve feature selection, quantization and compressive encoding. Some of the features that have been used in the literature include block intensity averages [120] or histograms [19,32,48], image edge information [86], DCT coefficients [58], the scale interaction model with the Mexican-Hat wavelets [8], Fourier-Mellion features for geometric resilience [105], median points from the Radon projections [35], geometry preserving local feature points [73], projections onto smooth random patterns [36], matrix invariants through singular value decomposition [53], and non-negative matrix factorization [74]. For applications that require a hash to be difficult to be guessed or forged, randomization is applied to different stages of hash construction, such as feature extraction, quantization, and/or encoding [36,74,105,109].

Robust image hashing can be used for image authentication, but a simple binary decision of authenticity is often inadequate. For example, an image that has undergone a small amount of local edits may be considered as a different image using existing image hashing, but the rest of the image content may still be trustworthy and contain valuable information. It is desirable to provide more information about the processing history of the multimedia data, so that the end users can have the flexibility in determining whether to trust the image content and how to utilize the received data for specific applications. Some recent work on robust image hashing can provide certain forensic capabilities. Roy and Sun [88] incorporate Scale Invariant Feature Transform (SIFT) features [61] into a hash for geometric registration of the received image with respect to the original and enable more reliable tampering localization than prior art. Lin et al. [59] apply distributed source coding to encode information about the original image and employ EM algorithm to estimate potential operations on the image. These prior techniques have some limitations: the use of SIFT feature in [88] results in considerable increase in the hash length and the image registration is not possible when the selected SIFT points are not available in the received image; and the EM algorithm used in [59] is computationally intensive because separate algorithms need to be applied for each specific type of operations that the image may have undergone.

Related work on multimedia forensics The research objective of multimedia forensics is to provide tools for analyzing the origin, processing history, and trustworthiness of multimedia information. Recent research in multimedia forensics can determine whether a received image/video has undergone certain operations without access to the original data. This is accomplished by analyzing intrinsic traces left by devices and processing, and by identifying inconsistencies in signal characteristics [31, 96]. Bayram et al. [6] and Swaminathan et al. [106] tried to identify the model of the camera by learning color filter array (CFA) interpolation patterns from images that are taken by the camera [6, 106]. Lukas et al. proposed to use the camera sensor imperfections as a unique signature to provide linkage between an image and its capturing device [70]. Popescu and Farid estimate the re-sampling factor of an image by examining the linear dependencies among image pixels resulted from the re-sampling process [84]. Malicious tampering of the image through cut-and-paste can also be detected by examining inconsistencies in signal statistics. Farid [30] performs image tampering localization by exploiting the inconsistencies of JPEG quality, as operations such as cut and paste often leave areas with different

JPEG quality factors in the same image. Another useful type of inconsistency is the directions of lighting and shadow. Johnson et al. evaluate the trustworthiness of an image by estimating the lighting direction in different parts of the image [49] . An authentic image has consistent lighting direction, while the cut-and-paste operations usually bring to the image some new content that has different lighting configuration. The above mentioned blind forensic work provide valuable tools to evaluate multimedia trustworthiness, but they have limitations in terms of the accuracy levels and the scope of forensic questions that can be answered. Such operations as cropping and rotation can be difficult to estimate without any side information about the original image. Many signal statistics and traces left by image operations may be removed or altered by further post-processing. A considerable amount of computational complexity is also involved in most blind forensic analyses.

Main idea and contribution of our work Given that the conventional image hashing only provides a binary authentication answer using simple distance comparison, and the blind forensics techniques have limitations in terms of the scope of questions that can be answered and the computational complexity, we propose the FASHION framework [65,67,68] to bridge these two research areas and combine their benefits. The FASHION framework uses side information called forensic hash to assist forensic analysis, and its relation with the other two research areas is shown in Fig. 2.1.

A forensic hash is designed to be nearly as compact as a conventional image hash, but instead of providing a binary decision, the generation and utilization of

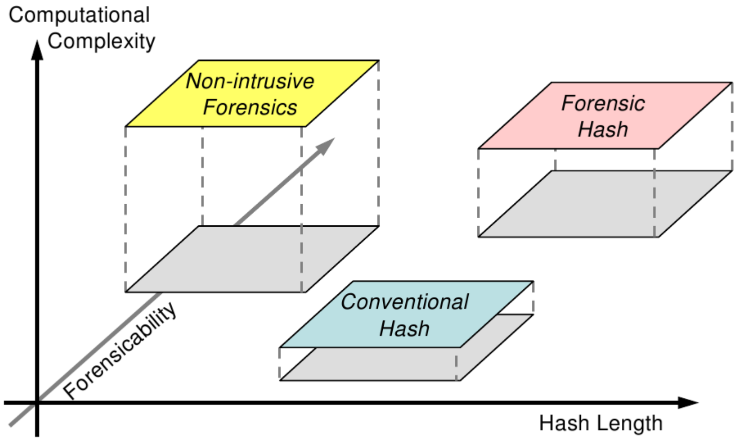


Figure 2.1: Forensic hash as compared to image hashing and blind forensics

forensic hash are designed to reveal more information about the processing history in terms of the likely types and the associated parameters of the processing operations. Compared to blind forensics, forensic hash has the advantage of being able to answer a broader scope of questions in a more accurate and efficient way.

The main contributions of our work include: a new framework of multimedia forensics by using side information represented via a compact hash; a modular design of forensic hash to address different forensic questions in a flexible and extensible way while avoiding the one-scheme-fits-all dilemma; two novel constructions of forensic hash that provide robust estimation of geometric transform such as rotation and scaling. The proposed forensic hash also achieves higher image discrimination capability than representative prior art.

The rest of the chapter is organized as follows: Section 2.2 explains the overall framework of multimedia forensics analysis using forensic hash. In Section 2.3, we present two constructions of forensic hash, based on Radon transform, scale space theory and visual words representation of SIFT features, respectively. In Section 2.4, we discuss the tampering localization capability of the forensic hash construction. The experimental results including image discrimination, geometric transform estimation, and tampering localization are presented in Section 2.5. Summary of the chapter is given in Section 2.6.

2.2 FASHION Framework

The objective of the proposed FASHION framework is to achieve efficient and accurate multimedia forensics by properly designing forensic hashes to capture important side information from the original image. We illustrate the role of forensic hash in forensic analysis and its modular design in Fig. 2.2 and Fig. 2.3 respectively.

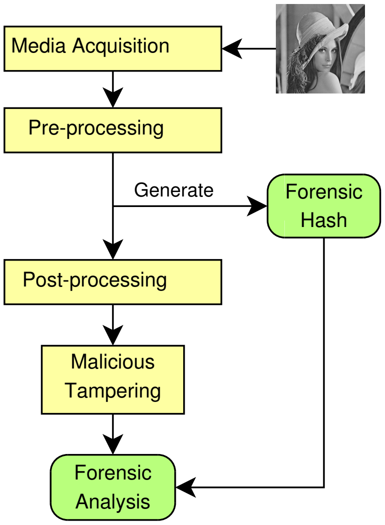


Figure 2.2: Flowchart of media processing

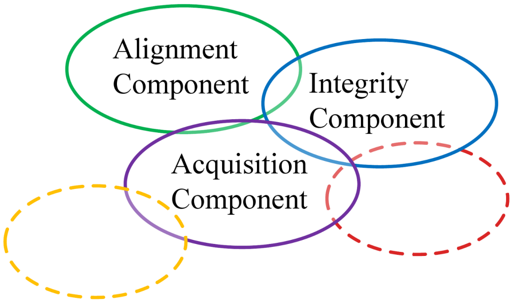


Figure 2.3: Modular design of forensic hash to reveal image processing history

After an image is captured by an imaging device, it may undergo certain preprocessing operations inside the device, such as white balance adjustment and color correction. The image will then be available in digital format for distribution and consumption. In order to evaluate the trustworthiness of a future received image and reveal potential operations during the distribution, a forensic hash can be generated at a point before the transmission. When the image is being distributed through different types of networks, such as P2P networks and mobile networks, to various receiving devices, some adaptations to the image format and content may occur. For example, the image may be resized and cropped for different screen sizes; logos may be inserted to image corners. In addition to these necessary adaptations, there can be malicious tampering that alters the image content to convey a different semantic meaning. The role of forensic hash is to be securely attached along with the transmitted image and assist the forensic analysis on the received image.

In order to better evaluate the trustworthiness of a received image, it is beneficial and sometimes necessary to gain more knowledge about the processing history of the data. For example, estimating the geometric transform that an image may undergone, such as rotation angle and scaling factor, is extremely helpful to align the received image with the original image so that further localization of the tampering can be easily performed. In traditional image hashing, the design goal is to extract features that are robust to allowable image operations. In many cases, it is difficult to find a feature that is robust to a wide range of operations while still be highly discriminative. For forensic hash, the goal is to detect the presence of image operations and estimate the associated parameters if possible. In order to avoid the problem of one-scheme-fits-all as in many prior image hashing work, a modular and multi-resolution design is desirable. Such a design includes several modules each focusing on some class of forensic tasks at various resolutions, while at the same time complementing each other and working synergistically. For example, an alignment component can be used to estimate geometric transform and provide a global view of an image to tell whether it is the same image modified from the original image or a completely different image. An integrity component then acts upon the global view to provide a finer resolution analysis on image integrity through localized approaches. Such a modular and integrated structure also enables easier design of other forensic modules to extend forensic capabilities in the future.

Unlike traditional image hashing, forensic hashes from similar images do not have to be similar in terms of smaller distance between the two hashes. Instead, efficient algorithms are designed to analyze both the forensic hash and the received image to identify potential operations that the image may have undergone and estimate the parameters of such operations. Desirable properties for forensic hash include robustness, scalability, and distinctiveness. Robustness means that the forensic hash can provide accurate analysis for images that have undergone multiple operations and strong tampering; scalability ensures that the performance of forensic analysis can be improved as we include more side information; distinctiveness implies that it is difficult to find two different images that give similar forensic hashes. The forensic hash is designed to be as compact as possible and enable forensic analysis with low complexity, which makes it possible to be used in band-limited channels and on small devices such as mobile phones.

2.3 Hash Construction for Geometric Alignment

The FASHION framework that we described in previous section is modular in nature. The forensic hash of an image can contain various forensic components designed for different forensic tasks at different resolutions and work together in a synergistic way. The modular design of forensic hash provides flexibility to suit for different applications. In this section, we first present the proposed algorithms for the alignment component of forensic hash, whose role is to estimate geometric transforms, such as rotation and scaling, and align the modified image with the original image to allow further forensic analysis. We then briefly present the constructions and usage of an integrity check component for cropping estimation and tampering localization.

2.3.1 Alignment based on Radon Transform and Scale-Space Theory

Geometric transforms such as rotation and scaling are common post-processing operations and the estimation of such transform parameters are important in order to compare the original and modified images on a common ground. Prior work [59, 88] on image hashing either result in considerable hash length to incorporate the geometric registration information or require high computational cost to estimate the transform parameters. In this subsection, we propose two constructions for the alignment component. The first construction is based on Radon transform and scale space theory, while the second one builds on robust SIFT features. The two constructions offer different trade-offs between robustness and compactness, which will be discussed below in more detail.

To align a received image with its original version, we exploit Radon transform for its nice property of separating scaling and rotation. Radon transform is a line integral of an image along certain directions and is a useful tool for image registration [115] and authentication [35]. Such line integral captures salient information about the image alone particular directions, and is robust to small variations in the image content, which may come from noise, moderate cropping, local tampering, and content preserving operations such as filtering, brightness/contrast adjustment. We use a compact summarization along the angular axis in the transform domain for rotation estimation and employ scale space theory to identify scale-resilient features along projections at different directions for scaling estimation. The overall block diagram for the alignment component construction is shown in Fig. 2.4, and we discuss the details next.

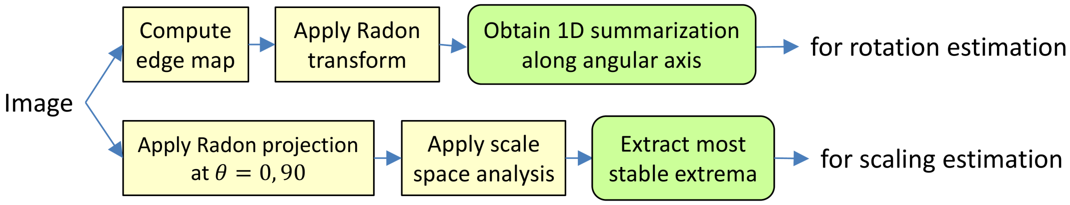


Image
Compute edge map
Apply Radon transform
Obtain 1D summarization along angular axis
for rotation estimation
Apply Radon projection at $\theta = 0, 90$
Apply scale space analysis
Extract most stable extrema
for scaling estimation

Figure 2.4: Block diagram for the alignment component based on Radon transform and scale space theory (Round corner boxes constitute the alignment component)

Rotation Estimation The direction of image edges can reveal information about image orientation. For an original image $I(x, y)$, we first compute its edge map $E(x.y)$ using Canny edge detector [17]. Radon transform is then applied on the edge map $E(x, y)$. Radon transform of an image is essentially a line integral of that image along certain directions, defined as follows:

$$R_E(\rho, \theta) = \int_{-\infty}^{\infty} E(\rho \cos\theta - u \sin\theta, \rho \sin\theta + u \cos\theta) du. \qquad (2.1)$$

Given image I' which is obtained from I by rotating α degrees counter-clockwise, its edge map E' would give a Radon transform $R_{E'}(\rho, \theta) = R_E(\rho, \theta + \alpha)$. Thus, in the transform domain, rotation becomes a shift along the angular axis. To estimate the rotation angle, we extract a 1-D summarization of the Radon transform along the angle axis. For Radon transforms $R_E(\rho, \theta)$ and $R_{E'}(\rho, \theta)$, the 1-D summarization is derived as $\mathbf{m}(\theta) = \max_\rho(R_E(\rho, \theta))$, and $\mathbf{m}'(\theta) = \max_\rho(R_{E'}(\rho, \theta))$. For compact representation, quantization and subsampling are applied to $\mathbf{m}(\theta)$. Since downsampling may cause aliasing, we first pass the signal $\mathbf{m}(\theta)$ through a low-pass filter $f_{low}(\cdot)$ to obtain $\hat{\mathbf{m}}(\theta) = f_{low}(\mathbf{m}(\theta))$. If an n-byte alignment component is desired, we downsample the signal $\hat{\mathbf{m}}(\theta)$ to obtain the forensic hash $\mathbf{h} = \{h(1), \cdots, h(n)\}$ with $h(i) = \hat{m}(\lfloor (i-1) \cdot \frac{180}{n} \rfloor)$, $i = 1, 2, \cdots, n$.

When estimating the geometric transform for image I', the Radon transform of its edge map $R_{E'}(\rho, \theta)$ and the 1-D summarization $\mathbf{m}'(\theta) = \max_\rho(R_{E'}(\rho, \theta))$ are generated accordingly. In order to compare with the alignment component $\mathbf{h}$ in the forensic hash of the original image, $\mathbf{m}'(\theta)$ will be passed through a low-pass filter and downsampled at different shift positions to obtain $\mathbf{h}'(\phi) = \{h'(1), \cdots, h'(n)\}$ with $h'(i) = \hat{m'}(\lfloor (i-1) \cdot \frac{180}{n} \rfloor + \phi)$, $\phi = 0, 1, \cdots, 179$. The shift amount that maximizes the cross-correlation between $\mathbf{h}$ and $\mathbf{h}'(\phi)$ is considered as the rotation angle between the two images I and I', i.e.

$$\alpha = \arg\max_{\phi} \sum_{i=1}^{n} h(i)h'(i+\phi),\ \ \phi \in \{0, 1, \cdots, 179\}. \tag{2.2}$$

To further compress the forensic hash, we apply ordinal ranking to $\mathbf{h}$ and store only the rank order information, i.e. $\text{rank}(\mathbf{h}) = \{r(1), \cdots, r(n)\}$ where $r(i) \in \{1, \cdots, n\}$ is the rank of $h(i)$. Given $\mathbf{h}'(\phi)$ of I', its rank order information is denoted by $\text{rank}(\mathbf{h}'(\phi)) = \{r'(1), \cdots, r'(n)\}$. The shift amount that minimizes the L_1 distance between $\text{rank}(\mathbf{h})$ and $\text{rank}(\mathbf{h}'(\phi))$ will be the estimated rotation angle between the two images I and I', i.e.

$$\alpha = \arg\min_{\phi} \sum_{i=1}^{n} |r(i) - r'(i+\phi)|,\ \ \phi \in \{0, 1, \cdots, 179\}. \tag{2.3}$$

Experimental results in Section 2.5.3 show that rotation estimation using rank order information gives performance comparable to estimation using cross-correlation, and a proper fusion of the two similarity metrics in (2) and (3) can lead to further improved estimation accuracy.

Scaling Estimation Given the original image I and its scaled version I' with scaling factor s, their Radon transforms have the property that the Radon projections at any particular angle θ, $f_\theta(\rho) = R_I(\rho, \theta)$ and $f'_\theta(\rho) = R_{I'}(\rho, \theta)$, have the same scaling factor s, i.e. $f_\theta(\rho) = s \cdot f'_\theta(s \cdot \rho)$. However, this ideal scaling relation may not be exactly satisfied when the image has undergone additional cropping, local tampering, and other image processing operations such as filtering and contrast enhancement. It is necessary to investigate a robust representation of Radon projections that is resilient to these operations.

We propose to use scale space features of the 1-D signals $f_\theta(\rho)$ and $f'_\theta(\rho)$ to address this problem. Scale space theory [60] is a powerful tool for analyzing signals at different scales, making it useful for automatic scale selection and scale invariant image analysis. Given $f_\theta(\rho)$ of the original image at a particular θ, we generate its scale space representation $L(\rho; t)$ by convolving $f_\theta(\rho)$ with a 1-D discrete Gaussian filter $g(\rho; t)$ at scale t:

$$L(\rho; t) = g(\rho; t) * f_\theta(\rho), \text{ where } g(\rho; t) = \frac{1}{\sqrt{2\pi t}} e^{-\rho^2/(2t)}. \tag{2.4}$$

The scale space representation is a 2-dimensional signal with higher value of t indicating coarser scale.

With $L(\rho; t)$ computed, we then locate the extrema of $L(\rho; t)$ at each scale t by detecting the zero-crossing positions of $\partial L(\rho; t)/\partial t$ for each t. For the Lena image shown in Fig. 2.6, its Radon projection $R(\rho, \theta)$ along the vertical direction $f_0(\rho)$ is shown in Fig. 2.7. The local extrema across scales in the scale-space representation of $f_0(\rho)$ are illustrated in Fig. 2.5, where horizontal direction represents the signal and the vertical direction represents the scale. Extrema positions are marked black, and the scale becomes coarser from top to bottom . Smoothing using Gaussian kernel has the property that no new extrema will be created in coarser scales [60], which means that the number of extrema will be fewer in the coarser scale and their evolution over scales will never cross each other.

We can see from Fig. 2.5 that smoothing can cause the extrema to drift at coarse scales. In order to locate each extremum accurately, we trace the extrema from the coarse scale to the finest scale and use its position at the finest scale.

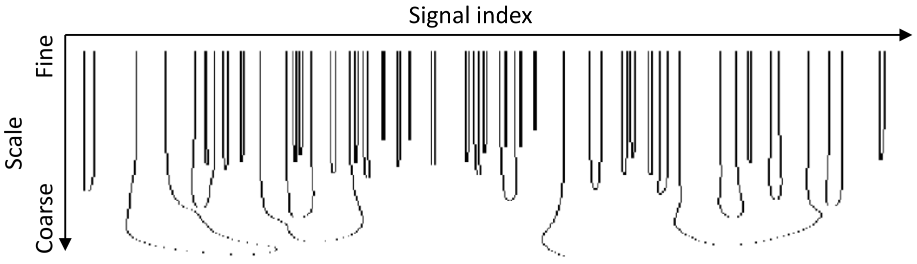


Figure 2.5: Space extrema across scales: horizontal direction represents the signal and the vertical direction represents the scale, with fine scale at the top and coarse scale at the bottom.

During tracing, we also compute the lifetime of a space extremum and denote it as $\log(t_D)$, where t_D is the scale at which the extrema disappears. Since the scale used in the Gaussian kernel is exponentially sampled, using logarithm ensures that signals at different scales are treated in a similar way [60]. Extrema with longer lifetime are expected to capture more important information about the signal and are more robust against local variations of the signal. The ten extrema of $f_0(\rho)$ with longest lifetime are shown in Fig. 2.7, which captures the most stable extrema while omitting small variations in the signal.

After computing $f_\theta(\rho)$ from the original image, the alignment component is augmented for scaling estimation by recording the positions of the n extrema of $f_\theta(\rho)$ that have the longest lifetime across scales, where n can be determined by the desired hash length. Since the extrema with long lifetime are expected to be stable after scaling even with cropping and local tampering, we use their positions to estimate the scaling factor s through the RANSAC algorithm [34]. Given the extrema positions of the two signals $f_\theta(\rho)$ and $f'_\theta(\rho) = s \cdot f_\theta(s \cdot \rho)$, we randomly choose

Figure 2.6: Lena image

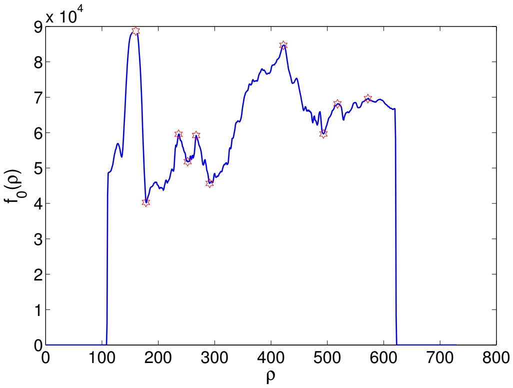


Figure 2.7: The 10 most stable extrema of $f_0(\rho)$ are shown in star. $f_0(\rho)$ is the Radon projection of the edge map of image Lena along the vertical direction.

two extrema x, y from $f_\theta(\rho)$ and two extrema x', y' from $f'_\theta(\rho)$ in each iteration of the RANSAC algorithm. An estimate $\hat{s}$ of the true scaling factor s is given by the ratio of $|x' - y'|/|x - y|$. We then scale the signal $f'_\theta(\rho)$ based on $\hat{s}$ and align it to the original extrema from $f_\theta(\rho)$ to count the number of matched extrema between the scaled $f'_\theta(\rho)$ and $f_\theta(\rho)$. After a given number of iterations or when the number of matched extrema exceeds certain threshold, the scaling factor that gives the maximum number of matched extrema between $f_\theta(\rho)$ and the scaled $f'_\theta(\rho)$ will be the estimated scaling factor s^*. By computing the Radon projections of the original image along both the vertical and horizontal directions, i.e., $\theta = 0$ and 90, we can obtain the scaling factors along these two directions using the above method. In Section 2.5, we report the experimental result which shows good performance of geometric transform estimation using the proposed alignment component.

2.3.2 Alignment based on Visual Words Representation of SIFT

Despite the good robustness and compactness, the alignment component based on Radon transform has some limitations. First, representing a two-dimensional image by a one-dimensional Radon projection may lose discriminative information of the image. Second, a large amount of cropping and local tampering may affect the Radon projections along all directions and make the geometric transform estimation less accurate. To overcome these limitations, we propose a second construction of the alignment component based on salient local features and their visual words representation. The basic idea is to extract more localized features across the image, and such features capture geometric information about the image and remain robust against common image processing and tampering operations. Using local features to complement global features such as Radon projection can improve the robustness of geometric transform estimation against larger amount of cropping and local tampering. The key challenge is to select proper local features and develop their compact representation. In this section, we propose a novel algorithm that uses robust SIFT points [61] as local features while encoding their high dimensional descriptors into a compact visual words representation for forensic analysis.

Visual words for FASHION The visual words representation for multimedia documents was proposed by Nistér and Stewénius [76] for efficient object recognition and retrieval over large databases. As shown in Fig. 2.8, to generate the visual words representation, visual features are first extracted from the image or video to describe its local appearance. Visual features can be color histograms, shape descriptors, and invariant region descriptors such as SIFT. Each of the feature vectors is then hierarchically clustered based on a vocabulary tree and assigned to the most similar leaf node in the tree. The vocabulary tree is trained using feature vectors from a set of training images. The leaf nodes in the tree are called "visual words", and a multimedia document is represented as a set of visual words, which is analogous to the bag of words representation in text retrieval.

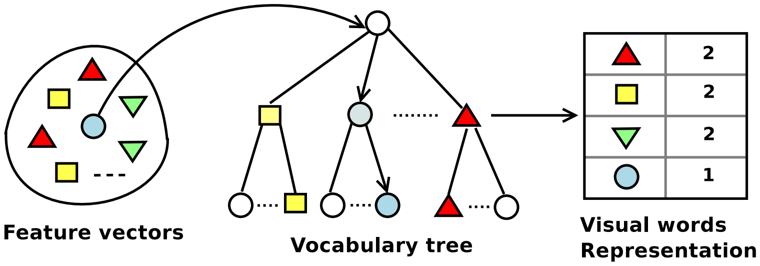


Figure 2.8: Visual words representation for multimedia

In this work, we use SIFT as the local feature to construct forensic hash. SIFT descriptors are designed to be invariant to affine transformations. A set of affine invariant salient points are first extracted from the scale-space representation of the image, and then a 128-dimensional descriptor is computed for each point using the gradient information of a region around the point. Each SIFT descriptor has an associated characteristic scale and dominant orientation. We propose to use the scale and orientation values between matched SIFT points of the original and the modified images to determine the scaling factor and the rotation angle between the two images. The overall block diagram for the alignment component construction is shown in Fig. 2.9, and we discuss next how to perform geometric transform estimation in the following paragraphs.

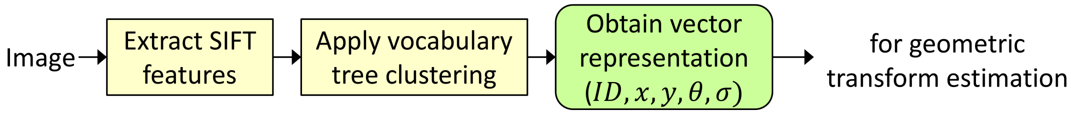


Image
Extract SIFT features
Apply vocabulary tree clustering
Obtain vector representation $(ID, x, y, \theta, \sigma)$
for geometric transform estimation

Figure 2.9: Block diagram for the alignment component based on visual words of SIFT features. (Round corner box constitutes the alignment component)

Since SIFT descriptors are high-dimensional vectors, the number of SIFT points that can be included in a compact hash representation is limited. Even by projecting the SIFT descriptor to a 60-bit string as suggested by a recent work by Roy and Sun [88], a hash of length 1000 bits can contain only around 5-10 points [88]. To address this challenge, we represent the SIFT features of an image using a bag of visual words so that only a compact form of the visual word labels but not the full descriptors need to be encoded, and such an encoding/compression strategy significantly increases the number of SIFT points that can be included for forensic analysis. This compact representation using visual words requires the same vocabulary tree to be available to both the sender who generates the forensic hash and the receiver who performs the forensic analysis. This assumption is reasonable because the vocabulary tree only needs to be generated once or can be made publicly available for download. As shall be seen from the experimental results in Section 2.5, using visual words representation plays a critical role in keeping the hash compact and providing more SIFT points for accurate and robust forensic analysis.

Geometric transform estimation Given an original image, we first extract its SIFT points and sort them based on their contrast values. SIFT points with higher contrast values are typically more stable against image operations such as rotation, scaling, and compression. For compactness, we select only SIFT points with contrast values above a certain threshold. This threshold can be adjusted to control the size of the alignment component. As an example, the most stable SIFT points for the Lena image are shown in Fig. 2.10.

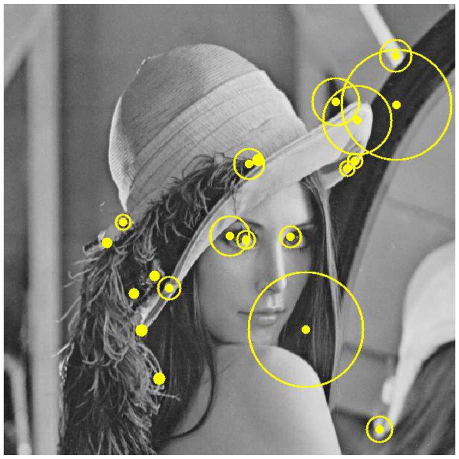


Figure 2.10: SIFT points with contrast value > 0.05. The size of the circle corresponds to the characteristic scale of the SIFT point.

The most stable SIFT points are then assigned to different visual words by hierarchically clustering using the vocabulary tree. Each point is represented by a vector of 5 parameters, which is denoted as $(id, x, y, \sigma, \theta)$: the visual word label, the x and y positions in the image, the scale at which the point is detected, and the dominant direction of the point. The vectors of all the selected SIFT points form the alignment component of the forensic hash. For a vocabulary tree with 1000 visual words and an image size of 1024x1024, each of the 5-parameter vector would take around 50 bits after proper quantization. Compared to the 80 bits per SIFT point in [88], our proposed alignment component can encode roughly *twice* the amount of SIFT points at the same hash length. Furthermore, it should be noted that the increased number of SIFT points can not only improve the robustness of geometric transform estimation, but also contribute to locating tampering and allow more compact representation of the integrity check component, which will be discussed in Section 2.4.

To estimate the geometric transform applied to a modified image, we extract SIFT points at the same contrast threshold and generate the 5-parameter vector for each of their SIFT points. We first find matching points between the modified image and the original image by identifying vectors with the same ID and that occur only once in both images. These points are denoted by $(p_1, \tilde{p}_1), (p_2, \tilde{p}_2), \cdots, (p_n, \tilde{p}_n)$, where p_i are salient points encoded in the alignment component and $\tilde{p}_i$ are salient points extracted from the modified image. Each matching pair gives an estimate of the scaling factor and rotation angle as $\sigma_i = \sigma(\tilde{p}_i)/\sigma(p_i)$ and $\theta_i = \theta(\tilde{p}_i) - \theta(p_i)$, where $\sigma(p)$ and $\theta(p)$ are the scale and orientation parameters of the SIFT point p. There can be false matching pairs because image processing operations can affect the SIFT descriptors, which may cause a SIFT point to be assigned to a different visual word or result in point addition and deletion. Robust estimation such as RANSAC [34] is then applied to estimate the actual scaling factor $\hat{\sigma}$ and rotation angle $\hat{\theta}$ from $\{\sigma_i, \theta_i\}$ and identify false matching pairs. This estimate is further refined by considering the remaining matched SIFT points that occur more than once in both images. Assuming that the majority of the matching points are correct matches or the false matches do not give consistent estimates, the robust estimation algorithm can provide accurate estimation for the geometric transform parameters. It should be noted that further savings in the hash length can be achieved by representing each point with only three parameters (id, x, y) and using only the point positions to estimate the geometric transform. The computational complexity of such geometric transform estimation will be higher but the hash can be more compact.

2.4 Hash Construction for Tampering Localization

The above proposed alignment component allows compact representation and enables robust estimation of geometric transforms that an image may have undergone. Estimating scaling and rotation is an important step in multimedia forensics because they are very common operations in image editing and tampering, but can be difficult and computationally expensive to estimate using traditional non-reference forensic techniques. Furthermore, knowledge of the transformation parameters allows us to compare the modified image with the original image on a common ground, and facilitate further forensic analysis. In this section, we describe how the alignment component of the forensic hash can enable efficient image cropping detection and facilitate the integrity component for tampering localization.

2.4.1 Cropping Estimation

Cropping can be used by an attacker to remove important information on the boundary of an image, and many image hashing schemes are sensitive to misalignment caused by cropping. Cropping is also difficult to detect using non-reference forensic techniques, because non-reference forensics typically rely on the traces or statistical changes left by certain operations, while cropping operation does not change statistical properties of the remaining part of the image. However, we will show that with the help from the alignment component proposed above, cropping operation can be easily detected and estimated.

With the alignment component based on Radon transform and scale space theory, the rotation angle of the modified image is first estimated using the 1-D summarization of the Radon transform along the angular axis. The scaling factors along the horizontal and vertical directions are estimated using the stable extrema from the Radon projections $f_0(\rho)$ and $f_1(\rho)$ of the image along vertical and horizontal directions, respectively. Cropping on the left and right boundaries of the image incurs the corresponding amount of cropping on the respective boundaries of the signal $f_0(\rho)$. Similarly, cropping on the top and bottom boundaries leads to the corresponding amount of cropping on the respective boundaries of the signal $f_1(\rho)$.

Given a modified image I', we compute its Radon projections along the vertical and horizontal directions to obtain $f_0'(\rho)$ and $f_1'(\rho)$. The positions of the most stable extrema of $f_0'(\rho)$ and $f_1'(\rho)$ are also computed. After a moderate amount of cropping, the majority of the extrema in $f_0(\rho)$ and $f_1(\rho)$ are still available in $f_0'(\rho)$ and $f_1'(\rho)$. We can align the original extrema with the extrema from f_0' and f_1' such that the number of matched extrema is maximized, as described in Section 2.3.1. An example of the alignment is shown in Fig. 2.11, where the modified image is a scaled and cropped version of the original Lena image. Once the two signals, f_i and f_i', $i \in \{0, 1\}$, are properly aligned, the amount of cropping can be obtained by comparing the distance between the corresponding boundaries of the two signals.

In Fig. 2.11, *dl* and *dr* are the estimated amount of cropping on the left and right boundaries of the original image.

With the alignment component based on visual words representation of SIFT features, rotation and scaling are estimated using the orientation and scale information between the matched pairs of SIFT features from the two images. Once the modified image is transformed using the estimated parameters, the two images can be aligned using the position information of the matched pairs of SIFT points and cropping can be immediately estimated from the differences in the image sizes. It should be noted that the accuracy of cropping estimation depends on the accuracy of the geometric transform estimation, since the modified image needs to be aligned with the original image through rotation and scaling.

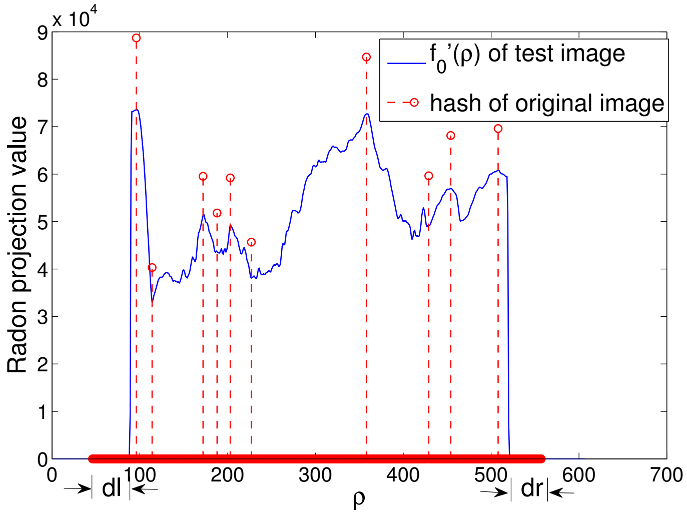


Figure 2.11: Cropping estimation by aligning extrema from original and modified signals

2.4.2 Block-based Tampering Localization

In addition to cropping, an image can also be locally tampered by operations such as cut and paste. These local tampering operations are important targets of multimedia forensic research. Traditional non-reference forensic techniques typically rely on the inconsistencies in image statistics caused by the local tampering operation. Such inconsistencies can be due to different JPEG quality factors [30], different color filter array (CFA) patterns [6], or different lighting directions [49] within the same image. Disadvantages of non-reference forensic techniques include relatively high computational complexity and that smarter forgeries may avoid introducing certain types of statistical inconsistencies and evade detection. If side information can be attached, the side information can include an integrity check component that is designed to allow for more efficient tampering detection and make it more difficult to evade detection.

Encoding block-based features from the original image has been commonly employed in the literature to detect local tampering. During integrity verification, block-based features are extracted from the testing image and block-wise comparison is conducted to reveal potential local tampering. Compared with tampering localization based on inconsistencies in image statistics, as used in non-reference multimedia forensics, block-wise comparison can locate tampering in a more efficient and accurate way. For this approach to work well, the two images need to be properly aligned before block comparison. The alignment component proposed in the previous section offers necessary geometric registration and therefore provides a common ground to enable accurate block-based tampering localization.

Since a tampered part of an image usually has significant difference from the original in terms of their gradient information, features such as edge direction histogram in a block have been used for tampering localization with very good results [88]. Using edge direction histogram and quantizing the pixel gradient into a few representative directions provide robustness against small rotation and scaling effect. In this work, we adopt such an approach to quantize pixel orientation into four directions (horizontal, vertical, diagonal, and anti-diagonal) and compute the edge direction histogram for each block. These edge direction histograms form the integrity check component of the forensic hash. In Section 2.5, we will examine the effectiveness of block-based edge histograms for tampering localization over a larger database and the effect of different quantization methods. We will show that compared to simple uniform quantization used in [88], non-uniform quantization of the histogram can provide enhanced performance without increasing hash length.

2.4.3 Hybrid Scheme for Tampering Localization

To achieve higher resolution of tampering localization using block-based feature, we need to use a smaller block size and thus a higher number of blocks, which will considerably increase the size of the integrity check component. We propose a hybrid construction for the integrity check component, which utilizes both the block based features and the alignment component to achieve a more compact representation. The alignment component based on visual words representation includes SIFT features with high contrast values. Since these SIFT features are robust against content preserving operations, if some SIFT features are missing or new SIFT features are introduced in the modified image, the location of these SIFT changes can be an indicator of potential image tampering. Block based features can then be used to detect tampering in regions that do not have highly stable SIFT points. The idea of hybrid construction is shown in Fig. 2.12, where SIFT points and their characteristic scales are marked using circles. Changes to the SIFT features will indicate potential tampering to the image regions covered by the dashed blocks, while tampering in the remaining areas will be revealed by block based features, as shown by the flowchart in Fig. 2.13.

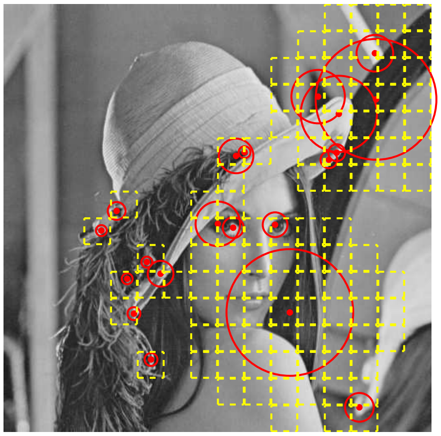


Figure 2.12: Hybrid construction of the integrity check component

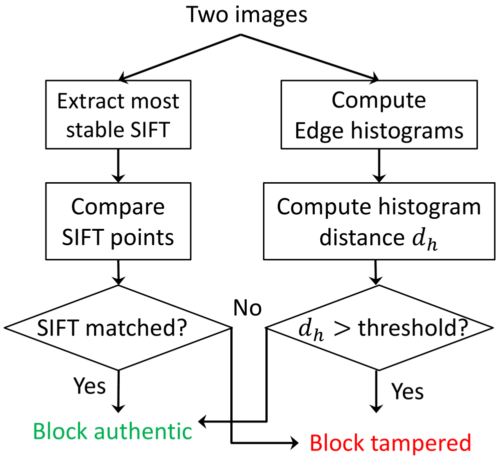


Figure 2.13: Flow diagram for integrity check using hybrid construction

SIFT points are typically extracted at different scales and the corresponding descriptors capture the gradient information over image regions of different size. Therefore, the addition or deletion of SIFT points indicates potential tampering to the image regions covered by the corresponding SIFT points. Utilizing SIFT points from alignment component for tampering detection helps reduce the size of integrity check component as we generate edge direction histograms only for blocks that are not covered by any detected SIFT points. Depending on the number of SIFT points selected and their distribution over the image, the savings of hash length may vary for different images, and this will be examined experimentally in Section 2.5.

2.4.4 Securing Forensic Hash

Forensic hash can be secured to prevent unauthorized forgery using established cryptography techniques such as symmetric encryption or public-key based digital signature. In the traditional two-party communication scenario, one party sends an image through an untrusted channel to the second party, who performs forensic analysis on the received image to evaluate its trustworthiness. This scenario is similar to image authentication considered in traditional image hashing schemes, where the sender either attaches the encrypted robust hash to the image or sends the hash through a separate secure channel to the receiver for authentication. The same strategy can be used here to secure the forensic hash. Since only the sender and receiver would know the secret key used for encrypting the forensic hash and the separate channel is considered secure, it is very difficult for an adversary to modify the image without being detected by the forensic analysis.

With the advancement in information technology, multimedia consumption has gone far beyond two-party communication. A more common scenario would involve multiple parties, where one authority party, such as news agencies or big media websites, provides multimedia data and distributes them over various network channels to different users. Some of the channels may be untrusted and involve potential tampering. The role of forensic hash is to be securely attached to the distributed images and allow easy verification on the trustworthiness of received image by different receivers. To allow verification, each receiver should be able to use the forensic hash and be sure that the hash has not been modified in any way. Digital signatures [87] can be used in such a scenario. The sender uses a secret private key to sign the forensic hash of an image to be distributed. Then given an image claimed to come from a specific source, any receiver can obtain a public key from the source to verify the authenticity of the forensic hash. If the hash is indeed signed by the trusted source, it can then be used to evaluate the trustworthiness of the received image. Since the private key used to generate the signature is kept secret, it is difficult for an adversary to forge an authentic hash for a tampered image claimed from a specific source.

In addition to cryptographic encryption, randomization in feature generation and hash construction can help further introduce uncertainty and make it difficult for an adversary to create forgeries and mislead forensic analysis. Some prior effort have been made for image hashing [105, 109], where randomization is introduced into feature extraction and/or quantization steps to improve the security of the hash. At the same time, randomization will reduce the robustness of the hash, and such an inevitable trade-off needs to be carefully studied. Our current work focuses on presenting the basic framework and algorithm of the new FASHION paradigm and demonstrating its performance. How to balance forensic accuracy and attack resilience through randomization are challenging research issues and will be considered in the future work.

2.5 Experimental Results

As discussed in Section 2.3, desirable properties of a forensic hash include compactness, scalability, robustness, and distinctiveness. A compact representation of forensic hash allows more information to be attached to the image. Scalability offers the flexibility to improve the forensic performance by increasing the hash length. Since an image may undergo multiple operations, robust estimation of the target operation is also important. Furthermore, forensic hash should be distinctive so that it is difficult to find two different images that have very similar hash content. In this section, we examine these properties of the proposed forensic hash through experiments.

2.5.1 Experiment Setup

In the experiment, we collect 1000 color images from the Corel database [1] with 10 different categories, such as beach, architecture, flower, etc. The image size is either 256x384 or 384x256. To evaluate the robustness of the forensic hash, we perform 26 operations for each of the 1000 images, generating a database of 27000 images in total. The operations are listed in Table 2.1, including rotation, scaling, cropping, local tampering, blurring, sharpening, and various combinations of these operations. For the local tampering operation, we randomly select and swap two blocks within the image, where the block sizes are 50x50 or 100x100. After swapping, proper blending is introduced to avoid the sharp transition at the boundary of the tampered regions. For each of the 1000 original images, we generate forensic hash composed of both alignment component and integrity check component, and then evaluate the forensic analysis performance over the 26000 modified images. For the alignment component based on visual words representation, a vocabulary tree of 1000 visual words is first trained using SIFT descriptors collected from all color images.

Table 2.1: Image operations and their parameters

Operations	Operation parameters	Variations per image / Total number
Rotation	$3°, 5°, 10°, 30°, 45°$	5 / 5000
Scaling	factor = 0.3, 0.5, 0.8, 1.2, 1.5	5 / 5000
Cropping	19%, 28%, 36% of image size	3 / 3000
Local tampering	block size 50x50, 100x100	4 / 4000
JPEG compression	Q=10	2 / 2000
Filtering	Gaussian filter, Median filter (3 x 3)	2 / 2000
Enhancement	Sharpening, Histogram equalization	2 / 2000
Combinations of rotation, scaling, cropping, and tampering		6 / 6000

2.5.2 Discriminative Capability of Forensic Hash

Before evaluating the forensic capability of the forensic hash, we first study its discriminative performance. We evaluate the capability of the forensic hash in answering a binary question whether a received image is the same original image except having possibly undergone certain operations or is a completely different image. This discrimination is important because performing alignment on two different images is not meaningful. The discrimination task that is carried out here has some resemblance to but is different from the binary authentication task typically considered in the existing image hashing literature. The authentication task answers the question whether a received image has undergone only allowable content preserving operation or it has been maliciously tampered, while in our discrimination task, an image having undergone local tampering is still considered as the same image rather than a different image. Binary authentication using traditional hashing often employs the Hamming distance or L_1 distance between two hashes and compares it to a threshold for differentiation. For forensic hash, simple distance comparison is not applicable. Instead, we exploit the confidence in geometric transform estimation as the metric for differentiation. Two images of the same content but undergo different operations should have a higher confidence score in geometric transform estimation than two different images.

In the forensic hash based on Radon transform, the confidence score from rotation estimation is the normalized cross-correlation between the 1-D summarizations of the two images. The confidence score from scaling estimation is the percentage of extrema points that are matched between the two Radon projections. For the forensic hash based on visual words, the confidence score is the percentage of matched SIFT points between the two images.

$$\begin{aligned}
\text{Confidence of Radon-rotation: } C_{Rr} &= \frac{\mathbf{h}\cdot\mathbf{h}'}{\|\mathbf{h}\|\cdot\|\mathbf{h}'\|}, \\
\text{Confidence of Radon-scaling: } C_{Rs} &= \frac{\#\text{ of matched extrema}}{\text{total }\#\text{ of extrema}}, \\
\text{Confidence of Visual words: } C_{VW} &= \frac{P_{01}}{\min(P_0, P_1)}.
\end{aligned}$$

Here, $\mathbf{h}$ and $\mathbf{h}'$ are the 1-D summarization of Radon transform of the original and received images, respectively, which are described in Section 2.3.1, P_{01} is the number of matched SIFT pairs between the two images, P_0 and P_1 are the number of SIFT points in the two images, respectively. For the discrimination task, $C_{Rr} + C_{Rs}$ is used as the confidence value for Radon based alignment, and C_{VW} is used as the confidence for the visual words-based alignment.

In the discrimination experiment, each image and its modified versions are considered as the same images. Confidence of geometric transform is computed both among same images and between different images. We show the discrimination performance in Fig. 2.14, where all the hashes have roughly the same length, around 700 bits. The ROC curves demonstrate the probability of correct discrimination and the probability of false alarm at different confidence or distance thresholds. We compare the performance with a few other representative image hashing schemes: Roy and Sun [88], Randon soft hash (RASH) [35], and image message authentication codes (IMAC) [120]. Among the three hashing schemes, [88] is most similar to ours in terms of the separation of alignment and tampering detection. The scheme in [35] generates a hash by taking the medium point of Radon projections along a total of 180 directions. The scheme in [120] first computes block average intensity of an image and then extracts the most significant bit from each average value to generate approximate authentication code.

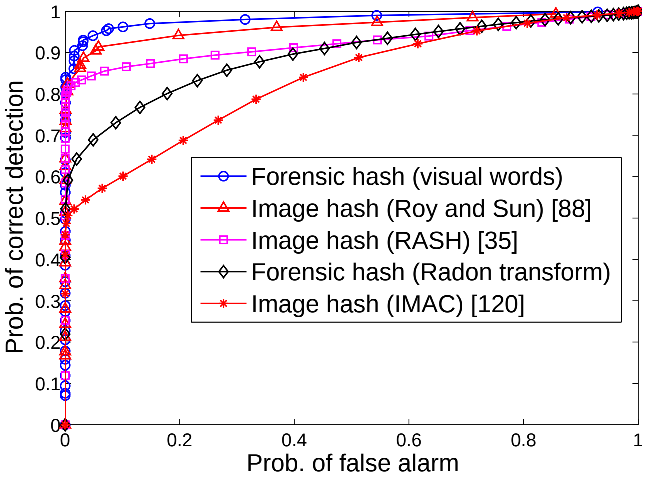


Figure 2.14: Comparison of discriminative performance. Here the discrimination task is to distinguish images that are modified from a source image vs. images that are completely different from the source image.

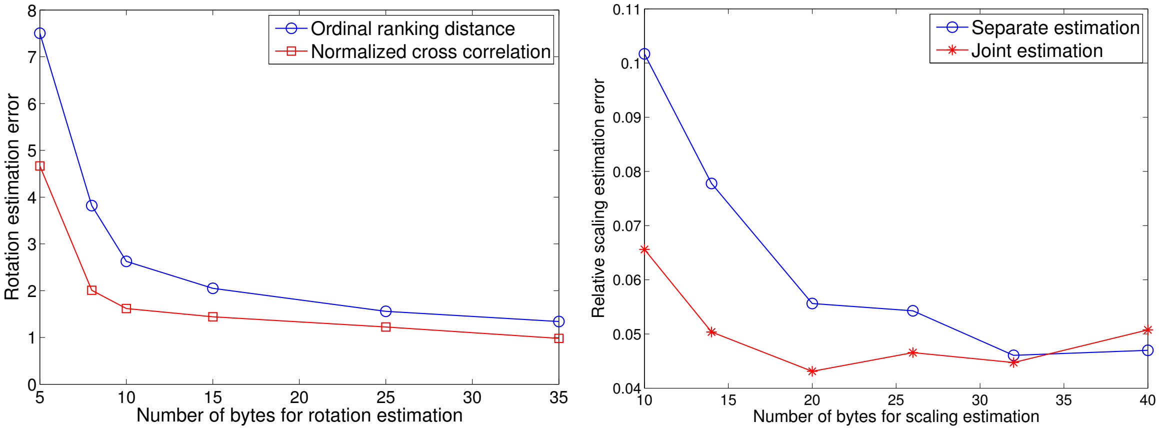
We can see that using the confidence value of geometric transform estimation, the forensic hash based on visual words achieves the best discrimination performance due to the distinctive power of SIFT features. The hash by Roy and Sun [88] has slightly lower performance, because although it also uses the discriminative SIFT features, the number of SIFT points is limited in their hash construction. The Radon soft hash takes the medium values at each of the 180 directions, while our Radon based construction for rotation estimation takes maximum value along only a small subset of the 180 directions. This explains that RASH performs better than our Radon based alignment component for discrimination. But it should be noted that the alignment component can provide geometric transform estimation, which when combined with additional block-based features, can be used to align a received image and significantly improve discrimination performance through blockwise comparison. The schemes in [120] utilizes block-based features and is designed to be robust against operations such as filtering, contrast/brightness adjustment, JPEG compression, but it is not expected to be robust against such operations as geometric transform and cropping that can cause misalignment. The advantage of forensic hash over traditional image hash is clear: at the same compactness, the forensic hash can not only achieve better discrimination between different images, but also provide robust estimation of geometric transform. The alignment component of the forensic hash thus offers a global view of the received image: to determine whether the received image is the same original image, and if so, to align it with the original image. This global view helps overcome the limitation of localized integrity features with respect to misalignment and allow for integrity check and tampering localization be carried out from aligned images in a meaningful way, which demonstrates that different components in the forensic hash complement each other and work together in a synergistic way.

2.5.3 Geometric Transform Estimation

In this subsection, we evaluate the performance of geometric transform estimation using the proposed alignment components. More specifically, we examine the estimation accuracy of the scaling factor and rotation angle with respect to the length of forensic hash, i.e., the amount of available side information.

Alignment performance based on Radon transform and scale space theory

In this construction of the alignment component, rotation and scaling are estimated using different side information. The 1-D summarization of Radon transform of the image is downsampled to assist rotation estimation, while the stable extrema in the Radon projection along horizontal and vertical directions are used for scaling estimation. By increasing the hash length, more sample points and more stable extrema can be included to improve the estimation performance. The rotation and scaling estimation accuracy are shown in Fig. 2.15a and Fig. 2.15b, respectively.



(a) Rotation estimation accuracy

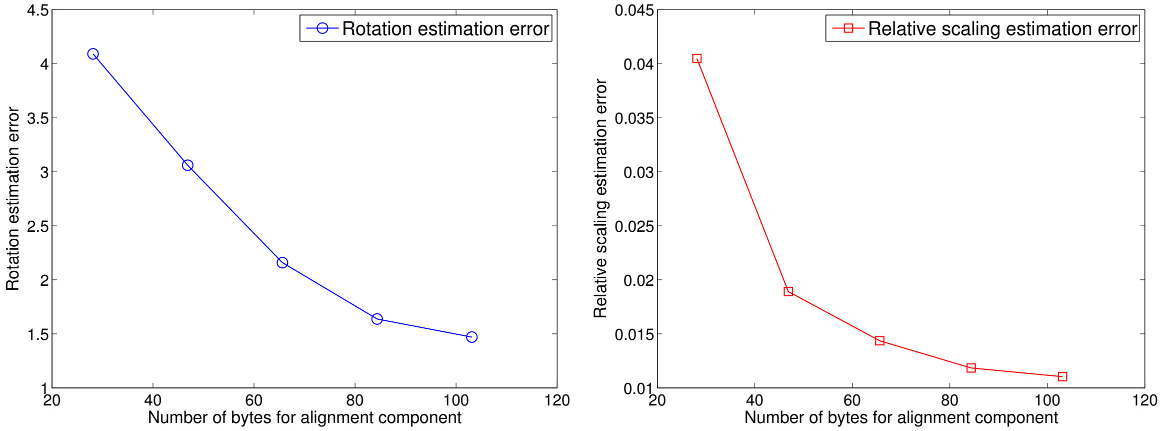
(b) Scaling estimation accuracy

Figure 2.15: Performance of geometric transform estimation using Radon transform and scale space theory

In the rotation estimation experiment, we use two distance metrics when estimating the rotation angle, namely, the normalized cross-correlation and L_1 distance of the ordinal ranking of the hash. As shown in Fig. 2.15a, the normalized cross-correlation performs consistently better than the ordinal ranking, especially at short hash lengths. At longer hash lengths, the estimation has similar performance for the two metrics. The rotation estimation accuracy can reach below 2 degrees with a hash length of 10 to 15 bytes. The estimation is robust for images that have undergone multiple operations, such as combinations of rotation, scaling, cropping, and local tampering.

For scale estimation, we use relative estimation error as the performance metric. When considering the horizontal and vertical directions separately, the estimation accuracy is shown by the curve with circle markers in Fig. 2.15b. Since scaling operation that maintains aspect ratio is done isotropically, we can estimate the scaling factors along the horizontal and vertical directions, respectively, and then select the one with higher confidence as the final estimation. This joint estimation result is shown by the curve with star markers in Fig. 2.15b, where less than 7% relative error can be achieved using only 5 extrema along each of the two directions, so overall it will only add 10 bytes to the alignment component. It should be noted that majority of the errors are contributed from images whose estimated rotation angles are different from the actual values by more than 5 degrees. For our database of 26000 modified images, only around 1.5% images have rotation error larger than 5 degrees, and this number can be further reduced by using longer hash lengths. For images whose rotation estimation is accurate, the scaling estimation error is typically below 1%.

Alignment performance based on visual words representation For the alignment component built on visual words representation, the most stable SIFT features from the original image are matched to the most stable SIFT features in the modified image to determine the scaling factor and rotation angle between the two images. In the experiment, we vary the number of SIFT points included in the alignment component and obtain the average estimation performance shown in Fig. 2.16.



(a) Rotation estimation accuracy

(b) Scaling estimation accuracy

Figure 2.16: Performance of geometric transform estimation using visual words of SIFT features

It should be noted that the number of bytes in Fig. 2.16 is for the entire alignment component, while that in Fig. 2.15a and Fig. 2.15b is just for rotation and scaling estimation, respectively. As can be seen from the figures, about 50 bytes of the alignment component based on visual words achieve rotation estimation error around 3 degrees and relative scaling estimation less than 2%. At the same hash length, the alignment component based on Radon transform has a similar rotation estimation error of 3 degrees but higher relative scaling estimation error of around 5%. We can see that the two constructions of alignment component give comparable estimation performance, and the alignment component based on visual words has better performance on scaling estimation. A possible source of this can be attributed to the fact that the characteristic scale associated with local SIFT points are more reliable than extrema points from 1-D Radon projections for the task of scaling estimation. By increasing the hash length, the estimation performance can be further improved. Our experiments have also found that the alignment component based on visual words is robust against various image operations such as cropping, local tampering, and combinations of such operations with rotation and scaling.

We carry out more comparison between the two proposed alignment components in terms of their advantages and limitations. For the alignment component based on Radon transform and scale space theory, the scaling estimation uses Radon projection of the image along horizontal and vertical directions. This requires the rotation angle of the image be accurately estimated, and thus the accuracy of rotation estimation will affect the performance of scaling estimation. For the alignment component based on visual words, estimating scaling and rotation are independent of each other. We compare the two proposed alignment components and list their scaling estimation performance under rotation estimation error of 3, 5, and 10 degrees in Table 2.2. We can see that scaling estimation performance degrades for the alignment based on Radon transform as the rotation estimation error increases, while alignment based on visual words maintains very good estimation accuracy. It should be noted that for over 90% of the modified images in our database, the

Table 2.2: Relative scaling estimation error using different alignment algorithms

Rotation error	3^o	5^o	10^o	36% cropping
Radon transform based alignment	7.17%	11.69%	18.58%	6.27%
Visual words based alignment	1.13%	1.54%	1.37%	0.26%
Robust hash by Roy and Sun [88]	2.62%	13.2%	10.21%	1.81%

rotation estimation error is less than 1^o and thus the scaling estimation using Radon transform based alignment is accurate for most cases.

Another difference between the two alignment constructions is that the Radon transform based alignment uses global Radon projection features, while the visual words based alignment uses local SIFT features. Radon transform based alignment is less robust to large amount of local tampering and cropping, because a large amount of tampering may change the Radon projection of the image along most directions significantly, while for visual words based alignment, SIFT features in the untampered regions can still provide useful information for forensic analysis. In Table 2.2, we can see that when 36% of the image is cropped, the alignment component based on Radon transform performs worse than the alignment component based on visual words.

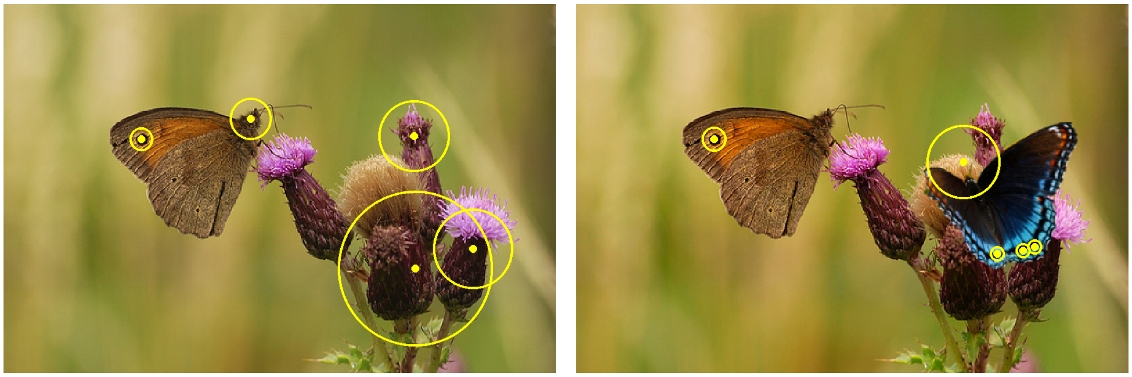
The performance of visual words based alignment depends on reliable matching of SIFT features between the testing and original images. At short hash length where only a few SIFT points can be included, the matching may not be reliable for images that have undergone tampering operations which may alter the original SIFT features. In such cases, the Radon-based alignment component may be preferred to obtain a rough estimation of geometric transform.

Overall, the two alignment components both provide accurate and robust geometric transform estimation. The visual words based construction gives better scaling estimation and more robust to large amount cropping and local tampering, while the Radon transform based construction can provide robust estimation at shorter hash lengths.

Comparisons with representative prior art We compare the performance of geometric transform estimation between our proposed alignment components and the robust image hashing work by Roy and Sun [88], which is most relevant to our work in terms of localization capability and global robustness through image registration. As reviewed in the introduction section, the image hash in [88] contains registration component to estimate the geometric transform before doing image authentication. Five most stable SIFT features are selected and their positions are included in the hash. To improve the compactness of the hash, the 128-dimensional SIFT descriptor is projected onto a Gaussian random matrix to generate a 60-bit binary string, which is similar in spirit to locality sensitive hashing. These 60-bit binary strings are included in the hash and used for matching purposes. Each SIFT point takes about 10 bytes and the registration component is around 50 bytes.

When implementing the Roy-Sun hash, we select most stable SIFT features based on their contrast values, because higher contrast values indicate better stability. When tested on the Corel image database, the Roy-Sun hash achieved an average of 10 degrees error on rotation estimation and 11% relative error for scaling estimation using the 50-byte registration component. At the same hash length, our proposed alignment components can achieve around 3 degree rotation error and 1% scaling error. Our proposed schemes also provide better scalability, as the estimation performance can be further increased when increasing the hash length. For Roy-Sun hash, each additional SIFT point will increase the hash length by 10 bytes, thus it requires roughly twice the length of our proposed schemes in order to achieve the same estimation accuracy.

The main limitation of the registration component in [88] is the limited number of SIFT points that are included. The most stable SIFT points can change after the image has undergone some operations, such as cropping, local tampering, and rotation. Some SIFT points may be removed from the image due to cropping or local tampering, and new points can be introduced into the image after local content change. Fig. 2.17 shows an example, in which we can see that the SIFT points with the highest contrast values can be changed or removed after small content modification. As a result, there can be fewer than 5 matching pairs between the original and modified images, and the reduced number of matched pairs leads to less robust estimation results or makes it difficult to provide any reasonable estimates. In contrast, the proposed alignment components are designed to be robust against common image operations. In the alignment component based on visual words representation, more SIFT points are included and help achieve more robust estimation performance.


(a) Original image

(b) Modified image

Figure 2.17: Five most stable SIFT points in the original image and the modified image.

2.5.4 Tampering Detection and Localization

In the modular design of the FASHION framework, the alignment component provides a global view of a received image to determine whether it is modified from the same original image and perform the necessary alignment. This global view facilitates other components such as the integrity check component in providing a finer resolution forensic analysis including tampering detection and localization.

Integrity check performance with block-based features As described in Section 2.4, the integrity check component can be constructed using block-based edge direction histograms from the original image. An example of tampering localization is shown in Fig. 2.19. For tampering localization, the choice of block size controls the trade-off between hash length and detection resolution. A larger block size gives a smaller hash length but can introduce higher false detection than a smaller block size. The edge direction histogram in each block is quantized in order to achieve a compact representation. We examine the tampering localization performance over our image database and compare the performance using uniform quantization and Lloyd quantization for the edge-direction histogram in Fig. 2.18. The block size is chosen to be 16 by 16. ROC curves are obtained by quantizing each component of the edge direction histogram to 4 bits using a uniform quantization and a Lloyd quantization, respectively. It can be observed from Fig. 2.18 that at the same hash length, the Lloyd quantization significantly improves tampering localization performance compared to using uniform quantization. This suggests that using Lloyd quantization leads to a more compact hash because we can show that using 1 byte per block with Lloyd quantization can achieve similar performance to that of using 2 bytes per block with uniform quantization. Furthermore, a 90% correct detection of tampered pixels with less than 3% false detection rate can be achieved, indicating accurate tampering localization. Not all tampered pixels may be detected because the tampered regions may not align with the image blocks, and a small number of tampered pixels in one block may not cause significant changes in the edge direction histogram of that block.

Integrity check performance with hybrid construction One limitation of the integrity check component with block-based features is that in order to achieve high localization accuracy and accommodate large image size, the number of blocks and thus the hash length will increase significantly. To ensure the compactness of the forensic hash, we have proposed a hybrid construction for the integrity check component in Section 2.4.3. The idea of the hybrid construction is to utilize informa-

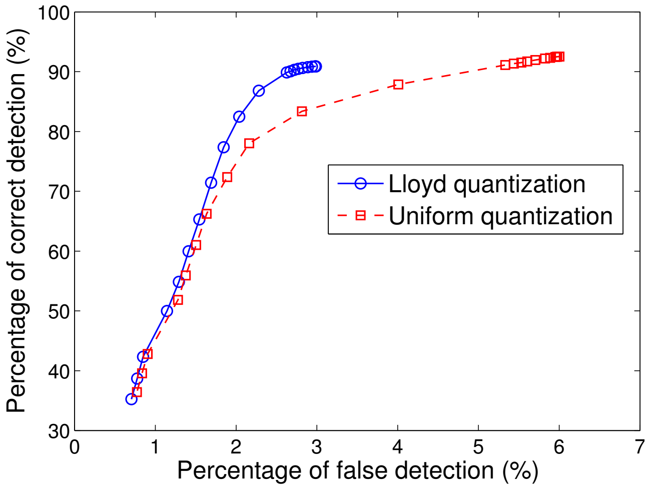
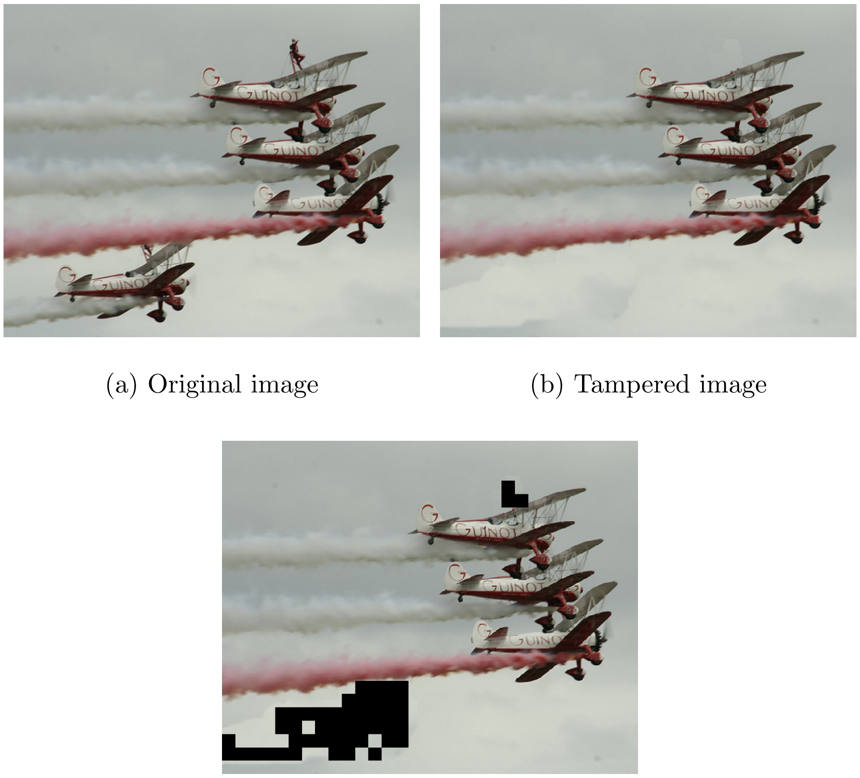


Figure 2.18: Tampering localization performance

tion from the alignment component to assist tampering localization, thus reducing the amount of information needed from the integrity check component. In our hybrid construction, the addition and deletion of SIFT points between the original and testing images is also used for tampering detection. Block-based features are used to detect tampering only for regions where no highly stable SIFT points are detected at the given contrast value, which substantially reduces the length of the forensic hash. For the Corel database, if block-based features are extracted from 32x32 blocks and a total of 20 SIFT points are encoded, we can reduce the number of block features by 30% on average. If the edge histogram is quantized to 1 byte per block using non-uniform quantization, we can save around 80 bytes on the integrity check component. The overall length of the integrity check component depends on the image size, block size, and quantization levels. An example of the tampering detection and localization using the hybrid construction is shown in Fig. 2.20. The tampered regions detected by block-based features are marked by solid blocks and



(a) Original image

(b) Tampered image

(c) Tampering localization

Figure 2.19: Example of tampering localization using block-based feature. (Original image source: Flickr)

regions detected by SIFT points are marked by dashed blocks. The SIFT points that do not match between the original and modified images are denoted by solid circles.

For tampering detection and localization using SIFT features, we have higher confidence that a region may be tampered if a larger number of SIFT points are not matched between the original and the modified images in that region. However, we have observed from experiments that a small amount of addition and deletion to SIFT points can occur due to content preserving operations, such as geometric trans-

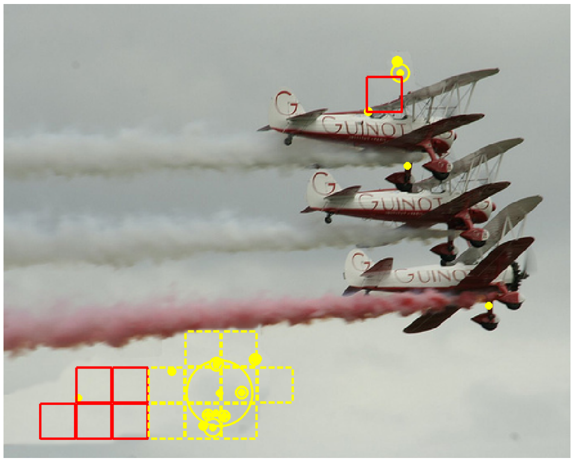


Figure 2.20: Tampering localization using the hybrid integrity check component (This figure is better viewed in color version of the dissertation)

form, filtering, compression, and others. These image operations may affect SIFT descriptors and can cause the descriptors to be assigned to different visual words during clustering. This may reduce the number of matched SIFT points between two images, which will in turn lower the confidence of tampering detection for regions with fewer matched SIFT points. We expect better clustering schemes such as soft clustering and more robust SIFT matching schemes can improve the robustness of visual words representation and provide more reliable tampering detection.

2.6 Chapter Summary

In this chapter, we proposed a new concept and framework of forensic hash as a compact representation of side information from the original image to facilitate forensic analysis. Two novel constructions of the hash are proposed, utilizing Radon transform, scale space theory and visual words representation of SIFT features. At the same compactness as the conventional image hash, the forensic hash can go beyond binary authentication and reliably detect geometric transform and estimate the transform parameters; when compared to the blind multimedia forensics, the forensic hash can answer a broader scope of forensic questions more accurately and efficiently. Furthermore, the performance of the forensic hash can be improved by moderately increasing the hash length, and the modular design of forensic hash provides flexibility and extensibility to suit for different applications. The geometric transform offered by the forensic hash serves as an important building block for further locating tampering using block-based features. We have also demonstrated very good image discrimination capability of the forensic hash. In future work, we plan to design more forensic modules to enrich the forensic capability of the forensic hash, and investigate means to further complement cryptographic approach in securing hash and mitigating attacks.

CHAPTER 3

Forensic Hash: Applications and Extensions

In this chapter, we demonstrate several extensions and applications of the FASHION framework proposed in Chapter 2. A good design of forensic hash should have good extensibility in answering a broader scope of forensic questions by introducing new modules or using existing modules. In the first part of the chapter, we show that the forensic hash that we proposed in the previous chapter is very effective at estimating advanced image editing operations such as seam carving, without changing the hash design. In the second part, we extend the FASHION spirit to the task of reduced-reference quality assessment on retargeted images, where compact side information is used to provide structure distortion analysis caused by retargeting operations.

3.1 Seam Carving Estimation using Forensic Hash

3.1.1 Background and Motivation

In recent years, mobile devices with multimedia capturing capability and social networks that provide media sharing and streaming services are rapidly emerging. Given the various screen sizes of the mobile devices, properly resizing multimedia data to better render them on smaller screen sizes becomes important. One group of techniques called retargeting [89], or content-adaptive resizing, address such problem by resizing an image or video without scaling down or distorting the salient content inside too much. Seam carving is one representative image retargeting technique, proposed by Avidan et al. [4]. It resizes an image in a way adaptive to its content by removing seams, which are eight-connected paths of low energy, from the image while keeping salient objects intact. Details on the seam carving algorithm can be found in [4]. Due to its capability of preserving salient objects and its aspect ratio after resizing, seam carving has found promising applications in rendering images and videos on smaller displays, such as mobile phones. In this section, we develop forensic techniques [69] to detect seam carving and estimate its parameters for the purpose of trustworthiness evaluation.

The main objective of seam carving is better image resizing, but as demonstrated in [4], seam carving can also be used to intentionally remove objects from the image. Such tampering brings challenges to forensic tasks. Traditional blind multimedia forensics try to detect potential tampering of digital images/videos without proactive aids such as signature attachment or embedded watermark [24]. This is accomplished by analyzing intrinsic traces, such as inconsistencies in signal characteristics, left by the processing operations [96]. However, the adaptive nature of seam carving makes it difficult to identify any traces or inconsistencies unique to the operation. Recent work by Sarkar et al. [94] and Fillion et al. [33] detect whether an image has undergone seam carving or not by using a machine learning framework based on intuitive features extracted from the image. Th accuracy of seam carving detection in the above mentioned works is around 80-90% for large amount of seam carving (e.g., number of seams larger than 30% of original image size), and becomes lower for a smaller amount of seam carving. In contrast to answering only a binary question of whether the image is seam carved or not using blind forensic techniques, in this section, we explore using compact side information to not only detect seam carving but also estimate the amount and location of seam carving. Such detailed information can help us better evaluate the trustworthiness of the image.

Utilizing compact side information for enhanced forensic analysis is the main spirit of the FASHION framework that we proposed in the previous chapter. A good design of forensic hash should good extensibility such that it can answer a broader scope of forensic questions by adding new forensic modules or reusing existing modules. In this section, we use the forensic hash based on visual words representation of SIFT features [67], which is proposed in Section 2.3.2 and extend it for seam carving detection and estimation. Experiments show that forensic hash can accurately estimate the amount of seam carving and their approximate locations. With the estimation results, we further explored reconstruction of original image from the seam carved image for tampering detection. Such a forensic analysis can provide a detailed trustworthiness evaluation of the image in terms of which part can be trusted and which part might be tampered.

3.1.2 Seam Carving Estimation

Below, we formally describe the seam carving estimation algorithm using forensic hash. We denote the original image as $\mathbf{I}$ and its forensic hash $\mathbf{h} = \{\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_k\}$, where $\mathbf{v}_i$ is the parameter vector of the i^{th} stable SIFT point in $\mathbf{I}$. Image $\mathbf{I}$ is transmitted and undergoes seam carving and potentially additional geometric transforms and tampering operations such as cut-and-paste. We denote the received image as $\tilde{\mathbf{I}}$ and its top stable SIFT points as $\tilde{\mathbf{h}} = \{\tilde{\mathbf{v}}_1, \tilde{\mathbf{v}}_2, \cdots, \tilde{\mathbf{v}}_k\}$. Given $\mathbf{h}$ and $\tilde{\mathbf{h}}$, we match their SIFT points based on their visual word IDs and denote the matched points as $(p_1, \tilde{p}_1)$, $(p_2, \tilde{p}_2)$,$\cdots$,$(p_n, \tilde{p}_n)$. Before seam carving estimation, we first need to make sure $\tilde{\mathbf{I}}$ is on the same scale and orientation as $\mathbf{I}$. This is achieved by estimating the rotation angle θ and scaling factor δ using the matched points, as described in [67]. The image $\tilde{\mathbf{I}}$ is then transformed to be $S(R(\tilde{\mathbf{I}}, -\theta), 1/\delta)$, thus aligned to $\mathbf{I}$. Here $R(\cdot,\cdot)$ is the rotation operator and $S(\cdot,\cdot)$ is the scaling operator. For simplicity, below we still denote the transformed image as $\tilde{\mathbf{I}}$.

To locate the vertical seams removed from the image, we sort the matched points based on their x-coordinates and then compare the distances between every adjacent matched pairs. An illustrative example is given in Fig. 3.1, where x_i, x_{i+1}, x_{i+2} are the x-coordinates of three adjacent points from original image $\mathbf{I}$ and $\tilde{x}_i$, $\tilde{x}_{i+1}$, $\tilde{x}_{i+2}$ are the x-coordinates of corresponding points in the resized image $\tilde{\mathbf{I}}$. We denote the distance between two adjacent points in $\mathbf{I}$ as $d_i = x_{i+1} - x_i$ and the distance between corresponding points in $\tilde{\mathbf{I}}$ as $\tilde{d}_i = \tilde{x}_{i+1} - \tilde{x}_i$, then the number of vertical seams removed in the horizontal range $[\tilde{x}_i, \tilde{x}_{i+1}]$ in $\tilde{\mathbf{I}}$ can be computed as $\Delta C_i = d_i - \tilde{d}_i$. A positive ΔC_i indicates seam removal and a negative one indicates seam insertion. For the example given in Fig. 3.1, we can see that there are seam insertions in the range $[\tilde{x}_i, \tilde{x}_{i+1}]$ and seam removal in the range $[\tilde{x}_{i+1}, \tilde{x}_{i+2}]$. After considering all adjacent matched pairs, we obtain the estimation results $\{\Delta C_i, \tilde{x}_i, \tilde{x}_{i+1}\}$ where $i \in \{1, \cdots, n\}$. Furthermore, in order to estimate the number of seams removed before the first point $\tilde{p}_1$ and that after the last point $\tilde{p}_n$, we can include the size of the original image into the forensic hash and obtain the complete estimation results as $\{\Delta C_i, \tilde{x}_i, \tilde{x}_{i+1}\}$ where $i \in \{0, \cdots, n+1\}$, $\Delta C_0 = x_1 - \tilde{x}_1$, $\Delta C_{n+1} = (w - x_n) - (\tilde{w} - \tilde{x}_n)$. w and $\tilde{w}$ are the width of images $\mathbf{I}$ and $\tilde{\mathbf{I}}$, respectively. To locate horizontal seams, we sort the matched points along y-coordinates and follow the same procedure.

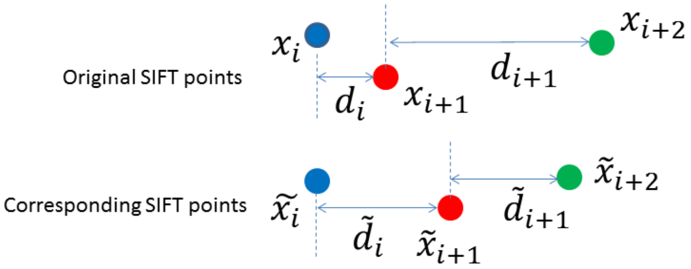
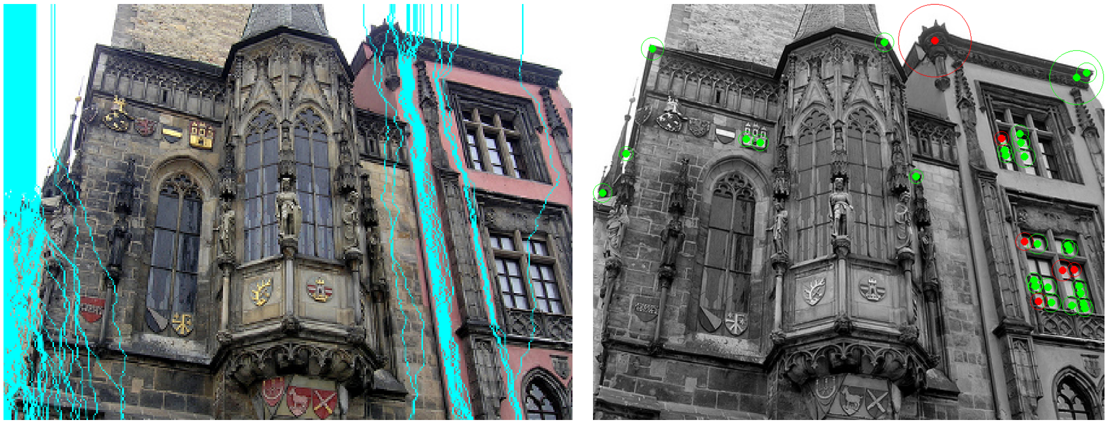


Figure 3.1: Vertical seam estimation using x-coordinates of matched SIFT point pairs

To give an example, we show the original image and its 50 vertical seams to be removed in Fig. 3.2a. The resized image and its stable SIFT points with contrast value larger than 0.05 are shown in Fig. 3.2b. The green circles are SIFT points matched with the original SIFT points encoded in the forensic hash and red ones are those not matched due to seam removal. By comparing the original and new distances of every adjacent pairs of matched points, we estimated that there are 30 vertical seams removed in the horizontal range of [1,9] in the resized image, 2 seams in the range of [9,30], 17 seams in the range of [285,363], and 1 seam in the range of [404,418]. Compared with ground truth knowledge, we have correctly estimated all 50 seams with no false alarm.



(a) Original image and its 50 vertical seams (b) Image after seam carving and its stable SIFT points

Figure 3.2: Illustration of seam carving estimation (This figure is better viewed in color)

It should be noted that such capability of estimating the amount and location of removed seams does not require modifying the hash construction in [67]. This shows that a good design of forensic hash can be used to answer a broad scope of forensic questions. The estimation here provides the regions in the received image where seam carving has occurred. This information is very helpful to reconstruct the original image and enable further forensic analysis such as tampering localization, as will be demonstrated in the next section.

3.1.3 Image Reconstruction and Tampering Localization

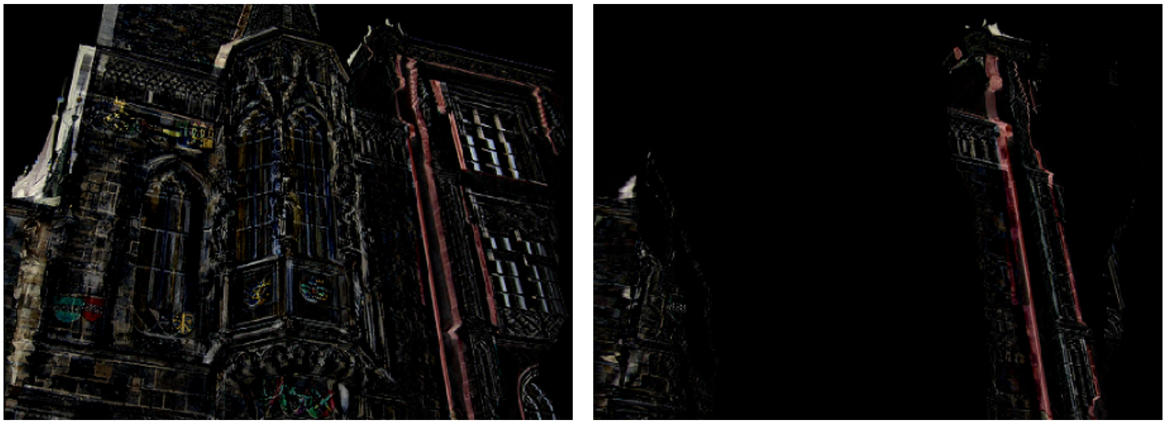
Knowing where and how many seams have been removed is an important first step to evaluate image trustworthiness. Places with seam removal are less trustworthy than regions without seam carving. In this section, we explore how to further use such information to adaptively resize a received image to align with the original image and thus enable tampering localization through block-wise feature comparison.

There are several ways to resize the seam carved image. Without any knowledge about the seam carving amount and location, a naïve resizing option is to resize the image or insert seams in the image. Such strategies cannot align the two images accurately and may make block-based comparison unreliable. In contrast, knowledge of the seam carving estimation can guide the reconstruction process by constraining the resizing or seam insertion to only those regions that have undergone seam carving and with only the necessary amount. More specifically, with estimation result $\{\Delta C_i, \tilde{x}_i, \tilde{x}_{i+1}\}$, i.e., ΔC_i seams have been removed in the horizontal range $[\tilde{x}_i, \tilde{x}_{i+1}]$ of image $\tilde{\mathbf{I}}$, we increase the width of $\tilde{\mathbf{I}}$ by ΔC_i through either rescaling the vertical strip $[\tilde{x}_i, \tilde{x}_{i+1}]$ or inserting ΔC_i seams into the same range. Since the shape of a seam is irregular and its pixels may not all fall into a vertical strip, we insert seams that have at least half of its pixels in the specified range.

Similarly for resized image due to seam insertion, we can reconstruct the original image by removing seams from the resized image.

To illustrate the accuracy of such reconstruction and alignment, we show an example below. For the same image in Fig. 3.2b, we apply simple seam insertion without any constraints and seam insertion with constraints based on the estimation results to resize the image to its original size. The difference between the resized results and the original image are shown in Fig. 3.3a and Fig. 3.3b, respectively, where the difference is shown as a color image and black color indicates zero error. We can see that the constrained seam insertion can provide accurate alignment between the resized image and original image, while simple insertion without any constraints mis-aligned the two images, causing large errors in many places of the image. The PSNR is 14.8 dB for simple insertion and 22.8 dB for constrained insertion.

Since the original image is not available during forensic analysis, compact block-based features can be encoded into the forensic hash as an integrity check component for tampering localization. The block feature used here is edge pixel direction histogram, quantized to four directions and has size of 1 byte per block, as described in [67]. Using block size of 32 by 32, the average block feature distance is 65.05 for simple insertion and 15.02 for constrained insertion. The accurate alignment achieved by adaptive reconstruction that utilizes seam carving estimation result is very important, as it enables tampering localization using block-based features, which are not robust to misalignment. As can be seen from Fig. 3.3a, misalignment may cause the block-based comparison to consider untampered regions as tampered.



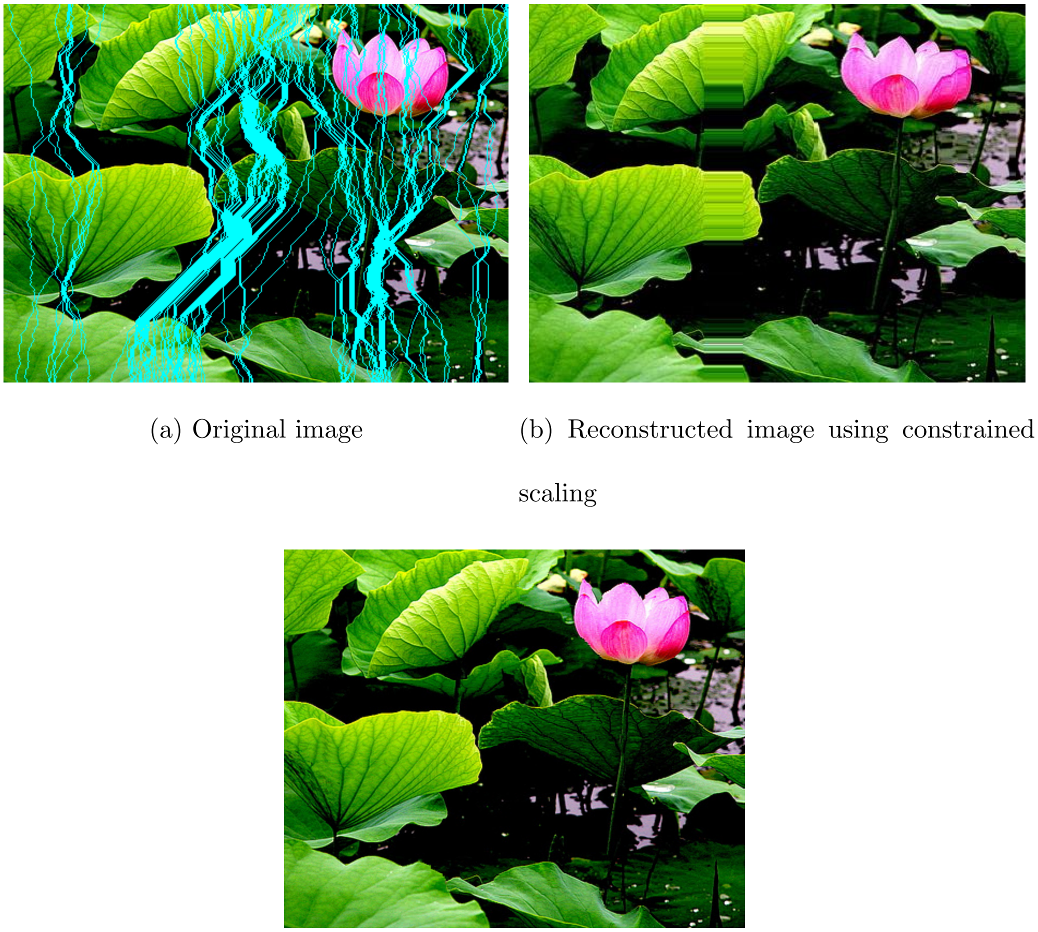
(a) Naïve reconstruction error

(b) Constrained reconstruction error

Figure 3.3: Difference between original and reconstructed image using seam insertion without and with constraints

With the knowledge of seam carving amount and locations, we resize each given range using simple scaling or seam insertion. For example, if seam carving estimation reveals that a vertical strip with horizontal range [10,29] has been carved 20 seams, we can resize this strip to a new width of 40 through scaling or inserting 20 new seams that pass through this vertical strip.

For large resizing amount, constrained scaling and constrained seam insertion both produce a blurred result and seam insertion may introduce additional distortion along edges. For small resizing amount, seam insertion is a better option than scaling in the sense that it can avoid blurness. Another case that seam insertion produces better reconstruction than scaling is when the removed seams have irregular shapes or going diagonal directions. In this case, scaling a vertical or horizontal strip will distort the image content. An example is given in Fig. 6. The original



(a) Original image

(b) Reconstructed image using constrained scaling

(c) Reconstructed image using constrained seam insertion

Figure 3.4: Image reconstruction example

image and its removed seams are shown in Fig. 3.4a. The reconstructed images using constrained scaling and constrained seam insertion are shown in Fig. 3.4b and Fig. 3.4c, respectively. We can see that the removed seams in the lower center region of the original image moves in a diagonal direction, therefore, scaling will produce a blurred strip in the center of the resized image (Fig. 3.4b) while the seam insertion can avoid such distortion in the image content (Fig. 3.4c).

Reconstructing the exact original image from a seam carved image is a challenging and open problem, but for the purpose of aligning the resized image with the original image for tampering detection, both the constrained scaling and constrained seam insertion work well and we will show more experiments in the next section.

3.1.4 Experimental Results

In this section, we perform several experiments to evaluate the performance of seam carving estimation and tampering detection using forensic hash. The image dataset used in the section includes 200 images from Flickr with 40 different tags, such as beach, building, flower, etc. The image size is about 500x300. For each of the 200 images, we perform seam carving along its larger dimension and generate four resized images whose modified dimension has 60%, 70%, 80%, and 90% of the original size, respectively. The total number of resized images is 800.

Robust geometric transform estimation To estimate seam carving, the modified image should be first aligned with the original image to the same orientation and scale. Such alignment can be achieved through geometric transform estimation using forensic hash. Here, we validate the robustness of such alignment against seam carving operation. For an image undergone seam carving operation, a robust geometric transform estimation should report a scaling factor close to 1. The estimation results on the 800 resized images are shown in Table 3.1, which shows the absolute rotation angle estimation error and relative scaling factor estimation error.

We can see that estimation errors remain low even at large amount of seam carving.

Table 3.1: Robustness of forensic hash against seam carving

Resize factor	90%	80%	70%	60%
Rotation error	0.06	0.26	0.97	2.82
Scaling error	0.18%	0.76%	1.88%	3.37%

Seam carving estimation After geometric alignment, we can estimate the amount and position of removed seams as described in Section 3.1.2. By comparing with the ground truth seam carving amount, we evaluate the estimation accuracy using probability of correct detection (P_d) and probability of false detection (P_f), which are defined as follows:

$$P_d = \sum_i \frac{\min(\Delta\hat{C}_i, \Delta C_i)}{\Delta C}, \; P_f = \sum_i \frac{\max(\Delta\hat{C}_i - \Delta C_i, 0)}{\Delta C}.$$

$\Delta\hat{C}_i$ is the estimated seam carving amount in the range given by the ith matched SIFT pairs, ΔC_i is the actual carving amount in the same range, and ΔC is the ground truth value of total number of seams that have been removed.

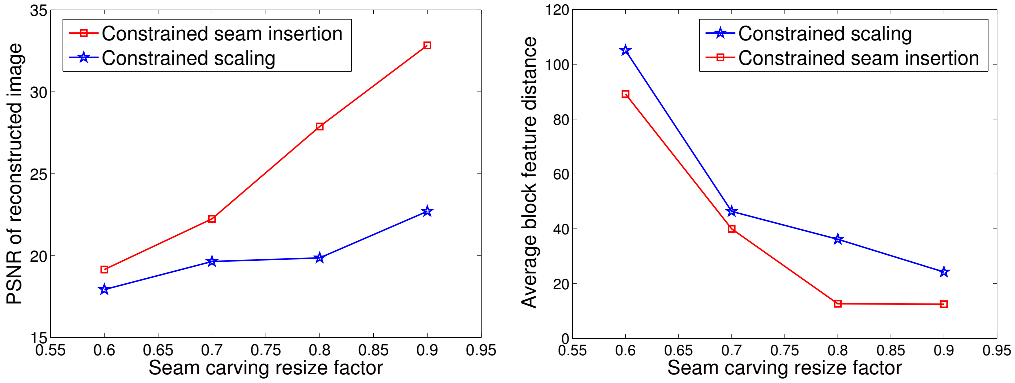
We perform estimation on the 800 images with different resize factors. The probability of correct detection and false detection at different hash lengths are shown in Table 3.2. With hash length at around 50 bytes, the average probability of correct detection is 99.4% and average probability of false detection is 2%. We can also see that longer hash length may not improve the estimation accuracy. Actually, when more SIFT points are used, there is higher chance of mismatch of SIFT points and we see slight decrease in estimation accuracy.

Table 3.2: Seam carving estimation performance

Hash length (bytes)	47	94	156	219
Prob. of correct detection	99.4%	98.6%	98.6%	98.3%
Prob. of false detection	1.99%	3.52%	4.26%	3.48%

Reconstruction and alignment We perform reconstruction using both the constrained scaling and constrained seam insertion guided by the seam carving estimation results over the 800 images. The PSNR and block feature distance between the original image and the reconstructed image using a forensic hash of 47 bytes are shown in Fig. 3.5. Fig. 3.5a shows the PSNR of the reconstructed image at different seam carving resize factors. We can see that constrained seam insertion consistently outperforms the constrained scaling, and the reconstruction quality is better if the seam carving amount is smaller, i.e., the resize factor is large. Fig. 3.5b, compares the average block feature distances of the two methods, and again we can see constrained seam insertion has better performance.

Tampering localization By adaptively resizing the seam carved image, we can accurately align it with the original image for tampering localization. We illustrate one example below. The original image is shown in Fig. 3.6a and its tampered version is shown in Fig. 3.6b. The tampered image has undergone seam carving to remove the central building and then cut-and-paste to insert a plane. Our seam carving estimation correctly identifies that there are 125 missing seams in the center of the tampered image. After adaptive resizing using constrained scaling, the result

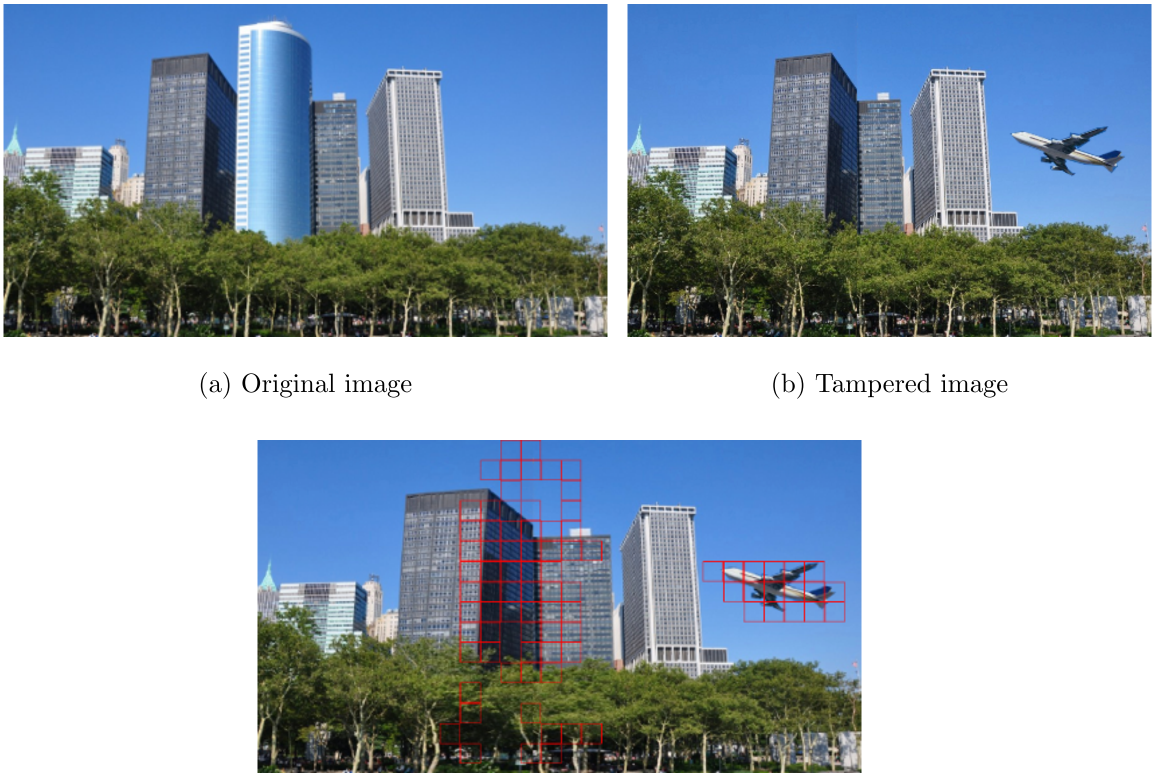


(a) PSNR

(b) Average block feature distance

Figure 3.5: Reconstruction and alignment performance

of block-wise comparison of edge-direction histogram is shown in Fig. 3.6c. Both the center region where the building has been removed and the inserted plane are correctly identified as tampered, as covered by red blocks. Therefore, using forensic hash and additional block-based features, we can provide a detailed report on the trustworthiness of an image, in terms of which part can be trusted and which part might be tampered.



(a) Original image

(b) Tampered image

(c) Tampering detection result

Figure 3.6: Example of reconstruction and tampering localization

3.2 Reduced-Reference Quality Assessment for Retargeted Images

The study of estimating seam carving using compact SIFT representations motivates us to further extend the FASHION spirit to the more general image retargeting operations, and in this section, we study the problem of using compact side information to evaluate quality of images that have undergone retargeting. Identifying and quantifying image quality degradation is very important in order to maintain and control image quality in various applications and online services. Image quality assessment research aims at developing techniques to predict image quality accurately and automatically. Below, we first review the related work on image quality assessment and image retargeting, then we discuss the main idea and contributions of this work.

3.2.1 Related Work and Our Contributions

The most reliable way of assessing image quality is by subjective evaluation involving human observers. The mean opinion score (MOS) is one of the commonly used and well regarded subjective measure for image quality assessment. However, involving human observers can be expensive and too slow to be useful for practical applications. Therefore, the goal of objective image quality assessment is to design computational models such that an estimated quality by the models correlates well with human subjectivity. Depending on the amount of available information of the original reference image, image quality assessment can be classified into full reference (the reference image is fully available), reduced reference (only partial information about the reference image is available), and no reference (no access to the reference image is allowed).

The simplest and most widely used full reference quality metric is the mean square error (MSE) and related metric of peak signal-to-noise ratio (PSNR). Although simple, MSE and PSNR are not well matched to the perceived visual quality of human beings. Taking advantage of known characteristics of human visual system (HVS), most state-of-the-art quality assessment works have adopted a two-stage structure, namely, local distortion measure and spatial pooling to get a final quality score. Some representative local distortion measures include structure similarity (SSIM) index [111, 113], block discrete cosine transform [116] and wavelet-based approaches [107, 117]. These local quality measures are then pooled together to maximize the correlation between objective and subjective image quality ratings [27, 55, 56, 114]. A comprehensive survey of image quality assessment techniques can be found in [112].

Image retargeting is one type of techniques that provide content-adaptive image resizing for better viewing images on screens of different sizes. The seam carving algorithm that we studied in the previous section in this chapter is one particular type of retargeting techniques. Image retargeting methods can be roughly classified as discrete or continuous [97]. Discrete approaches remove or insert unimportant pixels or patches from the interior of the images [4, 90, 91]. In such approaches, seams with minimum energy are removed from the image for resizing. The results are promising, although at large scale changes, visible discontinuities or aliasing artifacts can often be seen. Continuous retargeting techniques optimize a mapping (warp) from the source media size to the target size without explicit content removal. The key idea of such approaches is to scale visually important feature regions uniformly while allowing arbitrary deformations in unimportant regions of the image. Some representative works include [54, 110, 118]. A comprehensive comparison of state-of-the-art image retargeting techniques can be found in a recent work by Rubinstein et al. [89].

In this work, we study the problem of reduced-reference quality assessment for retargeted images. Exploring reduced-reference quality assessment is partly motivated by the work of forensic hash that we proposed earlier in this thesis. The idea of attaching a short signature about the original image before image transmission to assist forensic analysis can also be carried out to the task of quality assessment. Given that the original reference image is seldom fully available in many real-world applications, reduced-reference quality assessment will be a more feasible alternative, where the compact partial information can be embedded into the header file of the image and be used to monitor and assess image quality.

The motivation of applying quality assessment on retargeted images is two-fold. First, media consumption on mobile devices is becoming an inevitable trend, therefore, image retargeting techniques will see greater use in many real-world applications. Quality assessment on retargeted image can play an important role in monitoring and ensuring the experience of mobile media consumption. Second, conventional image quality assessment work typically focus on distortion caused by operations, such as compression, additive noise, blurring, and contrast/brightness changes. These types of distortions mainly alter the intensity and noise level of the image, but do not change the image size and the content structure inside the image. Therefore, exact correspondence can be established to compute local distortion in the full-reference scenario. Retargeting operations, on the other hand, bring unique challenges to quality assessment, because the internal structure of the image content will be changed and the distortion will be mainly structural rather than noise or intensity change. Such challenges call for new techniques that can explicitly measure image quality degradation due to structure changes.

The contributions of our work include: (1) We extend the idea of forensic hash for the application of quality assessment of retargeted images in a reduced-reference scenario. (2) We focus on distortions caused by image structure changes, which are different from distortions of noise and intensity changes typically considered in conventional quality assessment work. (3) We propose novel algorithms for the structural distortion analysis, and provide not only a single quality score, but also a detailed distortion map that can be used to assist human evaluation. The quality assessment results from our proposed algorithm can have many extended applications, such as classifying different retargeting operations and reconstructing the original image from the retargeted one.

In the rest of section, we first describe the framework of the quality assessment algorithm in Section 3.2.2. Then we present the details of the algorithm in Sections 3.2.3 and 3.2.4, including selecting partial information and structural distortion analysis using such partial information. Finally, we present some examples of the quality assessment results in Section 3.2.5.

3.2.2 Quality Assessment Framework

In this work, we consider reduced-reference quality assessment, which means that only partial information from the original image is available to evaluate quality degradation. As we mentioned earlier, the distortion considered here is mainly structural changes caused by the retargeting operations. Therefore, the partial information needs to compactly capture the structure information of the image and allow robust analysis of structure changes. The partial information used in this work is strong corner points from the original image. After retargeting, the corner points from the retargeted image will be matched against the corner points encoded in the partial information. After matching, we can obtain a detailed distortion map of how different parts of the image are changed from the original reference image. Finally, such a distortion map is analyzed by taking into account of the image saliency to produce a quality score that measures the amount of structural distortion. The overall framework of the proposed algorithm is shown in Fig. 3.7.

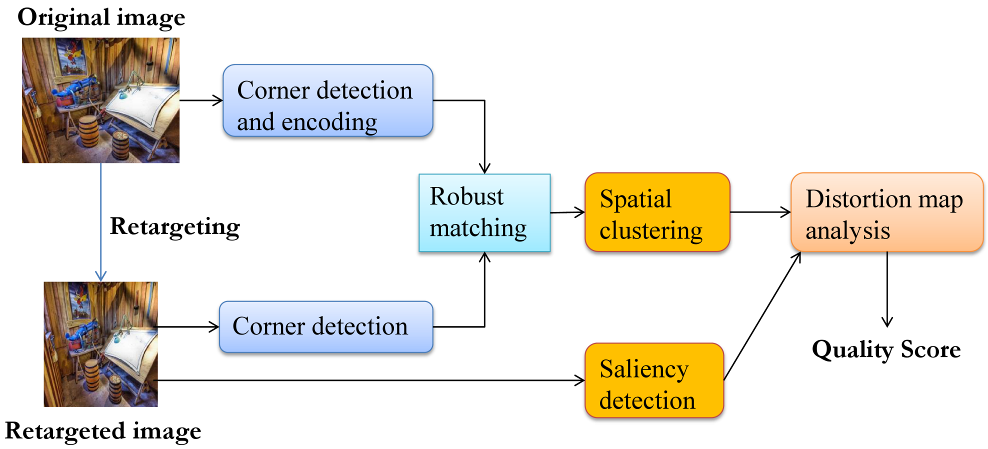


Figure 3.7: Overall framework of reduced-reference quality assessment for retargeted images

The output of this quality assessment algorithm includes the distortion map that describes the structural distortion of the retargeted image in detail and a quality score computed from the distortion map. The role of the quality score is similar to that in conventional quality assessment work, i.e., to predict the quality degradation as would be perceived by a human observer. In the case of structural distortion, recent study by Rubinstein et al. [89] found that even human beings have difficulties in judging quality of retargeted images and have large discrepancies on how important is each type of distortions, such as content loss, symmetry violation, distorted edges/lines, and deformed objects/faces, etc. Therefore, providing a detailed distortion map in addition to a single quality score can be very important to assist subjective quality assessment on the retargeted images. In the next two sections, we provide detailed discussion on each component in the proposed quality assessment algorithm.

3.2.3 Partial Information Selection

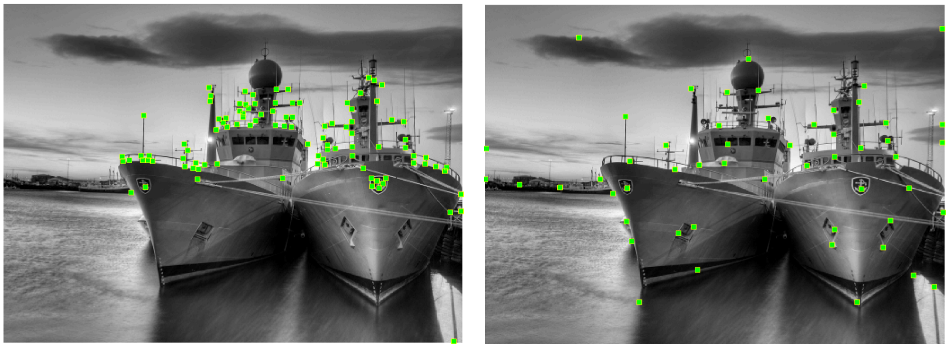
In order to evaluate structural distortion, the partial information that we select for quality assessment needs to capture the structure information of the reference image. Image structure can be captured by edges. However, using edge map as partial information is not compact enough. To compactly capture image structure information, we use corner points in this work. There have been wide variety of interest point and corner point detectors in the literature. They can be divided into three categories: contour based, intensity based, and parametric model based methods. A comprehensive survey and comparison of local interest point detectors can be found in [95].

In this work, we use the corner point detection method proposed by Harris and Stephens [44], where the corner points roughly capture the structure information of the image as they typically locate on dominant image structures, and they are also efficient to compute as compared to more computationally intensive local features such as SIFT. The Harris corner point detector identifies corner points using the gray value information of the local patch around the point. More specifically, a Harris matrix is computed for each point as

$$A = \sum_u \sum_v w(u,v) \begin{bmatrix} I_x^2 & I_x I_y \\ I_x I_y & I_y^2 \end{bmatrix}. \tag{3.1}$$

Here, I_x and I_y are the partial derivative of the image intensity along horizontal and vertical directions. $w(u, v)$ are the weights assigned to each neighboring pixel. A Gaussian weighting can be applied to provide an isotropic response. A corner point can then be characterized by analyzing the eigenvalues of the Harris matrix A. Denote the two eigenvalues of A as λ_1 and λ_2. If $\lambda_1 \approx 0$ and $\lambda_2 \approx 0$, the pixel (x, y) is not an corner point nor an edge point. If $\lambda_1(\lambda_2)$ has large positive value and $\lambda_2(\lambda_1) \approx 0$, then an edge point is found. If both λ_1 and λ_2 have large positive values, the pixel (x, y) is considered a corner point. For efficient detection of corner points, the following function $\det(A) - k \cdot \mathrm{trace}^2(A)$ involving only determinant and trace of A can be used instead of computing the eigenvalues.

Since the partial information needs to be compact, we will need to control the number of corner points that are selected. First, we would like the corner points to appear on major object or structure inside the image, rather than capturing small details such as texture on the background. To suppress noise and retain only corner points on salient objects, we perform smoothing on the image before doing the corner point detection. A large amount of blurring can remove noisy corner points that might appear on smaller edges that may not be useful for evaluating the global structure distortion. Second, to avoid large amount of corner points in a small neighborhood, local maximum suppression is necessary to retain only the dominant corner in a local patch. The larger window for the local maximum suppression, the more spread we can expect from the detected corner points. An example of corner point detection is shown in Fig. 3.8, where we can see that a larger maximum suppression window causes the detected corner points to spread out more.



(a) Corner points (window size = 3) (b) Corner points (window size = 15)

Figure 3.8: Corner points detected at different maximum suppression window sizes.

After corner points are detected, we need to compactly encode their positions. By considering the corner map as a binary image, where corner points are represented by value 1 and all other pixels in the image have value 0, we can use the JBIG2 encoding [78] to compactly represent such a binary image. JBIG2 is an international standard that provides a very good compression ratio for binary images and serves as a good candidate for our application. The size of JBIG2 compressed corner map with respect to different number of corner points is shown in Table 3.3. For comparison, we also show the size of the corner map encoded by PNG format. We can see that JBIG2 compression provides more than 40% savings in representing the partial information of corner points than simple PNG encoding.

Table 3.3: Compact encoding of corner points

# of corner points	50	80	100	120	150	200
JBIG2 size (byte)	225	276	310	339	391	469
PNG size (byte)	380	475	541	598	682	852

3.2.4 Structural Distortion Analysis

To evaluate structure distortion, we first need to measure how the corner points from the original reference image change in the retargeted image. Since retargeting operation will introduce highly non-linear and adaptive changes to the image structure, a robust matching between corresponding corner points between the original and retargeted images is very important to enable accurate distortion measure. After finding an accurate correspondence of corner points, we can then analyze the distortion and compute the quality score. Next, we describe each step in the structural distortion analysis.

Robust matching of corner points The algorithm of corner points matching is shown in Fig. 3.9. The first step is to find an optimal alignment between the binary corner map from the original image and the retargeted images. This step is important as it helps mitigate the effect of cropping which is a common operation in retargeting. In the second step, we establish corner point correspondence that minimizes some cost function. Finally in the third step, we perform outlier removal to remove false correspondences and get the final matching result.

Optimal alignment
Point correspondence
Outlier removal

Figure 3.9: Three steps in finding corner point correspondence

Since the retargeted image may have undergone cropping, we will first align the binary corner map with the retargeted image so that the point correspondence can be found more accurately. The process of finding such an optimal alignment is shown in Fig. 3.10. Given the retargeted image, we first compute its edge map and perform distance transform on the edge map. The distance transform can be represented as

$$dt_E(p) = \min_{p_e \in E} \|p - p_e\|_2. \tag{3.2}$$

Here, E is the edge map of the retargeted image, dt_E is the distance transformed image, p_e is a point in E, and p is a point in dt_E. After distance transform, we obtain an image where each pixel's value represents its distance to the closest edge. To find the optimal alignment, we overlay the corner map from the original image onto the distance transformed image at different shift positions. For each shift poisition

$s = (x, y)$, we compute the Chamfer distance [5] defined as

$$d(s) = \frac{1}{N} \sum_{p_c} dt_E(p_c + s). \tag{3.3}$$

Here, p_c is the corner point position, s is the shift, N is the total number of corner points encoded in the partial information. A smaller distance indicates that most of the corner points are located closer to the edges in the retargeted image, therefore, a better alignment. The optimal alignment is then computed as the shift s^* that gives the minimum Chamfer distance.

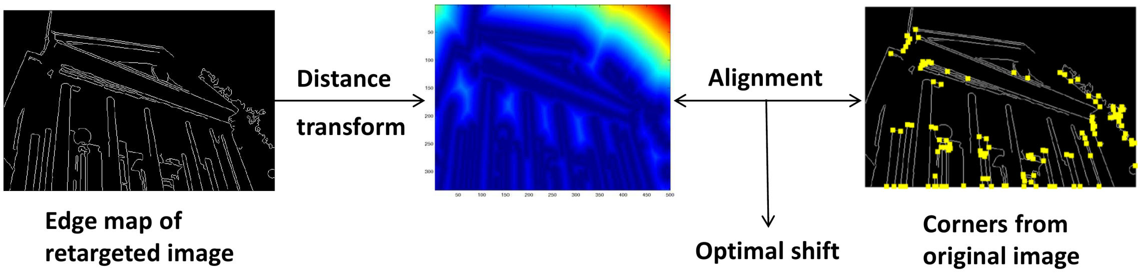


Figure 3.10: Finding optimal alignment between corner points and retargeted image

After alignment, we will compute the correspondence between the corner points from the original image to the corner points in the retargeted image. Retargeting operation changes the internal structure of the image by scaling, carving, and warping. Therefore, the task here is to find correspondence between two sets of points, where one set of points are deformed from the other set of points. This task is very similar to the task of shape matching, where one shape needs to be matched against another similar but deformed shape. Shape context proposed by Belongie et al. [7] is a useful technique for shape matching, and we use it here to find correspondence between corner points.

Shape context is a local descriptor associated with each shape point or corner point to describe the coarse distribution of the rest of corner points with respect to the current point. An illustration on how to compute shape context is shown in Fig. 3.11, where the two sets of points represent the same character “A” but has slightly different shapes. For each point, its shape context is essentially a histogram counting the number of other points that fall into each of the bins shown in the right-most part of the figure. Different bins capture shape information in different orientation and distance relative to the current point.

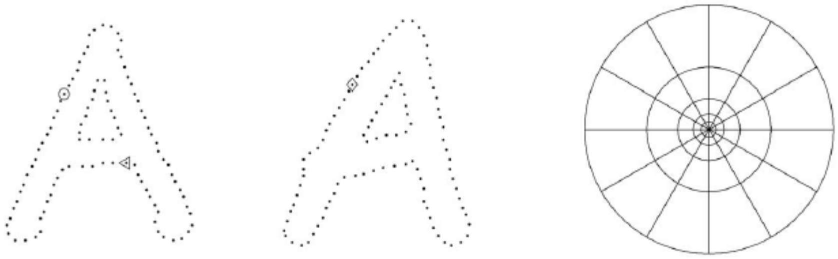


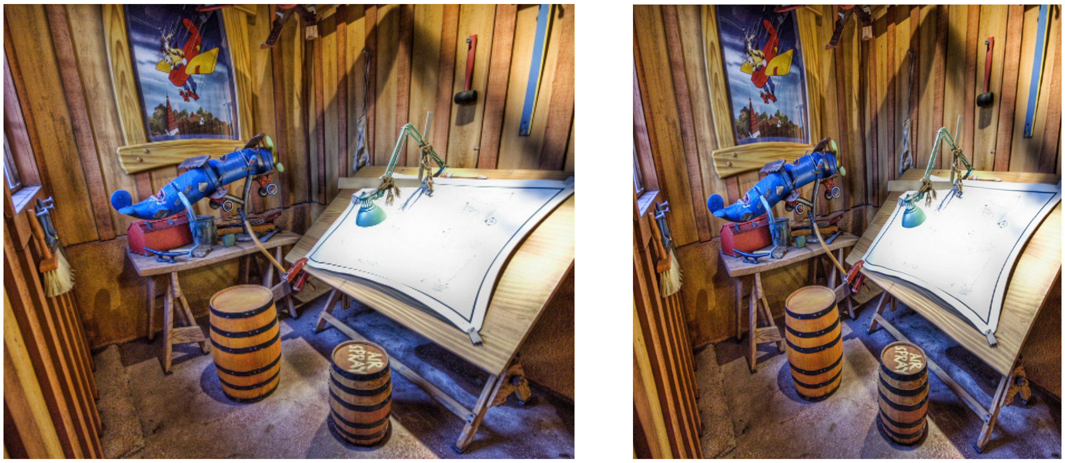
Figure 3.11: Illustration of shape context computation (Figure from Belongie et al. [7])

With shape context computed for each corner point, finding correspondence is then equivalent to finding for each corner point in one set the corner point on the other set that has the most similar shape context. The distance between two shape context can be computed using the χ^2 test statistic:

$$C(p,q) = \frac{1}{2}\sum_{i=1}^{K} \frac{[h_p(i) - h_q(i)]^2}{h_p(i) + h_q(i)}, \tag{3.4}$$

where $h_p(i)$ and $h_q(i)$ denote the K-bin normalized shape context histogram at points p and q, respectively. Given the shape context distance between all pairs of corner points p_i in the original image and corner points q_j in the retargeted image, we need to find a one-to-one correspondence such that the total cost of matching $\sum_i C(p_i, q_{\pi(i)})$ is minimized. Here, π is a permutation representing the one-to-one correspondence. In our current problem, we have aligned the corner points between original image and retargeted image, therefore, we expect that for the correct correspondence the distance between two matched points will be small. We can introduce an additional term into the cost function so that $C(p,q) = \alpha C_s(p,q) + (1-\alpha)C_d(p,q)$, where $C_s(p,q)$ is the shape context distance and $C_d(p,q)$ is the distance between the two points p and q. This optimal matching problem can be solved using the Hungarian method [81] in $O(N^3)$ time.

The correspondence found using above method minimizes the shape context difference and distance between matched corner points. However, there is no explicit cost term that minimizes the difference of displacements between neighboring matched points. More specifically, if one corner point in the original image is shifted to the left by 10 pixels, we would expect the neighboring corner points in the original image are also shifted to the left by similar amounts. This is because the retargeting operation typically tries to minimize the distortion when scaling down or removing parts of the image. Therefore, to refine the matching results, we perform an additional step of outlier removal. For each corner point p_i in the original image, we compute a displacement vector $v_i = q_{\pi(i)} - p_i$, which captures the displacement of its corresponding point in the retargeted image. Then we compare v_i with displacement vectors $\{v_j\}$ of points $\{p_j\}$ that fall into close neighborhood of p_i. If the difference between v_i and average of $\{v_j\}$ is larger than certain threshold, we consider $(p_i, q_{\pi(i)})$ a false match. An example of finding corresponding corner points are illustrated below. In Fig. 3.12, we show the original image and its retargeted version. In Fig. 3.13, we show the matching results without and with outlier removal. We can see that outlier removal helps us to obtain more accurate matching, so that distortion evaluation can be more accurate.

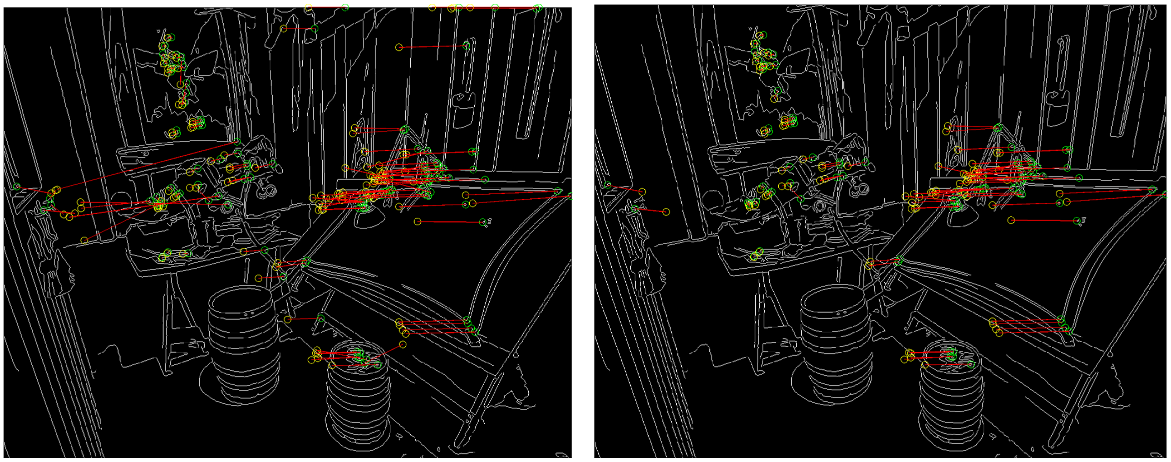


(a) Original image

(b) Retargeted image

Figure 3.12: Original image and its retargeted version. Images from RetargeMe database [2]

Global distortion measure Given the correspondence between corner points of original and retargeted images, we can measure the global structural distortion due to the retargeting operation. Several distortion measures are considered in this work. The first one is global affine cost. Since we know the positions of all the corner points, we can use the correspondence to estimate a global affine transform that transforms the set of corner points in the original image to the set of corner points in the retargeted image. If we only consider scaling and rotation, the affine transform matrix can be represented as a 2 by 2 matrix A. The singular value



(a) Without outlier removal

(b) With outlier removal

Figure 3.13: Correspondence between corner points in the original and retargeted images

decomposition of the matrix A can be denoted as $A = O_{\theta_1} D O_{\theta_2}$, where D is a diagonal matrix of singular values λ_1 and λ_2. The physical meaning of the singular value decomposition is that the transform basically scales along the direction θ_1 with factor λ_1 and scales along the orthogonal direction θ_2 with factor λ_2. If λ_1 and λ_2 are close to each other, the transformation roughly preserves the aspect ratio of the image; if λ_1 is much larger than λ_2, then large distortion can be expected because the aspect ratio of the image will be changed significantly. The global affine cost is defined as

$$C_{ga} = \log(\lambda_1/\lambda_2). \tag{3.5}$$

The second measure that we use here is called global bending energy. It is computed from the transformation estimated using the thin plate spline (TPS) model [25, 72]. Thin plate spline is a useful tool for interpolating surfaces over scattered data. Therefore, it can be used to estimate transformation over the entire image from a limited number of point correspondences. TPS model also includes the affine model as a special case. Bookstein [10,11] showed that TPS model is very effective to model biological shape changes as deformation. There, two thin plate spline mappings $[f_x(x,y), f_y(x,y)]$ are used, each mapping the x and y coordinates of corresponding points, i.e., $f_x(x,y) = x'$ and $f_y(x,y) = y'$, where (x,y) and (x',y') are matched points. Then the transformation $f(x,y) = [f_x(x,y), f_y(x,y)]$ is estimated by minimizing the bending energy defined as

$$C_{gb} = \int\int ((\frac{\partial^2 f}{\partial x^2})^2 + 2(\frac{\partial^2 f}{\partial x \partial y})^2 + (\frac{\partial^2 f}{\partial y^2})^2) dxdy. \tag{3.6}$$

The physical meaning of bending energy measures how twisted the estimated surface is. In this work, we use C_{gb} as the global bending energy to measure global structural distortion. A smaller distortion will generate smoother transformation and thus smaller bending energy.

The third global distortion measure used here captures the amount of content loss due to the retargeting operation. Since we only have partial information about the original image, we measure the content loss by the number of corner points that are not matched to any points in the retargeted image:

$$C_{cl} = \frac{N_o - N_m}{N_o}, \tag{3.7}$$

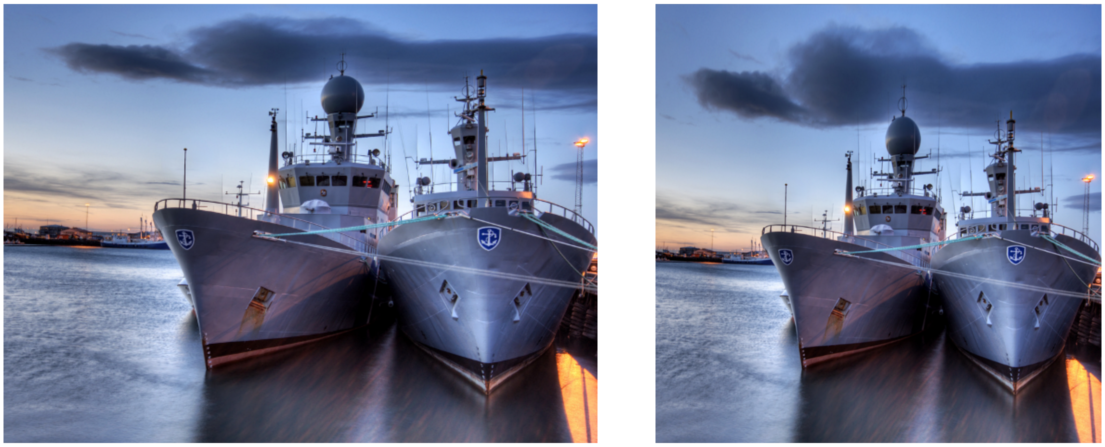
where N_o is the number of corner points in the original image and N_m is the number of matching points established between the original and retargeted images. Since stronger corner points tend to locate at salient objects or structures in the image, a smaller content loss indicates a better retargeting quality.

Local distortion analysis One of the advantages of reduced-reference quality assessment is that we have partial information from the original image, which can enable us to provide detailed analysis on quality degradation. In addition to the global distortion scores mentioned earlier, in this part, we provide analysis on local distortions caused by the retargeting operation.

The main idea of state-of-the-art retargeting operations is to resize different parts of the image differently through scaling, warping, or carving. Salient objects or regions will be scaled more uniformly so that the aspect ratio of the object is preserved and minimum distortion is introduced. For less important regions, the resizing operation will have less constraint because a large distortion can be tolerated. Such a main spirit of retargeting motivates us to perform quality assessment in a similar manner, i.e., to evaluate the distortions locally and differently for different regions of the image. Salient regions will have higher penalty on large distortions and non-salient areas will have smaller weight on their contribution to the overall distortions. We perform a two-step evaluation of the local distortion in this manner. In the first step, we perform spatial clustering on the matched corner points such that within each cluster, the corners will be spatially close to each other and have consistent displacement patterns. In the second step, we weigh the distortion of each local cluster with their saliency to get a final quality score.

Spatial clustering, which group similar spatial objects into classes, is an important component of spatial data mining [42]. Spatial clustering techniques can be classified into four categories: partitioning method, hierarchical method, density-based method and grid-based method. A detailed survey of spatial clustering and their classification can be found in [43]. In this work, we use K-meloid clustering to cluster corner points. K-meloid is similar to K-means clustering but instead of using average of cluster elements as centroid, K-meloid uses the most central element in the cluster as centroid. This makes K-meloid clustering more robust to noise and outlier data than the simple K-means clustering. We use an efficient K-meloid clustering method called Clustering Large Application using RANdomized Search (CLARANS) [75] for its efficiency and good quality of clustering. The basic idea of CLARANS is to perform a randomized search of node with minimum cost in a graph. The node here represents the selection of K meloids and the graph is composed of such nodes and neighboring nodes different at only one of the K meloids. The use of randomized search provides the advantage of efficient and the benefit of not confining the search to only a localized area. Both K-mean and K-meloid are partition-based clustering, therefore, they have the limitation of requiring specifying the number of clusters K at the beginning. To alleviate such a constraint, we take the following approach: we select a relatively large K to start with, and then merge neighboring clusters if their displacement vectors are very similar, and split a cluster if the displacement vectors within are diverse. An example of the spatial clustering is shown below, where Fig. 3.14 shows the original and retargeted images and Fig. 3.15 shows the corner point correspondence and their clustering result. Corner points of different clusters are labeled with different colors.

Each cluster of corner points obtained from the above spatial clustering method will have consistent displacement vectors, and thus they are expected to cover certain objects inside the image that have undergone resizing. In this work, we further



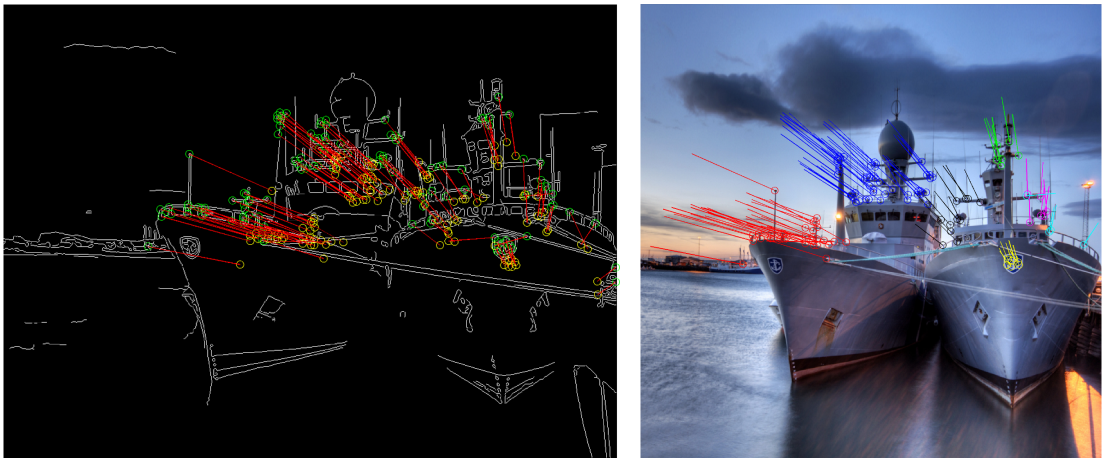
(a) Original image

(b) Retargeted image

Figure 3.14: Original image and its retargeted version.

utilize the image saliency information so that different regions/objects of the image contribute differently to the final quality score. Saliency is a subjective measure that captures what human observers consider as important in an image. Measuring saliency is an important step in image retargeting and different saliency metrics have been used in the literature, such as gradient magnitudes [118] and discontinuity of neighbors if a pixel is removed [90]. In this work, we use the saliency measure proposed by [110], which combines the gradient magnitude and the saliency map by Itti et al. [45]. The gradient information captures structural areas and the saliency map by Itti et al. captures attractive areas that have different color, intensity and orientation properties than their surroundings. The final saliency image is evaluated as $W = W_\alpha \times W_\beta$, where $W_\alpha = ((\frac{\partial}{\partial x}I)^2 + (\frac{\partial}{\partial y}I)^2)^{0.5}$, and W_β is the saliency map by Itti et al. An example of the saliency evaluation is shown in Fig. 3.16.

The clusters of corner points indicate how different parts of the image have been transformed during the retargeting operation. The saliency map tells us how



(a) Corner point correspondence

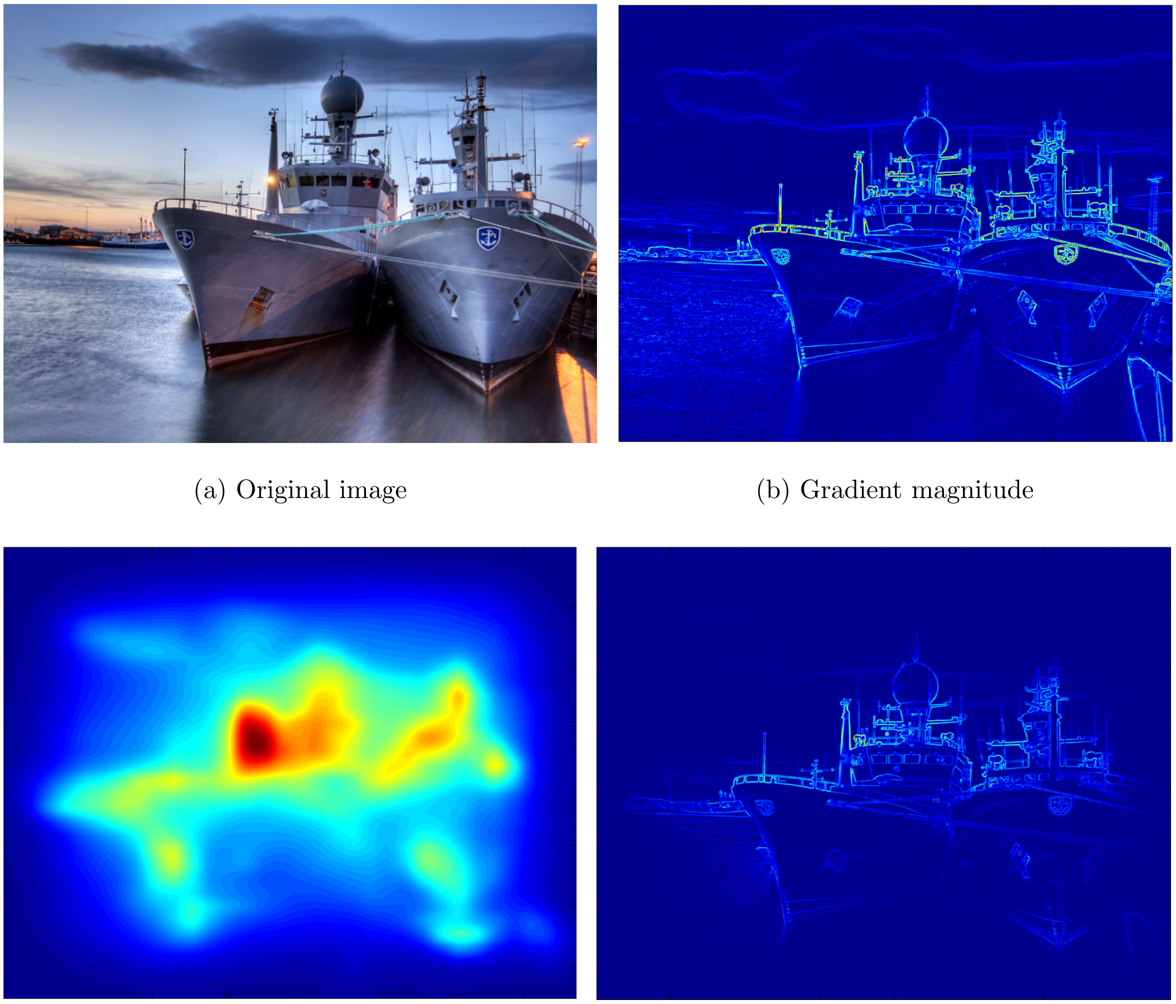
(b) Spatial clustering result

Figure 3.15: Corner point correspondence and spatial clustering result

important different areas of the image are. Combining the two, we can compute some quality score that measures local distortion of retargeting in terms of whether each region is resized uniformly and how different neighboring regions have been resized. Denote the set of corner point clusters as $C_1, \cdots, C_k$. For each C_i, we select l neighbor clusters $C_{i1}, \cdots, C_{il}$ that are within certain distance threshold to the current cluster. The distortion contributed by two neighboring clusters is computed as

$$d_{ij}(C_i, C_{ij}) = \|v_i - v_{ij}\| \cdot d_s e^{-\alpha \cdot d_e}. \tag{3.8}$$

Here v_i and v_{ij} are the displacement vectors of the meloids in cluster C_i and C_{ij}, respectively, d_e is the distance between the two clusters. $d_s = \frac{S_{ij}}{\max(S_i+S_j,S_{ij}}$ measures how likely the two clusters cover the same object. S_i and S_j are the saliency values of regions covered by C_i and C_{ij}, respectively and S_{ij} is the saliency of regions between C_i and C_{ij}. If two regions of high saliency are connected also by regions



(a) Original image

(b) Gradient magnitude

(c) Saliency map by Itti et al. [45]

(d) Final saliency map

Figure 3.16: Saliency map of an image

of high saliency, these two regions are likely to cover the same object. If they are connected by regions of low saliency, they may cover different objects therefore the weight assigned to the discrepancy between their displacement vectors will be lower. The quality score that measures such cluster-wise inconsistency is denoted as

$$C_{co} = \sum_{i=1}^{K} \sum_{j=1}^{l} d_{ij}(C_i, C_{ij}). \tag{3.9}$$

In summary, we have defined three quality metrics, namely global affine cost

C_{ga} in equation 3.5, global bending energy C_{gb} in equation 3.6, content loss C_{cl} in equation 3.7 and cluster inconsistency C_{ci} in equation 3.9. In the next section, we provide experimental results on using these quality scores on images undergone different retargeting algorithms.

3.2.5 Experimental Results

In this section, we present some experimental results on using corner points as partial information for reduced-reference quality assessment on images retargeted by different algorithms.

Experimental setup The retargeted image database that we used here is from the RetargetMe project [2], which contains 80 color images, each of which is retargeted by 8 different retargeting algorithms. The compared retargeting methods are: nonhomogeneous warping (WARP) [118], Seam carving (SC) [90], Scale-and-Stretch (SNS) [110], Multi-operator (MULTIOP) [91], Shift-maps (SM) [85], Streaming Video (SV) [54], and Energy-based deformation (LG) [50]. More details can be found in [89]. In addition to providing the images undergone different retargeting algorithms, the RetargetMe project has carried out an extensive subjective evaluation on the retargeted images by involving human observers through Amazon's Mechanical Turk service. In the last part of this section, we measure how well our proposed quality metrics correlate with users' subjective ratings.

Distortion analysis using partial information In this part, we demonstrate how the proposed quality scores correlate with human observation through several examples. In Fig. 3.17, we show the original image, the images retargeted by simple cropping and scaling (CR), seam-carving (SC), and streaming video (SV). We choose these three retargeting methods to compare here because they capture the three major types of retargeting effects. Cropping and scaling retains the aspect ratio of salient objects, and is found to be preferred by many human observers, especially when the original image is not available for comparison [89]. Seam carving resizes an image by removing seams of low energy. This will often change the aspect ratio of salient objects when the resizing factor is large. Streaming video represents the state-of-the-art retargeting that tries to scale different parts of the image different so that edge discontinuity as observed in seam carving can be reduced.

In Fig. 3.18, we show the point correspondence between the retargeted images and the original image. The correspondences here clearly demonstrate the different types of retargeting effect. For the image undergone cropping, we can see that the matched corner points have no displacement. For seam carving, we can see that the displacement of corners is mostly horizontal which means vertical seams have been removed from the image. Also, different regions have different displacement sizes, which indicates the different amount of carving at different regions. For image undergone streaming video, we can see the retargeting has a scaling effect, which scales down the salient region, i.e., the ship, quite uniformly so that we do not see the kind of structure distortion as we can see in the seam carved image.

The quality scores computed on these retargeted images is shown in Table 3.4.

We can see that cropping causes the least amount of affine cost and bending energy. This is expected because no structural distortion is introduced from the cropping operation. Seam carving has the highest affine cost, bending energy and cluster inconsistency, because the structural change caused by SC cannot be modelled well by affine. Streaming video achieves much lower C_{ga}, C_{gb} and C_{ci} because the structural change introduced by SV is smooth and can be well modelled by an affine transform. In this example, there is no corner points outside the ship, so cropping has no content loss. But in general, we will expect cropping to have larger amount of content loss than other methods. Overall, we can see that the proposed metrics correlate well with subjective expectations on this example.

Table 3.4: Quality scores on images retargeted by CR, SC, and SV methods

Quality metric	CR	SC	SV
Global affine cost C_{ga}	5.6×10^{-4}	0.32	0.09
Global bending energy C_{gb}	1.6×10^{-3}	7.34	0.62
Content loss C_{cl}	0%	2.5%	3.3%
Cluster inconsistency C_{ci}	0.04	0.14	0.05

Such a displacement field estimated from the corresponding corner points between the retargeted image and the original image not only can be used to compute several scores for automatic quality evaluation, but more importantly, it provides an important tool to assist human evaluation when the original image is not available. In such cases, the compact side information of the corner points actually provides very rich information on how the image has been retargeted. Since different persons will have different preference on various kinds of distortions, such as edge discontinuity, violated symmetry, content loss, etc., such an objective displacement map allows users to make their own judgement and preferences instead of being forced to accept a single quality score computed for them. This can be considered as one important advantage of the reduced-reference quality evaluation algorithm proposed in this work.

Correlation of objective quality score with subjective ratings We also compute the correlation of the objective quality score with subjective user ratings collected in [89]. For every pair of the eight retargeting methods compared in [89], a human observer is presented with two images retargeted using the two methods and asked about his/her preference in terms of which of the two images looks better. There are two scenarios: in the first one the user is given only the retargeted images and in the second scenario, the user is also given the original image as reference. Such experiment is repeated on 210 participants and leads to the collective subjective rating for each of the retargeting methods. In [89], the authors compute the correlation between the objective quality score with the subjective rating by ranking the eight retargeting methods and then computes the Kendall-τ correlation between the subjective ranking and objective ranking:

$$\tau = \frac{n_c - n_d}{\frac{1}{2}n(n-1)}, \tag{3.10}$$

where n_c is the number of concordant pairs and n_d is the number of discordant pairs over all pairs of entries in the two rankings. τ will have values between $[-1, 1]$, where a value of 1 indicates highest correlation or perfect agreement while a value of -1 is the case of perfect disagreement. Furthermore, to test the significance of the correlation values, a χ^2 hypothesis test is carried out against the null hypothesis that the observed correlations are zero mean, which means the subjective and objective scores are uncorrelated.

Below, we compute the Kendall-τ correlation scores and the p-values for the proposed quality metrics, as well as for some metrics in [89]. In Table 3.5, we show the results of correlation against the subjective ratings when users are not given the reference image, and Table 3.6 shows the results for the scenario of with the reference image. Bidirectional Similarity (BDS) is an image similarity measure proposed by [99]. For each patch in one image, a well-matched patch is sought in the other image and the distance of two images is defined as the mean distance in color space between corresponding patches. Bidirectional warping (BDW) is a similar metric with the exception that the mapping between the two images is constrained to be one-way. It is used in [91] as a similarity metric to optimize the retargeting process. The Earth Mover's Distance [82] measuring the similarity of two distributions by computing the minimum cost required to transform one distribution into the other. It should be noted that these three metrics all require the original image to be available and thus fall into the scenario of full-reference quality assessment.

From these two tables, we can see that the full-reference metric EMD performs the best among all metrics and the positive correlation is statistically significant (p-value ¡ 0.05). The proposed metric cluster inconsistency C_{ci} has positive correlation with subjective ratings and its performance comes close to BDS in the scenario

Table 3.5: Correlation with subjective evaluation without reference

Quality metric	Mean	std	p-value
BDS	0.12	0.28	0.0067
BDW	0.068	0.29	0.085
EMD	0.238	0.24	1e-6
Global affine cost C_{ga}	0.052	0.25	0.11
Global bending energy C_{gb}	0.031	0.26	0.24
Content loss C_{cl}	0.081	0.25	0.03
Cluster inconsistency C_{ci}	0.11	0.32	0.018

without reference and close to BDW in the scenario with reference. The proposed metric content loss C_{cl} also has positive correlations in both scenarios, and it is interesting to notice that in the scenario of with reference, the correlation of C_{cl} significantly higher than the without reference case, which indicates that when the user sees the original image, they may give loss of content a higher weight in the subjective quality evaluation process. The other two metrics global affine cost and global bending energy do not have significant correlation with the subjective ratings. Overall, we see that even the highest correlation by EMD is not very high, which indicates that it is very challenging to design a quality metric that can correlate well with human subjectives. Similar results are also demonstrated in [89], where it is shown that even humans tend to disagree with each other on quality assessment of retargeted images.

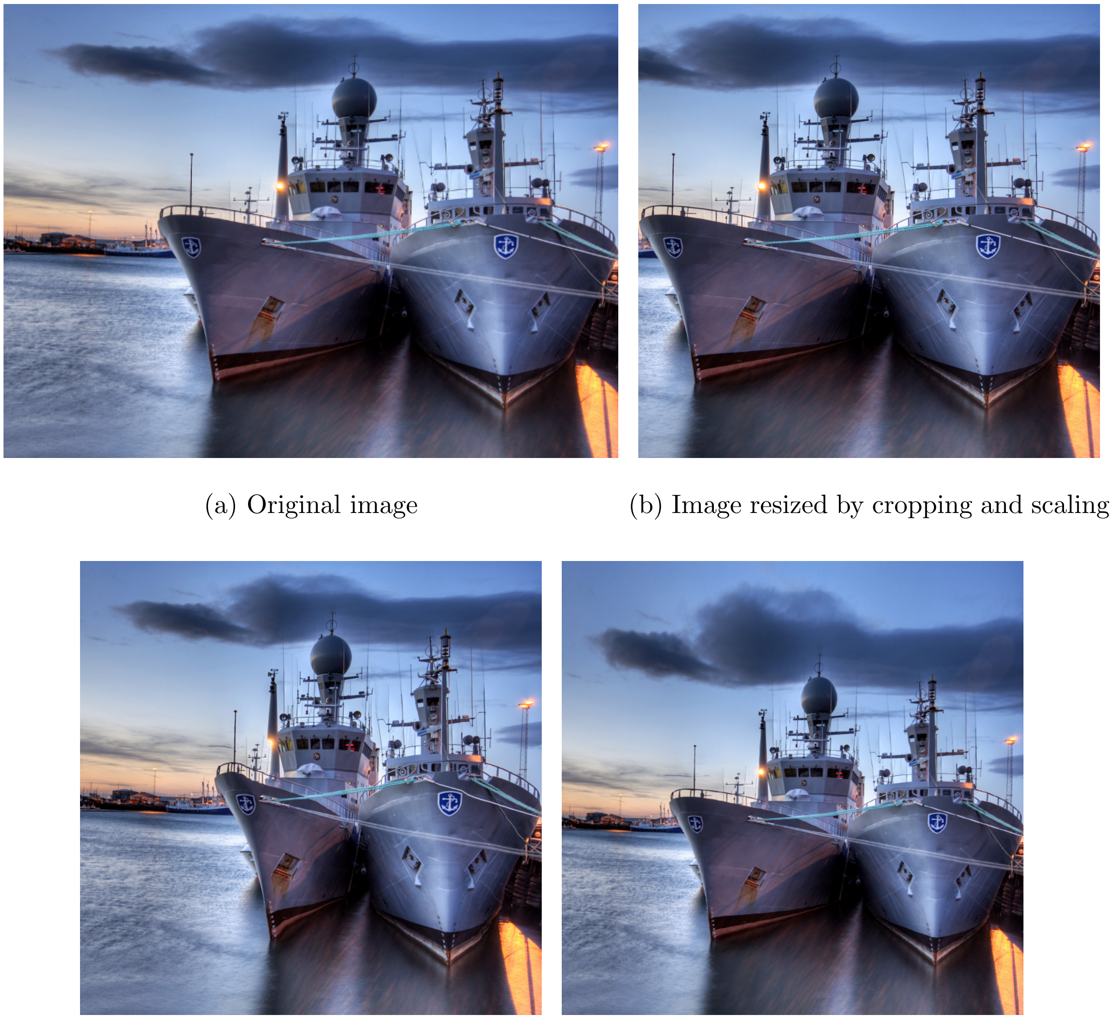
Table 3.6: Correlation with subjective evaluation with reference

Quality metric	Mean	std	p-value
BDS	0.06	0.30	0.11
BDW	0.10	0.32	0.026
EMD	0.25	0.28	1e-6
Global affine cost C_{ga}	0.07	0.25	0.052
Global bending energy C_{gb}	0.03	0.29	0.28
Content loss C_{cl}	0.16	0.25	0.0003
Cluster inconsistency C_{ci}	0.12	0.28	0.0081

Nevertheless, we want to stress that the quality scores proposed in this work only uses very compact partial information from the original image. They cannot achieve the best performance of a full-reference quality metric (EMD), but still shows good correlation with users' subjective ratings and have performance similar to some of the full-reference quality metrics (BDS, BDW). More importantly, as we showed earlier in this section, the proposed quality evaluation algorithm does not only give a single quality score, but also provides detailed distortion map that can assist human observers in quality assessment when the original image is not available.

3.3 Chapter Summary

In this chapter, we first studied the problem of estimating seam carving using forensic hash. The adaptive nature of seam carving allows effective image tampering against traditional blind forensic techniques. However, we demonstrate that using the forensic hash proposed in Section 2.3.2 without changing its design, we can reliably estimate both the amount and the location of seam carving, and further enable accurate alignment and tampering localization on a modified image. Such detailed information of trustworthiness provided by the forensic hash is important for better utilization of online multimedia information. In the second part of this chapter, we further extended the FASHION spirit to evaluate quality of images that have undergone more general image retargeting operations. Given the increasing popularity of mobile devices and media consumption on different screen sizes, ensuring and monitoring quality of retargeted image can find important applications. We proposed to compactly encode corner points as partial information and compute correspondence of corner points to estimate a detailed distortion map due to the retargeting operation. Quality metrics that capture the global and local structure distortions are proposed. Experiment results show that some of the proposed metrics (content loss and cluster indistinguishability) have statistically significant correlation with human subjective evaluations. Furthermore, the distortion map provides an important tool to assist users in evaluating image quality with their own preferences instead of being forced to accept a single quality score.



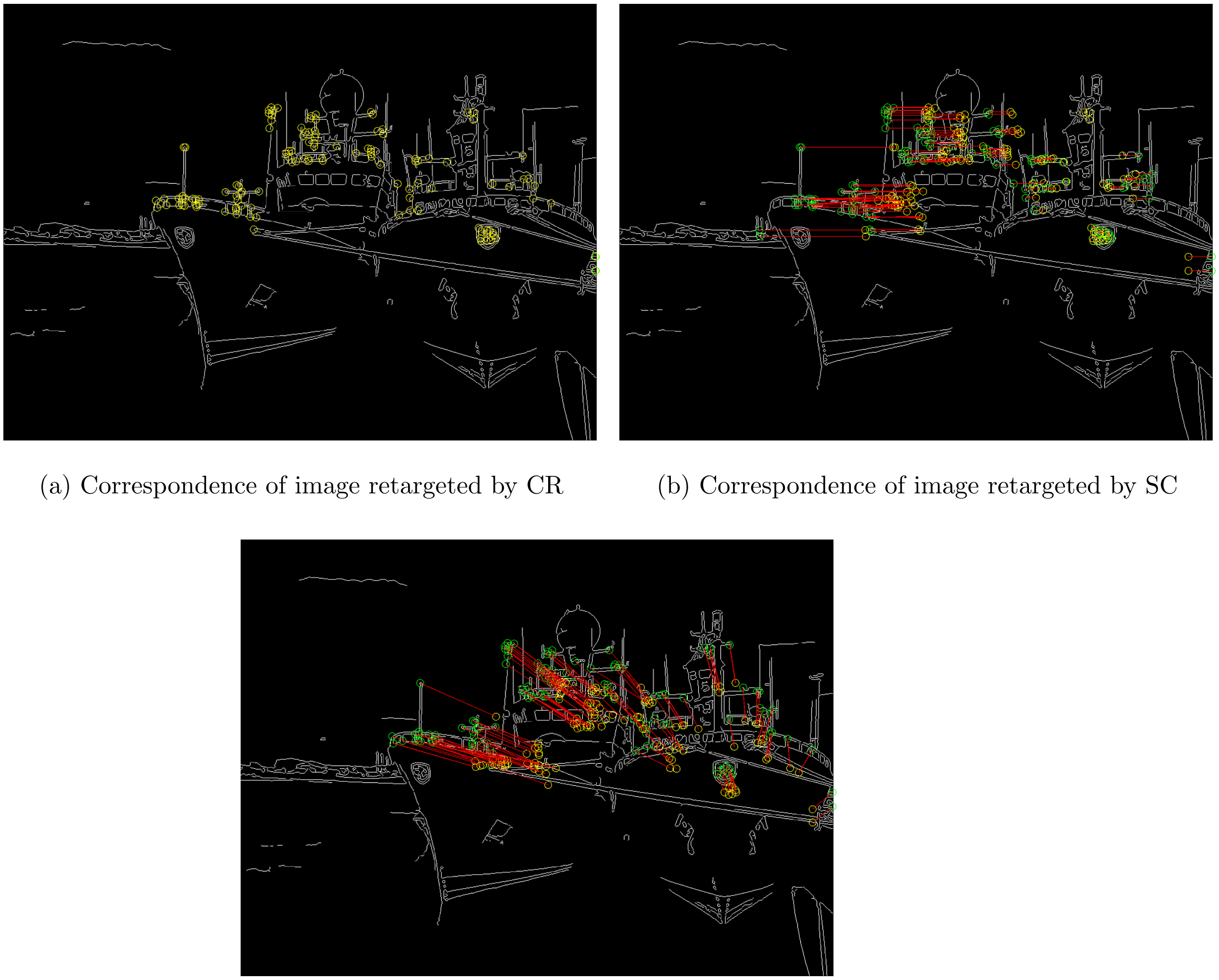
(a) Original image

(b) Image resized by cropping and scaling

(c) Image resized by seam carving

(d) Image resized by streaming video

Figure 3.17: Original image and its retargeted versions



(a) Correspondence of image retargeted by CR

(b) Correspondence of image retargeted by SC

(c) Correspondence of image retargeted by SV

Figure 3.18: Correspondence between retargeted images and original image

CHAPTER 4

Confidentiality-Preserving Search of Multimedia

4.1 Introduction

The advancement of information technology has been rapidly integrating the physical world where we live and the online world that we rely on for retrieving, sharing, and managing information. Online services and web applications emerge everyday and benefit our life in almost every aspect: from information retrieval using search engines to sharing user generated content through social networks, and from personal information management, such as webmail and online photo albums, to online backup services. With the arrival of the cloud computing paradigm, the Internet stores not only information for sharing, but also sensitive personal data that should be carefully protected against any unauthorized access. Secure management of personal data stored online is an increasingly important issue that can help achieve the data confidentiality and availability requirements of cloud computing. Technologies that can enable secure online data management are going to play a critical role in the future of the internet.

Traditional privacy protection for online personal data focuses on access control and secure data transmission, which ensure that the data can be securely transmitted to the server and no unauthorized people can access the data. However, once the data arrives at the server, the server decrypts the data and operates on plaintext in order to provide services to users, such as categorization, search, and data analysis. This makes the user's private information vulnerable to untrustworthy service providers and malicious intruders. For example, most personal emails are stored online as plaintext data and can be viewed by the system administrator. Given the trend that an increasing amount of personal data will be stored at a third-party server, it is both desirable and necessary to develop technologies that can better protect users' privacy without sacrificing the usability and accessibility of the information.

Information retrieval over encrypted databases is a promising technological capability for privacy protection in online information management. Encryption of the data stored on the server helps protect content privacy against untrustworthy service providers and malicious intruders, but using traditional cryptographic ciphers alone makes it difficult for the server to process the data, and for the user to retrieve information from the encrypted database. The goal of information retrieval over an encrypted database is to provide efficient and accurate search capability over encrypted documents without decrypting them first. An example application is that a user stores his/her private data in encrypted form on web servers and later wants to search and retrieve data in a privacy preserving manner. The server here merely provides the storage and search capability, and should not be able to decrypt the private data. For privacy protection, the amount of information that the server can learn from the user's data set should be kept minimal. Due to the widespread use of digital cameras and portable camcorders, multimedia data has become a significant part of today's personal data collections. Storing and managing this large volume of multimedia data online is a desirable option for convenient data access anywhere anytime. Technologies that can enable content-based retrieval over encrypted multimedia databases will play an important role in helping people manage their multimedia data both effectively and securely. The main focus of the current work is to explore efficient techniques for such an application.

Related Work

Prior work in the area of information retrieval in the encrypted domain focused on text documents. Song et al. [102], Brinkman et al. [14], and Boneh et al. [9] explored Boolean search to identify whether a query term is present in an encrypted text document. Swaminathan et al. [104] proposed a framework for rank-ordered search over encrypted text documents, so that documents can be returned in the order of their relevance to the query term. In that work, several protocols are studied to address different operational constraints such as different communication cost allowed to perform the secure search. Secure text retrieval techniques can also be applied to keyword based search of multimedia data. However, keyword search relies on having accurate text description of the content already available, and its search scope is confined to the existing keyword set. In contrast, content-based search over an encrypted multimedia database provides more flexibility, whereby sample images, audios or videos are presented as queries and documents with similar audio-visual content in the database are identified.

An emerging area of work related to confidentiality preserving multimedia retrieval is secure signal processing, which aims at performing signal processing tasks while keeping the signals being processed secret. Erkin et al. [29] provided a review of related cryptographic primitives and some applications of secure signal processing in data analysis and content protection. However, applying cryptographic primitives to content-based multimedia retrieval is not straightforward. Effective multimedia retrieval typically relies on evaluating the similarity of two documents using the distance between their visual features, such as color histograms, shape descriptors, or salient points [23]. By design, traditional cryptographic primitives do not preserve the distance between feature vectors after encryption. Given the much larger data volume for multimedia data than that of text and other generic data, efficiency and scalability are also critical for multimedia retrieval but can be difficult to achieve using cryptographic primitives alone. Another work by Shashank et al. [98] addresses the problem of protecting the privacy of the query image when searching over a public database, where the images in the database are not encrypted. By appropriately formulating the query message and response message during multiple rounds of communication between the user and the server, the server is made oblivious to the actual search path and thus unaware of the query content.

Recent work in the area of secure computation for privacy protection addressed related but different problems under various application settings [28, 47, 79, 92, 119, 121]. Yiu et al. [121] considered privacy preserving range query over geospatial coordinates using a k-dimensional tree. Extending such a technique to multimedia retrieval is difficult because features used for content-based multimedia retrieval are high dimensional vectors and kd-tree is known to be inefficient in high dimensional spaces. Wong et al. [119] proposed secure k-NN computation that can determine which of two encrypted database entries has a smaller distance to the query, while keeping the actual distance value secret. This work can potentially be used for rank-ordered multimedia retrieval, but the efficiency is limited because each comparison only answers a binary question of which one among the two being larger or smaller. Erkin et al. [28], Sadeghi et al. [92] and Osadchy et al. [79] studied privacy preserving face recognition, where one party wants to verify the existence of a given face image in a database hosted on another party's servers. The two parties want to keep their own data secret from each other. Additive homomorphic encryption schemes are used to allow similarity computation in the encrypted domain. Similar techniques are also used by Jiang et al. [47] to identify the existence of similar text documents between two parties. As there have been no efficient homomorphic encryption schemes yet that allow both addition and multiplication, multiple rounds of communication between the two parties are required to compute the Euclidean distance between the query and each database entry.

There are several major differences between our current work and the above mentioned works [28, 47, 79, 92]: (1) we are considering rank-ordered search where the server needs to return the documents ranked according to their similar to the query, while existing secure computing work typically focus on a binary matching problem, such as biometric matching and keyword search, and the server may be made oblivious to the binary matching result; (2) in our secure search problem, the user owns all the data and the server merely offers storage and search functionality, while in secure multi-party computation scenario, both parties have their own private data that need to be kept secret from each other when computing a joint function; (3) we consider retrieval over large volumes of multimedia data using high dimensional visual features, which requires efficient processing to be practical; (4) existing work typically exploit homomorphic encryption and cryptographic protocols that involve high computation and communication cost, which can be formidably expensive for content-based retrieval over a large multimedia database, while our work seeks highly practical and efficient schemes without incurring heavy communication. (5) The application considered in this work is more consumer-oriented, which has less stringent requirement on security but demands highly efficient solutions and least user involvement; while the applications considered in secure computation literature typically involve very sensitive data such as biometrics, thus demanding very high security at the cost of heavier computation and communication cost.

Contributions and Chapter Organization

Given the different application settings and different requirements on security-efficiency trade-off, it is not feasible to employ techniques such as homomorphic encryption and cryptography protocols, which are arguably too heavy-weight for consumer applications. As such, we are trying to approach the problem from a practical perspective and explore what we can do now as possible solutions to help protect confidentiality of online personal data. By jointly exploiting areas of cryptography, image processing, and information retrieval, we propose efficient confidentiality preserving search techniques [66], without multiple rounds of communications between the user and the server. The search capability over encrypted multimedia database is achieved by designing proper scrambling or randomization schemes for visual features and search indexes generated from the multimedia data, while the multimedia data can be encrypted by any established cryptographic ciphers. We propose two types of confidentiality preserving search techniques: the first group of techniques focuses on visual feature protection that allows similarity comparison among scrambled features; while the second group of techniques aim at randomizing the search indexes directly, where the search indexes are typically generated from visual features and carefully designed to enable efficient search over large multimedia databases. The two groups of techniques are complementary and represent different trade-offs between user-side computational complexity and communication overhead.

We demonstrate the proposed techniques using image databases in this work, although these techniques are applicable to other multimedia modalities such as video. Our experiments show that privacy preserving retrieval can achieve comparable performance as traditional plaintext retrieval. The proposed schemes also demonstrate good efficiency and reveal as little information as possible to an honest but curious server. It should be noted that we are not designing highly secure encryption schemes, but we are exploring, from an interdisciplinary point of view, efficient algorithms that can be practical and offer certain amount of randomization to preserve data confidentiality. We also provide quantitative study on such a security-efficiency trade-off. Since a user's private photo collection may not be as sensitive as his/her biometric data, a highly efficient scheme with reduced security provides a reasonable practical solution for applications that do not require the highest level of protection or cannot afford the computation and communication cost required by traditional cryptographic schemes. To the best of our knowledge, this work is among the first endeavors on confidentiality-preserving content-based multimedia retrieval and can have promising applications in secure online multimedia management.

This chapter is organized as follows: Section 4.2 formulates the problem of confidentiality preserving multimedia search and discuss possible solutions. Sections 4.3 and 4.4 presents the proposed search schemes, based on feature protection and index randomization, respectively. Section 4.5 summarizes experimental results on retrieval over an encrypted color image database. Summary of the chapter is given in Section 4.6.

4.2 Problem Formulation

We now use image as an example modality to discuss problem formulation. In order to protect data privacy, images need to be encrypted before being transferred to the remote server. Image encryption can be done using state-of-the-art ciphers such as AES or RSA by treating images as ordinary data, or using image specific encryption techniques such as selective and format-compliant encryption [71], [41], [52] to enable post-processing such as transcoding of encrypted images. As these techniques are built upon established cryptographic encryption tools, it is computationally difficult for an adversary to decrypt the encrypted image files.

In modern image retrieval techniques, content similarity is typically evaluated using search indexes or visual features, such as color histograms and salient points, instead of comparing images pixel by pixel. Therefore, encryption of images alone is not sufficient for privacy preserving retrieval because search indexes or image features in plaintext may reveal information about image content. For example, a color histogram with large values for the blue components would indicate the presence of sky or ocean, and SIFT descriptors [61] may reveal information about distinctive objects in the image. In order to be able to search through the encrypted database without leaking information from the plaintext search indexes or image features, we devise schemes to generate and appropriately randomize image features or indexes on the user side using a secret key and then transfer them to the server, where the randomized features or indexes are used to evaluate image similarity by the server. A system model for the secure search scenario is shown in Figure 4.1, where the left part depicts the database construction stage and the right part depicts the retrieval stage. There are two entities in this model: a user who owns the private image collections, and a server who stores the encrypted data and performs retrieval based on a given encrypted query. During database construction, the user encrypts the images using standard ciphers and protects visual features or search indexes using the schemes proposed in this work. After encryption, the user sends the encrypted data to the server for storage. During retrieval, the user randomizes the visual feature or search index from the query image and sends the randomized index to the server, who performs retrieval using the randomized index to return similar images in their encrypted form. The block “Build search index” corresponds to randomizing the features in the feature protection schemes or building secure indexes in the secure indexing schemes, which is the focus of this work and will be described in Sections 4.3 and 4.4.

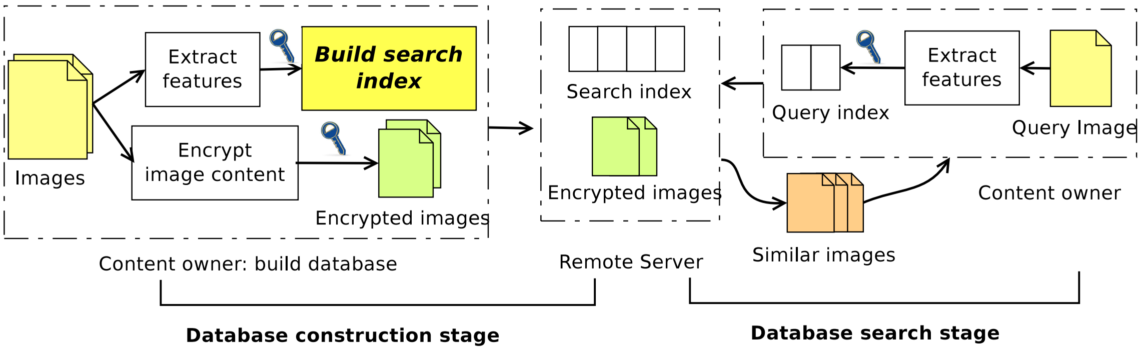


Figure 4.1: System model for secure image retrieval

The first approach for secure image retrieval is to randomize the feature vectors of each image and store those randomized features on the server, as described in Section 4.3. The server can use these randomized features as naïve indexes if the database is small, or the server can build efficient indexes upon the randomized features for improved search efficiency. Since the similarity of two images is typically measured by computing the distance between features extracted from them [23], the randomization of image features should approximately preserve their distances. Suppose we represent image features as vectors in $\mathbb{R}^n$, we seek a randomization function $\mathcal{E}(\cdot) : \mathbb{R}^n \to \mathbb{R}^m$, such that given two feature vectors $\mathbf{f}$ and $\mathbf{g}$, $d_{\mathcal{E}}(\mathcal{E}(\mathbf{f}), \mathcal{E}(\mathbf{g})) \approx c \cdot d(\mathbf{f}, \mathbf{g})$, where $d_{\mathcal{E}}(\cdot, \cdot)$ and $d(\cdot, \cdot)$ are some appropriate distance measures on the randomized and the raw features, respectively, and c is a constant scaling factor. The approximate distance preserving randomization scrambles the visual features for content protection and allows servers to perform similarity comparison in the encrypted domain.

Since efficiency and scalability are critical aspects for retrieval from a large database and rely on the design of search indexes, the second approach for secure image retrieval explores the possibility of randomizing the state-of-the-art multimedia search indexes without affecting their search capability. During retrieval, the content owner who knows the secret key can generate a properly randomized query index from the query image. The server then compares the randomized query index with the stored indexes and returns the encrypted files of the most similar images to the user for decryption and viewing. Without knowing the secret key used for randomization, it should be difficult to search the database or infer the database content. The randomized index also helps protect the privacy of the query image.

Secure image retrieval through feature and index randomization are closely related. Image features themselves can be considered as a special form of search index, where each image is represented by its feature vectors and the query image's feature is compared to all features in the database. On the other hand, modern indexing schemes are built upon image features and allow efficient retrieval by reducing the number of images that need to be compared. Since the randomized features preserve the capability of similarity comparison, they can be used to build efficient indexing schemes. The content owner has the flexibility either to provide the server with randomized features and let the server perform the time-consuming index generation, or to generate the secure index on his/her side to reduce the amount of information that needs to be sent to the server. Therefore, the two kinds of approaches represent different trade-offs between user-side computational complexity and communication overhead.

4.3 Visual Feature Protection

In this section, we propose two categories of secure retrieval schemes. As discussed in the previous section, most state-of-the-art content-based image retrieval techniques utilize low-level visual features to represent and compare image content, and these visual features can potentially reveal important information about the image content. We first discuss feature protection schemes that enable similarity comparison between features in the encrypted domain. The randomized features along with encrypted images can protect image content privacy against untrustworthy service providers and malicious intruders.

The ability to generate randomized indexes on the user side provides an alternative for secure retrieval with reduced communication overhead. In the second part of this section, we discuss the protection of search indexes by exploiting the visual words representation of images [77]. Visual words method hierarchically clusters features into a vocabulary tree, following which each image is indexed based on this vocabulary tree and represented as a bag of visual words. Experiments on object recognition in the recent literature [77, 83] show that visual words based representation can be scaled to large databases. We propose two secure indexing schemes based on inverted index [122] and min-hash [16]. These two schemes protect information about the image content from the adversary and at the same time achieve efficient and scalable search capability.

Three feature protection schemes are proposed in this work with different trade-offs among computational complexity, retrieval performance, and security.

4.3.1 Bitplane Randomization

The most significant bits (MSB) of an image capture important information about image appearance. The concept of processing bit-planes from MSBs to LSBs has been used in multimedia signal processing such as scalable encoding to provide fine granular trade-off between bit-rate and quality. Feature vectors with small distances are likely to have similar patterns among their MSB bit-planes. This motivates us to investigate the scrambling of feature vectors based on a secret key, such that the patterns in the MSB bit-planes of the feature vectors are preserved for similarity comparison, but without knowing the secret key, the bit-planes cannot be decrypted to reveal the image content.

Given a feature vector $\mathbf{f} = [f_1, \cdots, f_n] \in \mathbb{R}^n$, each component f_i is represented in its binary form as $[b_{i1}, \cdots, b_{il}]^T$, where b_{i1} is the first MSB, b_{il} is the least significant bit (LSB) or the lth MSB, and l is the total number of bit-planes. For example,

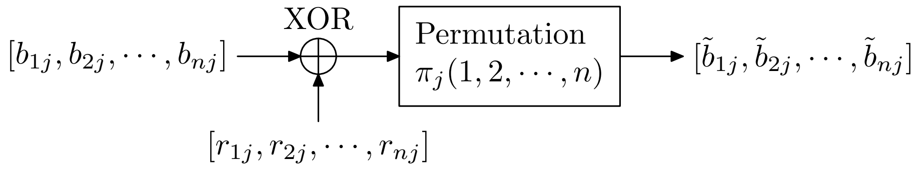

XOR
$[b_{1j}, b_{2j}, \cdots, b_{nj}]$
Permutation
$\pi_j(1, 2, \cdots, n)$
$[\tilde{b}_{1j}, \tilde{b}_{2j}, \cdots, \tilde{b}_{nj}]$
$[r_{1j}, r_{2j}, \cdots, r_{nj}]$

Figure 4.2: Randomization of the j^{th} bit-plane

the 8-bit representation of “148” is 10010100, so the 1st MSB to the 8th MSB of 148 are 1, 0, 0, 1, 0, 1, 0, 0, respectively, and the LSB is 0. The j^{th} bit-plane of $\mathbf{f}$ is composed of the j^{th} MSB of the n feature components, denoted as $[b_{1j}, b_{2j}, \cdots, b_{nj}]$. The Hamming distance between two bit-planes is preserved when each bit-plane is XORed with the same binary vector or when each is permuted using the same permutation pattern. We exploit this property to randomize the top k MSB bit-planes of the feature vectors while preserving the Hamming distances among randomized bit-planes.

The randomization of the j^{th} bit-plane of a given feature vector is illustrated in Fig. 4.2. The bits comprising the bit-plane are first XORed with a pseudorandom binary sequence, which essentially randomly flips the value of each bit. As a result, each bit in the resulting bit-plane will be equally likely to be 0 or 1 and the number of ‘1’s in the bit-plane will be approximately the same as the number of ‘0’s. Hiding the original number of ‘1’s in each bit-plane maximizes the entropy of the randomized bit-planes and thus improves security. The resulting bits are then randomly permuted to obtain the randomized bit-plane.

All the randomized bit-planes are reassembled to form the randomized feature vector $\mathcal{E}(\mathbf{f}) = [\tilde{f}_1, \cdots, \tilde{f}_n]$. Since the values $\{\tilde{f}_1, \cdots, \tilde{f}_n\}$ are completely random, traditional L_1 or L_2 norm does not capture the similarity between randomized features. Instead, we compute the distance between two randomized feature vectors $\mathcal{E}(\mathbf{f})$ and $\mathcal{E}(\mathbf{g})$ using a weighted sum of Hamming distances between their individual bit-planes:

$$d_{\mathcal{E}}(\mathcal{E}(\mathbf{f}), \mathcal{E}(\mathbf{g})) = \sum_{i=1}^{n} \sum_{j=1}^{l} |\tilde{b}_{ij}^{(\mathbf{f})} - \tilde{b}_{ij}^{(\mathbf{g})}| \times w(j). \tag{4.1}$$

Here, $\tilde{b}_{ij}$ is the i^{th} bit in the j^{th} randomized bit-plane, and $w(j)$s are the weights assigned to the bit-planes to reflect their unequal importance, which are chosen to be 2^{-j} in this work. Since using the same permutation and XOR pattern on corresponding bit-planes of two feature vectors preserves their Hamming distance, we have

$$d_{\mathcal{E}}(\mathcal{E}(\mathbf{f}), \mathcal{E}(\mathbf{g})) = \sum_{i=1}^{n} \sum_{j=1}^{l} |b_{ij}^{(\mathbf{f})} - b_{ij}^{(\mathbf{g})}| \times 2^{-j} \geq \sum_{i=1}^{n} \left| \sum_{j=1}^{l} (b_{ij}^{(\mathbf{f})} - b_{ij}^{(\mathbf{g})}) \times 2^{-j} \right| = \|\mathbf{f} - \mathbf{g}\|_1. \tag{4.2}$$

The distance $d_{\mathcal{E}}(\cdot, \cdot)$ between randomized features is an upper bound on the L_1 distance between the original features. The bound is mostly tight but occasionally large errors between $d_{\mathcal{E}}(\mathcal{E}(\mathbf{f}), \mathcal{E}(\mathbf{g}))$ and $\|\mathbf{f} - \mathbf{g}\|_1$ may arise as some feature vectors with small L_1 distance may have large distance under $d_{\mathcal{E}}(\cdot, \cdot)$. For example, $8 = (1000)_2$ and $7 = (0111)_2$ have L_1 distance 1 but $d_{\mathcal{E}}(8, 7) = 15$. Fortunately, such cases occur with a relatively low probability, and experimental results in Section 4.5 show that bit-plane randomization leads to only a slight reduction in retrieval accuracy, as a trade-off for security.

4.3.2 Random Projection

An alternative to treating the feature vector as separate bit-planes is to consider the vector as a whole and perform some random transformation that preserves the distance. Random projection accomplishes this goal based on the idea that close points in a high dimensional space will be mapped to close points in a low dimensional space with high probability. Due to this property, random projection has been used as a building block in locality sensitive hashing [39] for efficient search over large multimedia databases.

The idea of random projection is briefly described as follows. Given a feature vector $\mathbf{f} \in \mathbb{R}^n$, we generate a key-dependent Gaussian random matrix $\mathbf{R} \in \mathbb{R}^{m \times n}$ with independent standard Gaussian components. The randomized function is defined as

$$\mathcal{E}(\mathbf{f}) = \mathbf{R} \cdot \mathbf{f} = [\mathbf{r}_1 \cdot \mathbf{f}, \cdots, \mathbf{r}_m \cdot \mathbf{f}] \in \mathbb{R}^m, \tag{4.3}$$

where $\mathbf{r}_i \cdot \mathbf{f}$ is the dot product between the i^{th} row of $\mathbf{R}$ and $\mathbf{f}$. The distance preserving property of random projection can be derived by considering the L_1 distance of randomized features, i.e., $d_{\mathcal{E}}(\mathcal{E}(\mathbf{f}), \mathcal{E}(\mathbf{g})) = \|\mathcal{E}(\mathbf{f}) - \mathcal{E}(\mathbf{g})\|_1$. Using equation (4.3), we have

$$\begin{aligned} d_{\mathcal{E}}(\mathcal{E}(\mathbf{f}), \mathcal{E}(\mathbf{g})) &= \sum_{i=1}^{m} |\mathbf{r}_i \cdot \mathbf{f} - \mathbf{r}_i \cdot \mathbf{g}| = \sum_{i=1}^{m} |\mathbf{r}_i \cdot (\mathbf{f} - \mathbf{g})| = \sum_{i=1}^{m} \|\mathbf{r}_i\|_2 \cdot \|\mathbf{f} - \mathbf{g}\|_2 \cdot |\cos(\theta_i)| \\ &= \|\mathbf{f} - \mathbf{g}\|_2 \cdot \sum_{i=1}^{m} \|\mathbf{r}_i\|_2 \cdot |\cos(\theta_i)| \approx c \cdot \|\mathbf{f} - \mathbf{g}\|_2 \end{aligned} \tag{4.4}$$

Here, θ_i is an independent and identically distributed random variable representing the angle between the vector $\mathbf{f} - \mathbf{g}$ and the random vector $\mathbf{r}_i$. By the law of large numbers, $\|\mathbf{r}_i\|_2 \approx$ const and $\sum_{i=1}^{m}|\cos(\theta_i)| \approx$ const. Thus, the distance $d_{\mathcal{E}}(\cdot,\cdot)$ between randomized features is proportional to the L_2 distance between the original feature vectors with high probability. By increasing the dimension m of the projected feature vector, the error $|d_{\mathcal{E}}(\mathcal{E}(\mathbf{f}),\mathcal{E}(\mathbf{g})) - c \cdot \|\mathbf{f}-\mathbf{g}\|_2|$ can be made arbitrarily small, leading to better approximation of the original L_2 distance. The projection dimension m controls the trade-off between retrieval performance and storage, as will be shown by the experimental results in Section 4.5.

In image retrieval literature, L_1 norm is also widely used and is shown to achieve slightly superior performance over L_2 norm in retrieval based on color histogram [39]. Random projection can also be used to preserve the L_1 distance between the original feature vectors when the projection is performed on the square-root of the feature vector,

$$\mathcal{E}(\mathbf{f}) = \mathbf{R} \cdot \sqrt{\mathbf{f}}, \text{ where } \sqrt{\mathbf{f}} = \left[\sqrt{f_1}, \cdots, \sqrt{f_n}\right]. \tag{4.5}$$

To prove that the randomization in (4.5) preserves L_1 distance, we introduce the concept of unary encoding of an integer vector. Given $\mathbf{f} = [f_1, \cdots, f_n]$, its unary encoding $\mathcal{U}(\mathbf{f})$ is defined as follows:

$$\mathcal{U}(\mathbf{f}) = [\mathcal{U}(f_1), \mathcal{U}(f_2), \cdots, \mathcal{U}(f_n)], \text{ where } \mathcal{U}(f_i) = \underbrace{11\cdots11}_{f_i}\underbrace{00\cdots00}_{M-f_i}. \tag{4.6}$$

Here M is the maximum possible value for any component of $\mathbf{f}$. Since the L_1 and L_2 norms for a binary vector are the same, we can perform random projection on $\mathcal{U}(\mathbf{f})$ so that

$$\|\mathbf{R} \cdot \mathcal{U}(\mathbf{f}) - \mathbf{R} \cdot \mathcal{U}(\mathbf{g})\|_1 \approx c \cdot \|\mathcal{U}(\mathbf{f}) - \mathcal{U}(\mathbf{g})\|_2 = c \cdot \|\mathbf{f} - \mathbf{g}\|_1. \tag{4.7}$$

Note that the projection of $\mathcal{U}(\mathbf{f})$ onto a vector of standard Gaussian random variables results in a Gaussian random variable with variance $\sum_{i=1}^{m} f_i$. This is equivalent to the projection of $\sqrt{\mathbf{f}}$ onto a vector of standard Gaussian random variables.

The security of random projection based scheme is due to the fact that without knowing the secret key and therefore the projection matrix $\mathbf{R}$, it is extremely difficult to reconstruct the exact original features from the projected ones. For $m < n$, $\mathbf{y} = \mathbf{R} \cdot \mathbf{x} \in \mathbb{R}^m$ is an under-determined equation and there are infinitely many $\mathbf{x}$ that can give the same output $\mathbf{y}$. For $m \geq n$, the equation $\mathbf{y} = \mathbf{R} \cdot \mathbf{x}$ can be solved by using pseudo-inverse, but a different choice of $\mathbf{R}$ will give a different $\mathbf{x}$. Therefore, without knowing $\mathbf{R}$, it is extremely difficult to obtain the exact $\mathbf{x}$ by knowing $\mathbf{y}$.

4.3.3 Randomized Unary Encoding

Key-dependent random projection is an efficient algorithm for feature protection and preserves the distance between feature vectors with high probability. However, since the projection is a linear operation, using a reasonable amount of known plaintext features and their randomized versions, an attacker can obtain the projection matrix. As will be shown by the security analysis in the next chapter, this poses security threats in the known plaintext attack model (KPA), where the attacker has access to a set of plaintext and ciphertext pairs. This motivates us to add an additional layer of security by introducing non-linear operations into the feature randomization.

Given a feature vector $\mathbf{f} = [f_1, \cdots, f_n]$, we perform unary encoding $\mathcal{U}(\mathbf{f}) =$

$[\mathcal{U}(f_1), \mathcal{U}(f_2), \cdots, \mathcal{U}(f_n)]$. The non-linearity of the randomization is achieved by performing XOR of $\mathcal{U}(\mathbf{f})$ with a vector of binary random variables $\mathbf{r}$ and then randomly permuting the resulting binary vector. As discussed in Section 4.3.1, XOR and random permutation preserve the Hamming distance among $\mathcal{U}(\mathbf{f}), \forall \mathbf{f}$, which also equals the L_1 distance between original feature vectors. Denoting the randomization by XOR and permutation as $\mathcal{E}_1(\cdot)$, we have $\|\mathcal{E}_1(\mathcal{U}(\mathbf{f})) - \mathcal{E}_1(\mathcal{U}(\mathbf{g}))\|_2 = \|\mathbf{f} - \mathbf{g}\|_1$. By using efficient shuffling algorithms, $\mathcal{E}_1(\cdot)$ takes $O(nM)$ time, where M is the maximum possible value for any component of $\mathbf{f}$. One disadvantage of using unary encoding is the storage increase from $O(n \log M)$ bits to $O(nM)$ bits. To reduce storage, we further apply random projection on $\mathcal{E}_1(\mathcal{U}(\mathbf{f}))$, which also plays an important role in enhancing the security of the scheme, as will be shown in Section 4.5.

Denote the randomization by XOR and permutation as $\mathcal{E}_1(\cdot)$ and random projection as $\mathcal{E}_2(\cdot)$. The overall randomization function $\mathcal{E}(\cdot)$ is now $\mathcal{E}_1(\cdot)$ followed by $\mathcal{E}_2(\cdot)$, and can be written as $\mathcal{E}(\mathbf{f}) = \mathcal{E}_2(\mathcal{E}_1(\mathcal{U}(\mathbf{f}))) \in \mathbb{R}^m$. Considering L_1 distance of randomized features, we have the approximate distance preserving property

$$d_{\mathcal{E}}(\mathcal{E}(\mathbf{f}), \mathcal{E}(\mathbf{g})) \approx c \cdot \|\mathcal{E}_1(\mathcal{U}(\mathbf{f})) - \mathcal{E}_1(\mathcal{U}(\mathbf{g}))\|_2 = c \cdot \|\mathbf{f} - \mathbf{g}\|_1. \tag{4.8}$$

The randomized unary encoding scheme effectively preserves the L_1 distance of original feature vectors with high probability and provides enhanced security, as will be shown in Chapter 5.

4.4 Secure Search Indexes

Once the image features are randomized using the above methods, they can be stored onto the server and provide search capability in the encrypted domain without revealing information about the database content. However, since the image features are usually high dimensional vectors, comparing every pair of such vectors is computationally prohibitive for a large database. Modern image retrieval techniques often achieve efficiency and scalability through well-designed search indexes. In the following, we propose two secure indexing schemes, – secure inverted index and secure min-hash, which can retain the efficient search capability of the plaintext search indexes.

4.4.1 Secure Inverted Index

Inverted index [122] is a widely used indexing structure in text document retrieval, where each keyword has an associated inverted index listing the documents that contain this keyword and the number of occurrences of this word in each of these documents. Only those documents that appear in the query word's inverted index need to be considered during retrieval. By utilizing the visual words representation of images [77], inverted index can be constructed for image documents and facilitates efficient search and retrieval over large image databases.

In order to protect the privacy of query image and minimize the amount of database information leaked to the server during the search process, inverted indexes should be generated and protected on the user side before being transferred to the server. In order to generate inverted index, a vocabulary tree is first created, where each node in the tree denotes a representative feature vector and each leaf node represents a visual word. Generating a vocabulary tree requires large set of training images and computationally intensive hierarchical clustering. Therefore, we assume that the vocabulary tree will be generated by the service provider, who usually has large computational resources, and is then provided to each user. To build secure search indexes from the vocabulary tree, the content owner extracts the visual features from each image, assigns each feature to the closest visual word in the vocabulary tree, and finally updates and randomizes the inverted indexes for those visual words. This procedure of index generation is illustrated in Figure 4.3.

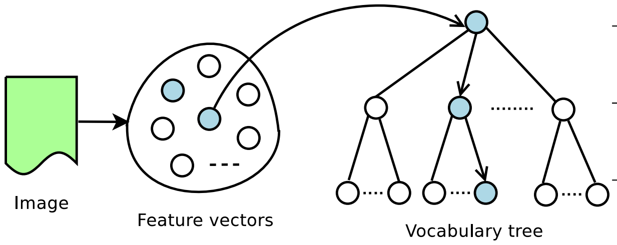


Figure 4.3: Inverted index generation by content owner

Word ID	i			
Image ID	I_1	I_2	$\cdots$	I_{N_i}
Word frequency	w_1	w_2	$\cdots$	w_{N_i}

Figure 4.4: Data structure of inverted index

Consider a total of N visual words and suppose N_i images contain the i^{th} visual word. Figure 4.4 shows the content of the inverted index of the i^{th} visual word, where w_j is the number of times the i^{th} word appears in image I_j. Given any query image Q and database image D, their bags of visual words are denoted as $\{Q_1, Q_2, \cdots, Q_N\}$ and $\{D_1, D_2, \cdots, D_N\}$, respectively. Here Q_i and D_i are the number of times the i^{th} word appears in the query and the database image, respectively. Normalization is typically applied so that $\sum_{i=1}^{N} Q_i = \sum_{i=1}^{N} D_i$ for all database images. After normalization, Q_i and D_i can take non-integer values and represent the relative frequency of occurrence of the i^{th} word. In the conventional non-secure setting, $\{D_1, D_2, \cdots, D_N\}$ is used to update the inverted indexes during index generation and $\{Q_1, Q_2, \cdots, Q_N\}$ is used to search the database for similar images.

Randomization of Inverted Index: Given that the service provider typically creates and thus has the knowledge of the vocabulary tree, inverted indexes in their plaintext form can potentially reveal information about the visual content of the images. We protect the inverted index by first performing a random permutation $\tau(\cdot)$ on the word IDs so that the i^{th} word will now have an ID $\tau(i)$. Computing random permutation takes $O(N)$ time and needs to be done only once on the user side. However, the server needs to guess the correct IDs from $O(N!)$ possibilities, which is computationally infeasible given the typically large value of N.

Scrambling word IDs alone is not secure, because the server can still use visual word frequencies to identify the words that appear more frequently. An example is given in Figure 4.5, showing the distribution of word frequencies for local color histograms extracted from the Corel image dataset of 1000 images. This statistical information can be exploited to identify many words and makes the random permutation less secure. We apply order preserving encryption (OPE) [3] to alleviate this problem. OPE has the property that for two values x and y, if $x < y$, after encryption $\mathcal{E}(\cdot)$, the order relation is preserved so that $\mathcal{E}(x) < \mathcal{E}(y)$. By applying OPE on the word frequency values, we can make the distribution of encrypted frequency values close to a uniform distribution in order to reduce the amount of information leaked to the server. At the same time, the preservation of the order information ensures that image similarity can still be compared in the encrypted domain.

To perform order preserving encryption, we map each frequency value w to an integer uniformly selected from an interval $[l_w, u_w]$. The length of each such interval is chosen by the content owner to be proportional to the number of times that the value w occurs in all inverted indexes. Note that inverted index of a particular word only stores information about images that contain this word, therefore only positive w will be considered. Intervals for different word frequency values are non-overlapping and order preserving, i.e., for $w < v$, their corresponding intervals $[l_w, u_w]$ and $[l_v, u_v]$ satisfy $u_w < l_v$, while for $w = v$, they will be mapped to two values randomly chosen from the same interval $[l_w, u_w]$. These intervals form a partition of a large overall interval, and we use [0,7800] as the overall interval in our experiments. Figure 4.6 shows that after order preserving encryption, the distribution of word frequency values is closer to a uniform distribution over the large overall interval.

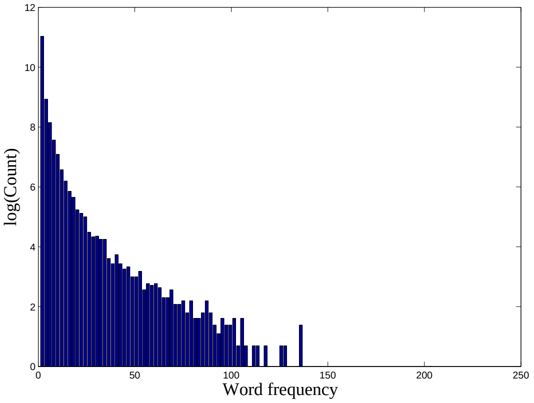


Figure 4.5: Histogram of word frequencies before OPE

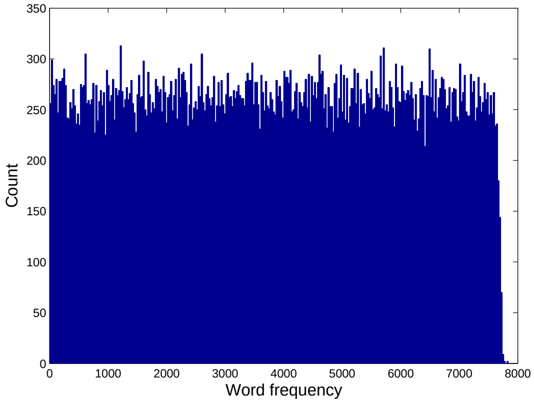


Figure 4.6: Histogram of word frequencies after OPE

***Retrieval using Randomized Index**:* After randomization, the visual words representations of the query image and an image in the database are denoted by $\{\mathcal{E}(Q_1), \cdots, \mathcal{E}(Q_N)\}$ and $\{\mathcal{E}(D_1), \cdots, \mathcal{E}(D_N)\}$, respectively, where $\mathcal{E}(\cdot)$ represents the order preserving encryption. Since visual words that are common in many images carry little discriminative information, we weigh the OPE encrypted version of each frequency value $\mathcal{E}(Q_i)$ and $\mathcal{E}(D_i)$ by its inverse document frequency (IDF) [93]. IDF is defined as $\text{IDF} = \log(\frac{M}{N_i})$, where M is the total number of images in the database and N_i is the number of images containing the word i. Commonly occurring visual words will have low IDF and receive small weights in similarity comparison. After encryption and weighting, we represent the query image and database image as

$$\begin{aligned} Q_{OPE} &= \{\tilde{Q}_1, \tilde{Q}_2, \cdots, \tilde{Q}_N\}, \text{ where } \tilde{Q}_i = \mathcal{E}(Q_i) \log\left(\frac{M}{N_i}\right), & (4.9) \\ D_{OPE} &= \{\tilde{D}_1, \tilde{D}_2, \cdots, \tilde{D}_N\}, \text{ where } \tilde{D}_i = \mathcal{E}(D_i) \log\left(\frac{M}{N_i}\right). & (4.10) \end{aligned}$$

The similarity of two images Q_{OPE} and D_{OPE} after OPE is measured by the Jaccard similarity between $\{\tilde{Q}_1, \tilde{Q}_2, \cdots, \tilde{Q}_N\}$ and $\{\tilde{D}_1, \tilde{D}_2, \cdots, \tilde{D}_N\}$:

$$\text{Sim}(Q_{OPE}, D_{OPE}) \triangleq \frac{|Q_{OPE} \cap D_{OPE}|}{|Q_{OPE} \cup D_{OPE}|} \triangleq \frac{\sum_{i=1}^{N} \min(\tilde{Q}_i, \tilde{D}_i)}{\sum_{i=1}^{N} \max(\tilde{Q}_i, \tilde{D}_i)}. \qquad (4.11)$$

The Jaccard similarity measures the similarity between two sets and has been used for near duplicate detection of text and image documents [15,20]. The set operations $\cap$ and $\cup$ are extended in Equation (11) to measure the similarity between two sets of word frequency values. The functions $\min(\cdot, \cdot)$ and $\max(\cdot, \cdot)$ return the minimum and maximum value of the two input arguments, respectively. As the order information used in $\min(\cdot,\cdot)$ and $\max(\cdot,\cdot)$ is preserved by the order preserving encryption, the Jaccard similarity computed from the encrypted sets reflects the similarity of the plaintext sets, thus allowing similarity comparison in the encrypted domain.

As order preserving encryption preserves the order among encrypted word frequency values, some information about the randomized index may be revealed. The other limitation of using OPE on the inverted index is that the length of intervals used in OPE is determined by the distribution of word frequency values and this distribution can change when many more images are added to or deleted from the database. For example, if one word frequency value appears much more often in the newly added set of images, the corresponding OPE interval will have higher probability in the word frequency distribution than other intervals. Such a change in the distribution may reveal some of the interval ranges used in OPE and make OPE less secure. As the storage size of image indexes is typically much smaller than that of the images, this security problem with dynamic database changes can be alleviated by periodically downloading the indexes from the server to the user side, decrypting them, and encrypting them again using the new distribution information.

4.4.2 Secure Min-Hash

The min-Hash algorithm, first proposed by Broder et al. [16], provides another efficient way to compute the Jaccard similarity between the visual words representations of two images. The min-Hash algorithm was originally developed for near duplicate detection of text documents [15]; extensions to near duplicate detection of images have been proposed recently by applying min-Hash to visual words representations [20,21]. Here, we focus on the security aspect of the min-Hash algorithm and use it for secure ranking of image similarity.

The basic idea of the min-Hash algorithm is as follows: For any given set $\mathcal{A}$ such as the visual words representation, its min-Hash is defined as $m(\mathcal{A}, f) = \arg\min_{x \in \mathcal{A}} f(x)$, where f is a randomized hash function[1] with the property that $\Pr[f(x) < f(y)] = \Pr[f(x) > f(y)] = 0.5$, $\forall x, y \in \mathcal{A}$ and $x \neq y$. The probability that two sets have the same min-Hash value is given by their Jaccard similarity defined in Equation (4.11).

To compare the similarity between a given query image and an image in the database, we use their visual words representations:

$$\begin{aligned} Q_{MH} &= \{\hat{Q}_1, \hat{Q}_2, \cdots, \hat{Q}_N\}, \text{ with } \hat{Q}_i = Q_i \log\left(\frac{M}{N_i}\right), && (4.12) \\ D_{MH} &= \{\hat{D}_1, \hat{D}_2, \cdots, \hat{D}_N\}, \text{ with } \hat{D}_i = D_i \log\left(\frac{M}{N_i}\right), && (4.13) \end{aligned}$$

where Q_i and D_i are the number of times the i^{th} visual word appears in the query and the database image, respectively. Non-zero components in Q_{MH} and D_{MH} suggest the existence of the corresponding visual word and represents the number of occurrence scaled by the inverse document frequency. In order to apply min-Hash to measure the Jaccard similarity between the sets Q_{MH} and D_{MH}, we follow the

[1] The hash function used here is different from a cryptographically secure hash function in that it does not need to be strongly collision free.

method of Chum et al. [21] and represent Q_{MH} and D_{MH} as the following sets:

$$\begin{aligned} \mathcal{A}(Q_{MH}) &= \{X_1^1, \cdots, X_1^{\hat{Q}_1}, X_2^1, \cdots, X_2^{\hat{Q}_2}, \cdots, X_N^1, \cdots, X_N^{\hat{Q}_N}\}, && (4.14) \\ \mathcal{A}(D_{MH}) &= \{X_1^1, \cdots, X_1^{\hat{D}_1}, X_2^1, \cdots, X_2^{\hat{D}_2}, \cdots, X_N^1, \cdots, X_N^{\hat{D}_N}\}. && (4.15) \end{aligned}$$

Here, X_i^j is a unique element indexed by i and j. The min-Hash values generated from $\mathcal{A}(Q_{MH})$ and $\mathcal{A}(D_{MH})$ are essentially elements randomly selected from the two sets, and they satisfy

$$\Pr[m(\mathcal{A}(Q_{MH}), f) = m(\mathcal{A}(D_{MH}), f)] = \mathrm{Sim}(Q_{MH}, D_{MH}) = \frac{\sum_{i=1}^{N} \min(\hat{Q}_i, \hat{D}_i)}{\sum_{i=1}^{N} \max(\hat{Q}_i, \hat{D}_i)}. \tag{4.16}$$

In order to obtain a reliable estimate of $\mathrm{Sim}(Q_{MH}, D_{MH})$, k independent hash functions $f_1, f_2, \cdots, f_k$ are used to generate k min-Hash values for $\mathcal{A}(Q_{MH})$ and $\mathcal{A}(D_{MH})$, respectively. The concatenation of the k min-Hash values for $\mathcal{A}(Q_{MH})$ forms a *sketch* of the image Q_{MH}, and a sketch of the image D_{MH} is formed similarly. The number of identical values in their sketches, denoted by $s(Q_{MH}, D_{MH}) = |\{i : 1 \leq i \leq k | m_i(Q_{MH}) = m_i(D_{MH})\}|$, follows a binomial distribution

$$\Pr[s(Q_{MH}, D_{MH}) = l] = \binom{k}{l} [\mathrm{Sim}(Q_{MH}, D_{MH})]^l [1 - \mathrm{Sim}(Q_{MH}, D_{MH})]^{k-l}.$$

Thus, the maximum likelihood estimate for the similarity of two images $\mathrm{Sim}(Q_{MH}, D_{MH})$ is the fraction of identical values in their sketches, $s(Q_{MH}, D_{MH})/k$.

To implement the randomized hash function, we use the input value as part of the seed to a pseudo random number generator and map the input value to a random number between $(0, 1)$ as the output. The min-Hash $m(\mathcal{A}, f) = X_{i_0}^{j_0} = \mathrm{argmin}_{i,j} f(X_i^j)\ \forall X_i^j \in \mathcal{A}$. In our implementation, $X_{i_0}^{j_0}$ is the output of a trapdoor function $g(i_0, j_0)$ uniquely determined by i_0 and j_0 so that it is easy to compute in one direction to obtain $g(i_0, j_0)$ given i_0 and j_0, but it is computationally difficult to compute in the opposite direction, i.e., to determine i_0 and j_0 given $g(i_0, j_0)$. Therefore, the original word frequency information, as captured by the parameter j_0, can be protected from the adversary who has knowledge of only the min-hash values. The addition of trapdoor function is different from the normal min-hash and ensures that the original feature information is securely protected. The trapdoor function can be implemented through a trapdoor permutation function [51].

During index generation, the content owner creates min-Hash sketch for every image using a secret key and stores these sketches on the remote server. During retrieval, the query image is processed similarly by the content owner, who has the secret key to generate its min-Hash sketch. This sketch is then sent to the server for comparison with the sketches of the database images. Similarity between two images is computed as the percentage of identical values in their min-Hash sketches. Retrieval efficiency can be further improved by organizing similar sketches into the same bucket of another hash table [100] and comparing only to sketches with similarity higher than a certain threshold.

4.5 Experimental Results

Two desirable properties of a secure image retrieval scheme are good retrieval performance that is comparable to state-of-the-art plaintext retrieval schemes and provable security so that content privacy is protected against adversaries. In this section, we demonstrate the retrieval performance of the proposed secure search schemes, and in the next section, we analyze the security of these schemes under different attack models.

4.5.1 Experiment Setup

We perform search and retrieval experiments on an image database containing 1000 color images from the Corel dataset [1]. These images are grouped by content into 10 categories, with 100 images in each category: African, Beach, Architecture, Buses, Dinosaurs, Elephants, Flowers, Horses, Mountain, and Food. Image sizes are either 256 × 384 or 384 × 256. This database has been used as ground-truth for evaluating color image retrieval [46] and image annotation [18]. Sample images from the database are shown in Figure 4.7.

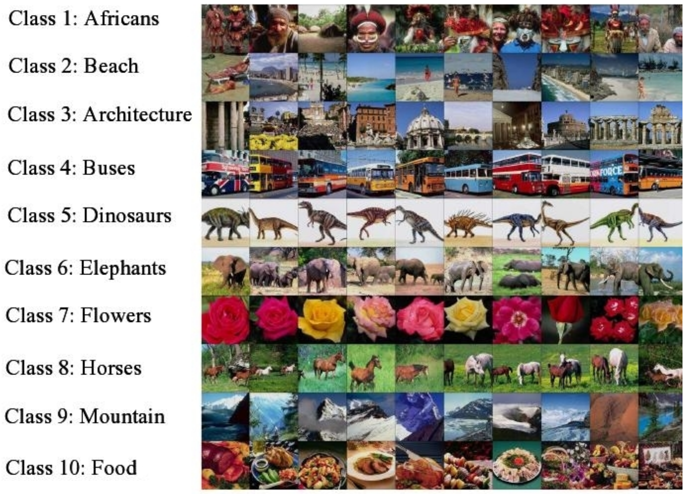


Figure 4.7: Selected content of the Corel dataset (Figure from [46])

We use global color histogram for the feature randomization based schemes and localized color histogram for the index randomization based schemes. The color histograms are in the HSV color space. For localized color histogram, we divide an image into 256 blocks and extract a 128-dimensional color histogram from each block by quantizing the three channels of hue, saturation, and intensity value into 8, 4, and 4 levels, respectively, where finer quantization is allocated to hue as suggested by Jeong et al. [46]. For feature randomization based schemes, we have one histogram for each image, while for index randomization based schemes, we obtain a training set of 256,000 histograms from the entire database and perform hierarchical clustering to build the vocabulary tree. During clustering, we use L_1 norm to measure the distance between color histograms and take the average of each cluster as its representative feature. Each node in the vocabulary tree except the leaf nodes has 10 children and the tree has height 3, which gives 10^3 visual words.

During search and retrieval, images in the database are returned in the descending order of their similarity to the query. Retrieval performance is evaluated using precision-recall curves, where precision and recall are defined as

$$\begin{aligned} precision &= \frac{\#\text{ of positive images among returned images}}{\#\text{ of returned images}}, \\ recall &= \frac{\#\text{ of positive images among returned images}}{\#\text{ of positive images in the database}}. \end{aligned}$$

A higher precision value at a given recall value indicates better retrieval performance. Our experiments use every image in the database as a query, and positive images are those images in the same category as the query.

For comparison with prior art on color histogram based image retrieval, we choose Jeong et al.'s work [46], where different settings for image retrieval using color histograms are compared and the best retrieval performance is achieved by comparing image similarity using the intersection of global color histograms in the HSV space. Given two color histograms H_1 and H_2 in the d-dimensional space, their intersection $I(H_1, H_2)$ is defined as

$$I(H_1, H_2) = \frac{\sum_{i=1}^{d} \min[H_1(i), H_2(i)]}{\min[\sum_{i=1}^{d} H_1(i), \sum_{i=1}^{d} H_2(i)]}.$$

Images with higher intersection values are considered more similar. During retrieval, the color histogram of the query image is compared with every histogram in the database and images with higher similarity are returned. When the L_1 norms of the histograms are the same, retrieval based on histogram intersection is equivalent to retrieval by L_1 distance of the histograms.

4.5.2 Retrieval Results based on Randomized Features

In contrast to the conventional retrieval scheme that uses plaintext color histograms as features for similarity comparison, we use the randomized versions of the same features in the secure retrieval scheme. Recall that the distance defined in Section 4.3.1 between features after bit-plane randomization is an upper bound on the original L_1 distance between features, while random projection and randomized unary encoding preserves the original L_1 distance with high probability. Thus, we would expect our secure retrieval schemes based on randomized features to have performance comparable to conventional schemes. The retrieval performance of the three feature protection schemes are illustrated in Fig. 4.8.

For comparison, we show the retrieval performance using histogram intersection and L_2 distance between plaintext histograms as the top and bottom curves

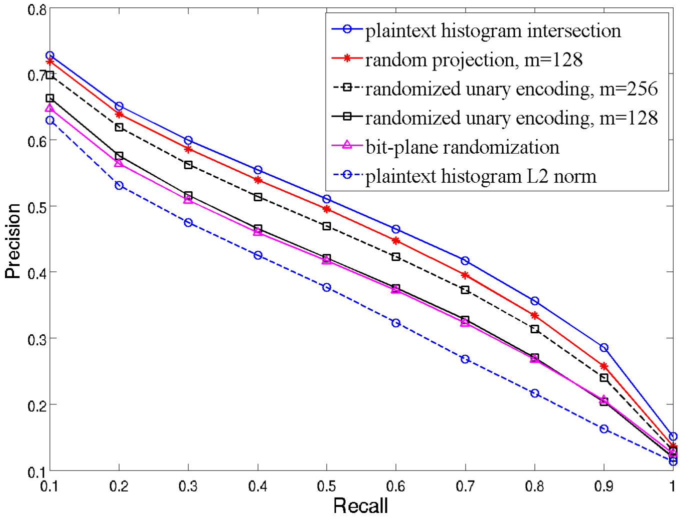


Figure 4.8: Retrieval performance based on randomized features

in Fig. 4.8, respectively. For each of the precision-recall curves in Fig. 4.8, a higher precision value at a given recall value indicates better retrieval performance. We see that the retrieval performance based on randomized features is better than plaintext histogram based on L_2 norm and is close to plaintext histogram intersection. This is expected because the original L_1 distance between color histograms is approximately preserved. By searching over randomized features, we only need to retrieve about $1\% - 9\%$ more images to obtain the same number of relevant images as in plaintext search. This shows that secure retrieval can be achieved by slightly trading off retrieval accuracy.

Among the three feature protection schemes, we observe a trade-off among retrieval performance, storage, and computational complexity. Bit-plane randomization has the largest gap to the plaintext intersection method among the three schemes. This is because the distance between features after bit-plane randomization is an upper bound to the original L_1 distance and there is some discrepancy between the distances for certain cases, as discussed in Section 4.3. However, bit-plane randomization has the lowest complexity $O(kn)$ as compared to $O(mn)$ of random projection and $O(mnM)$ of randomized unary encoding, where k is the number of bit-planes to randomize, m is the dimension of the projected features, and M is the largest value of the feature vector. Random projection and randomized unary encoding preserve L_1 norm with high probability, so their performance can be made arbitrarily close to plaintext scheme by increasing the projection dimension m. By doubling the projection dimension m from 128 to 256, the gap between the curves of plaintext and randomized unary encoding can be reduced by half, and the performance of random projection can be made almost the same as plaintext search (random projection with $m = 256$ is not shown in the figure). With the same m, random projection outperforms the randomized unary encoding because the latter projects the much longer unary encoded version than the original feature. M in randomized unary encoding can be quantized to a much smaller value to reduce complexity. In this work, we quantize M from 98304 to 128 with no loss in retrieval performance. The higher complexity of randomized unary encoding is a trade-off for better security, which will be analyzed in Chapter 5. Compared to traditional non-secure retrieval scheme, the additional step in our schemes is to randomize the features using pseudo random permutations or random projection, which are computationally efficient and take less than 1 second per image on a dual-core 3.0GHz PC with 4GB RAM in our experiments.

4.5.3 Retrieval Results based on Secure Indexes

The two secure indexing schemes are based on the visual words representation of the image. To establish the baseline retrieval performance of the visual words representation, we first demonstrate retrieval using the inverted index without any randomization and compare that with the performance of plaintext histogram intersection. In visual words representation, local color histograms are extracted from blocks of the image. By utilizing the vocabulary tree, each image is then represented as a bag of visual words $Q = \{\hat{Q}_1, \cdots, \hat{Q}_N\}$. Here, $\hat{Q}_i$ takes the form $\hat{Q}_i = Q_i \log\left(\frac{M}{N_i}\right)$, as shown in equations (4.12) and (4.13), where Q_i is the term frequency value and $\log\left(\frac{M}{N_i}\right)$ is the inverse document frequency weighting. In Fig. 4.9, we can see that plaintext histogram intersection and inverted index with term frequency-inverse document frequency (TF-IDF) weighting achieve very similar performance.

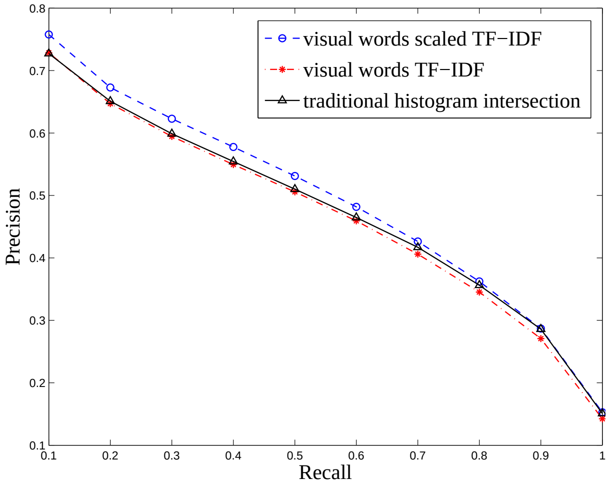


Figure 4.9: Baseline retrieval performance of visual words representation

Considering that w occurrences of a word may not necessarily carry w times the significance of a single occurrence, we apply the following scaled TF-IDF weighting,

$$\hat{Q}_i = \begin{cases} (1 + \log(Q_i)) \log(M/N_i), \text{ if } Q_i \neq 0, \\ \\ 0, \text{ if } Q_i = 0, \end{cases} \tag{4.17}$$

and find that the inverted index using visual words representation outperforms the histogram intersection by about 3% in precision. The comparison in Figure 4.9 shows that visual words representation can be used for rank-ordered retrieval of color images, while its success for object recognition using SIFT [61] features has been reported in [77, 83].

In the secure indexing scheme based on inverted index, the inverted indexes are randomized by order preserving encryption and random permutation of word IDs. We perform the same retrieval experiment using randomized inverted indexes and compare in Figure 4.10 its precision-recall curve with that of the baseline inverted index without any randomization. We can see that randomization of the index has very little impact on the retrieval performance, and the precision-recall curves before OPE and after OPE are very close to each other. This can be attributed to the use of Jaccard similarity, which is approximately preserved after the order preserving encryption. Compared to the conventional non-secure setting, generating randomized indexes imposes additional computational cost on the content owner, but this cost is small. When performed on a dual-core 3.0GHz PC with 4GB RAM, the tasks of extracting features, creating visual words representation, and randomizing inverted indexes can be done within 2 seconds per image, and search and retrieval over the entire database of 1000 images takes less than 1 second. The use of inverted index ensures that retrieval can be efficiently scaled to larger databases.

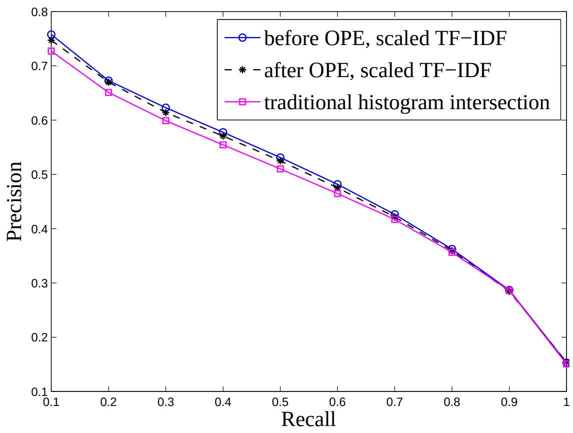


Figure 4.10: Retrieval performance of OPE

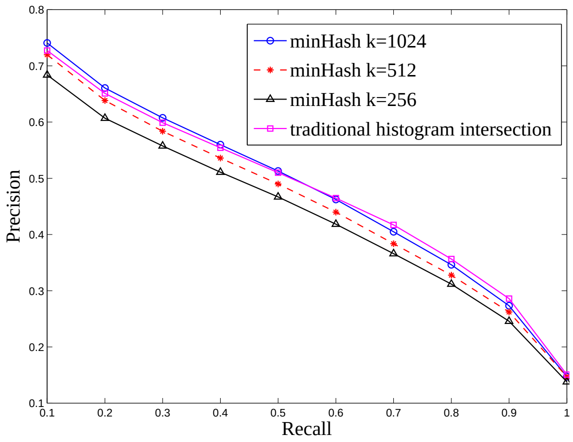


Figure 4.11: Retrieval performance of min-Hash

In the secure indexing scheme based on the min-Hash algorithm, each image is represented by a sketch $\{m_1, m_2, \cdots, m_k\}$, where m_i is the min-Hash value generated by the i^{th} randomized hash function. Images are returned in the descending order of their similarity to the query, measured by the percentage of identical values between their min-Hash sketches. Retrieval performance using min-Hash sketches is shown in Figure 4.11, where we can see that using min-Hash sketches gives a retrieval performance comparable to those of the histogram intersection method and the inverted index scheme. This is expected because the number of identical values in two min-Hash sketches preserves the Jaccard similarity with high probability. As the length of the sketch k increases, the estimate for image similarity based on the percentage of identical values in two min-Hash sketches becomes more accurate, leading to better precision-recall curves. A sketch length of 1024 gives performances similar to that of the inverted index scheme. Min-Hash sketches can be computed very efficiently on the user side, taking less than 1 second per image on a Dual-Core 3.0GHz PC with 4GB RAM. During retrieval, we compare sketches of all the images in the database in order to obtain the precision-recall curve. In practice, typically only the most similar images are of interest, so additional hash tables can be constructed for those sketches to further improve retrieval efficiency.

4.6 Chapter Summary

In this chapter, we studied the problem of confidentiality-preserving content-based search of images. The application that we are considering is secure online services that help manage personal multimedia collections. Such applications typically do not have very high security requirement but demands good efficiency and least user involvement. We address the problem from a joint signal processing and cryptography point of view and explore possible efficient solutions. We proposed two complementary approaches: one is to scramble visual features of multimedia documents and allow similarity comparison of the randomized features; and the other is to randomize state-of-the-art search indexes without affecting their search capability. Scalability and efficiency of the search indexes are retained after the randomization. We have shown through experiments that retrieval performance comparable to plaintext retrieval can be achieved. In the next chapter, we will provide an in-depth study on the security-efficiency trade-off achieved by the proposed techniques, and quantitatively compare such trade-off with cryptography-based approaches.

CHAPTER 5

A Comparative Study for Confidentiality-Preserving Multimedia Search Techniques

The techniques that we proposed in the previous chapter is from a joint signal processing and cryptography point of view. We have demonstrated the great efficiency of such techniques and in this chapter, we further study the security aspect of the confidentiality-preserving search problem, and provide a comparative study of our proposed technique with primarily cryptography techniques in terms of the security-efficiency trade-off offered by them. Since there is no existing work in the secure computation literature addressing the problem of confidentiality-preserving multimedia search, we will first discuss how existing additive homomorphic encryption and the recent advancement in fully homomorphic encryption can be potentially used for multimedia search. We compare these two types of techniques in terms of their search accuracy on an actual encrypted image database, as well as their security strength and computational efficiency. We hope such a quantitative comparison between these two types of techniques for the problem of secure search can reveal some insights in practical design of secure computation techniques for real-world applications involving digital multimedia.

5.1 Review of Cryptographic Techniques

Semantically secure homomorphic public-key encryption schemes are central cryptographic tool for many secure multi-party computation problem. Below, we briefly review the basics of simple additive homomorphic encryption and recent advance of fully homomorphic cryptosystems, then discuss how such techniques can be applied for the problem of secure multimedia search.

Additive Homomorphic Encryption In an additive homomorphic cryptosystem, given encryptions $[a]$ and $[b]$, the encryption of their summation $[a + b]$ can be computed by $[a + b] = [a][b]$, where all the computations are performed in the encrypted domain, without decryption. Following the above property, the multiplication of an encrypted value $[a]$ with a known constant b in the clear can be computed as $[ab] = [a]^b$.

One of the representative additive homomorphic cryptosystem is proposed by Paillier [80], which is based on the decisional composite residuosity problem. Let $n = pq$ of size k, where p and q are large prime numbers and k is from the range

$1000 - 2048$. Randomly select a base g ($g = n + 1$ will do). Then to encrypt a plaintext message $m \in \mathbb{Z}_n$, the user will select a random value $r \in \mathbb{Z}_n$ and computes the ciphertext $c = g^m r^n \bmod n^2$. The parameters (n, g) are the public keys and the pair (p, q) serves as the private key. Given a ciphertext c, its plaintext message m can be obtained by $m = \frac{L(c^\lambda \bmod n^2)}{L(g^\lambda \bmod n^2)} \bmod n^2$, where $L(u) = \frac{u-1}{n}$. It is easy to see that the Paillier is additively homomorphic and for an encryption $[m]$, re-randomizing it can be done without knowing the private key by $[m]r^n \bmod n^2$. More details of the Paillier cryptosystem can be found in [80].

Fully Homomorphic Encryption Earlier homomorphic cryptosystems [22, 37, 40, 80] support either addition or multiplication between encrypted values, but not both operations at the same time. This brings challenges to many secure computation problems because many operations such as computing the Euclidean distance between two encrypted vectors require both addition and multiplication. With only additive or multiplicative homomorphic cryptosystem, a cryptography protocol that involves communication between the two computing parties is typically required.

More recently, in a breakthrough work, Gentry [38] constructed a fully homomorphic encryption (FHE) scheme capable of evaluating an arbitrary number of additions and multiplications (thus compute any function) on encrypted data. The mathematics and construction details in [38] are quite involved, but the basic idea can be summarized as follows. An initial “somewhat homomorphic” scheme based on ideal lattice is constructed to allow evaluation of essentially unlimited addition and a certain amount of multiplication. This initial scheme is somewhat homomorphic because the errors in the ciphertext grows with more operations, as such only a limited amount of multiplication can be supported. To achieve fully homomorphic encryption, the ciphertext has to be re-encrypted through a technique called ”bootstrapping”, so that errors in the ciphertext can be cleaned and unlimited number of operations can be allowed.

Following this first construction of fully homomorphic encryption, there have been subsequent developments that try to improve the efficiency of FHE [12,13,101, 103,108]. Although the most recent solutions of FHE have improved upon the initial construction of Gentry, with more efficient encryption and shorter ciphertexts, there is still a long way to go before FHE can be practical for real-world applications. As such, Lauter et al. [57] discussed the possibility of using a somewhat homomorphic encryption, which is more efficient than their FHE counterparts, for applications that require only a limited amount of multiplication.

Using Homomorphic Encryption for Multimedia Search As discussed in Section 4.2, the application of rank-ordered image search has different settings from many secure computation work such as privacy-preserving face recognition [28, 92]. The challenge here is that the database has access only to the encrypted images and encrypted features, and rank-ordered search results rather than a binary exact matching is required. To the best of our knowledge, there is no existing work that address the problem of rank-ordered multimedia search using homomorphic encryption. Below, we discuss possible scenarios and constructions of using homomorphic encryption for secure multimedia search.

We first provide some notations. We assume that there are N images in the database, and each image has a visual feature $\mathbf{f}_i \in \mathbb{R}^n$. The query image is denoted as Q and its visual feature is $\mathbf{q} \in \mathbb{R}^n$. Paillier homomorphic encryption of a plaintext message m is denoted as $[m]$, and fully homomorphic encryption is denoted as $[[m]]$. The encryption of a feature vector is just the encryption of its individual components, i.e., $[\mathbf{f}] = \{[f_1], [f_2], \cdots, [f_n]\}$.

(1) Scenario-1: Additive homomorphic with encrypted query: In this baseline scenario, we use additive homomorphic encryption to encrypt the visual features of both the database images and the query image. Since the database will return a list of encrypted images similar to the query image, encrypting the feature of the query image is important to prevent the server from inferring the content of returned images using the query feature.

The computational task in this scenario is to compute distance between encrypted vectors $[\mathbf{f}]$ and $[\mathbf{q}]$. Take the commonly used L_2 distance as example, we need to compute $\sum_{i=1}^{n}(f_i - q_i)^2$ using only encrypted values $[f_i], [q_i]$. Unfortunately, with additive homomorphic encryption alone, such computation is impossible without decryption because the computation involves both addition and multiplication. Since the database holds only the encrypted features without knowing the secret key, in order to proceed with the computation, the database needs to send back all the encrypted features $[\mathbf{f}_i],\ i \in \{1, \cdots, N\}$ to the user. The user then decrypts all the features and compute distances on his/her end. The ranking result on the computed distances will be sent back to the database to retrieve similar images. Although the visual features typically have smaller size than the image itself, this naïve base-line scenario is still highly impractical because each query will require the database sending back the entire database of encrypted features. To be more efficient, the user might as well stores all the visual features on his/her local machine and computes similarity by his/herself. This alternative costs storage space and computational burden on the user and fails to utilize the computation power of online services.

(2) Scenario-2: Additive homomorphic with plaintext query: In order to fully utilize the computational power of the cloud, we need to minimize the computation and involvement on the user side. In this scenario, we make a relaxation such that the query feature is not encrypted but sent in plaintext to the database.

The computational task in this scenario is to compute distance between an encrypted feature $[\mathbf{f}]$ and a plaintext feature $\mathbf{q}$. This can be done directly in the database without communication with the user. We give two examples with dot product and L_2 distance, respectively. Computing dot product between a plaintext vector and an encrypted vector is directly supported by additive homomorphic. To see this, the dot product $\mathbf{f} \cdot \mathbf{q} = \sum_{i=1}^{n} f_i q_i$ can be computed in the encrypted domain as $[\mathbf{f} \cdot \mathbf{q}] = \Pi_{i=1}^{n} [f_i]^{q_i}$, where $\mathbf{q}$ is the plaintext query feature. For L_2 distance, $\|\mathbf{f} - \mathbf{q}\|_2 = \sum_{i=1}^{n} (f_i - q_i)^2 = \sum f_i^2 - 2 \sum f_i q_i + \sum q_i^2$. The encrypted distance value thus can be computed as $[\sum f_i^2] \cdot (\Pi [f_i]^{q_i})^{-2} \cdot [\sum q_i^2]$. To allow the database compute the distance without interacting with the user, the user can provide the database an encrypted value $[\sum f_i^2]$ for each feature in the database.

The N encrypted distance values between the query feature and every database feature will then be sent back to the user for decryption and ranking. The security consideration of allowing the query feature in clear is that the database can infer the content of the query image and the final images returned from the search. To mitigate such a security concern, the user can add some noise to the ranking result, so that not all requested images will be similar to the query. Adding noise increases security at the cost of less accurate search.

(3) Scenario-3: Fully homomorphic with encrypted query: In this last scenario, we consider that FHE is used to encrypt features from both the query and database images. Despite that there is no efficient FHE implementations available, this scenario still helps us understand how FHE, if efficiently available in the future, can help address the problem of confidentiality-preserving multimedia search.

With both query and database features encrypted by FHE, the computation of any distance function between $[[\mathbf{f}]]$ and $[[\mathbf{g}]]$ can be done directly in the encrypted domain without interaction with the user. However, the ranking of the encrypted distance values cannot be done alone by the database. This is because a semantically secure FHE should prevent the database from learning any information from the ciphertext, therefore, the database cannot learn ranking information from the encrypted distances without interacting with user. To obtain the final ranking, the database can either send N encrypted distance values to the user or send $N(N-1)/2$ encrypted binary values indicating the pair-wise relation of encrypted distances. The user then computes the ranking and requests similar images from the database. We can see that even FHE cannot completely eliminate the interaction with the user in order to complete the task of content-based image search.

5.2 Comparison on Search Accuracy

We first compare the homomorphic based technique with our proposed randomization techniques in terms their search accuracy. The experimental setup is the same as in Section 4.5. For conventional content-based image search without any protection, color histograms can be compared using L_1 distance. In the confidentiality-preserving search, the color histogram is either encrypted using homomorphic encryption or scrambled using feature/index randomization techniques.

Homomorphic encryption operates on integer values. This implies that if the feature vector is in floating point, it has to be properly scaled and quantized. This will bring quantization error to the distance computation, although such error can be made quite small and with little impact on the search performance. The color histogram used in this experiment contain only integer values, so homomorphic encryption can be applied without causing quantization error and the distance between encrypted features will be exactly the same as that of their plaintext versions. Therefore, we expect confidentiality-preserving search using homomorphic encryption to have the same performance as the conventional search.

Feature/index randomization technique scrambles the visual features or search indexes, and approximately preserves the distance between original features. The approximate distance preserving property ensures that the search accuracy is preserved with only slight degradation. In Fig. 5.1, we compare the search accuracy of different confidentiality-preserving techniques. Overall, we can see that different techniques achieve search accuracies that are close to each other. Since feature randomization operates on global color histogram while index randomization utilizes the indexes generated from local color histograms, we discuss them separately. For search indexes, both the homomorphic encryption and secure inverted index retain the accuracy of using plaintext indexes, therefore, we only show the curve of secure inverted index for clarity. The secure min-hash technique has a slight performance drop at hash length 256, but its performance can be made close to the plaintext index by increasing the hash length. It should be noted that the distance metric used in these two schemes are Jaccard similarity and number of identical elements, respectively. Computing such distance metrics between vectors encrypted by homomorphic encryption is involved and requires heavy communication with the user.

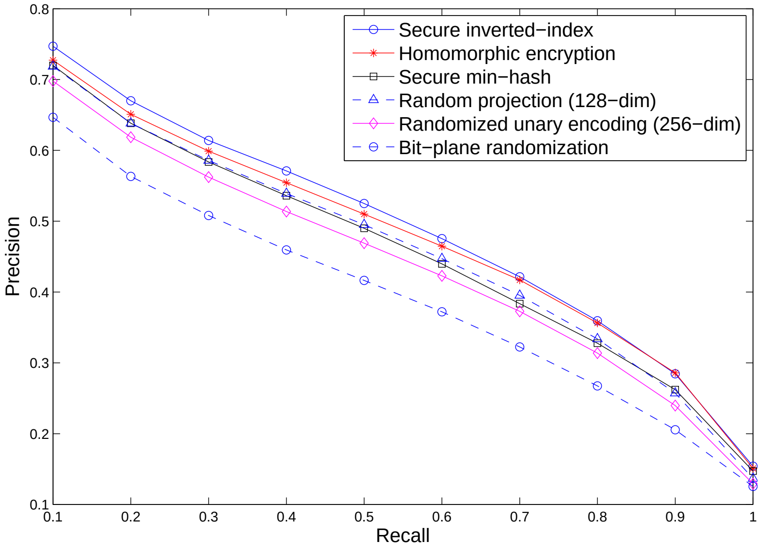


Figure 5.1: Comparison of search accuracy of different techniques

The other four curves capture the search accuracy of using global color histogram protected by homomorphic encryption, bit-plane randomization, random projection, and randomized unary encoding, respectively. The search accuracy using homomorphic encryption technique is the same as the search accuracy using plaintext features, and is the best among the four. We can see from this figure that random projection and randomized unary encoding preserve the search accuracy with only a slight degradation. Furthermore, the search accuracy of using random projection and randomized unary encoding can also be made arbitrarily close to the performance of plaintext search by increasing the feature dimension. Among the three feature randomization techniques, bit-plane randomization has relatively larger degradation on the search accuracy because the distance between randomized features is only an upper bound on the original L_1 distance.

The comparison above demonstrates that homomorphic encryption can retain the exact search accuracy of a conventional scheme that operates on plaintext features, while the index and feature randomization techniques also achieve performance very close to that of the homomorphic encryption. The gap between the two can be made arbitrarily small by increasing the feature dimension for techniques such as random projection, randomized unary encoding, and secure min-hash. It should be noted that homomorphic encryption will greatly expand the encrypted feature size, which we will discuss later in this chapter, so at the same protected feature size, the performance between the homomorphic encryption and feature/index randomization techniques should be negligible.

5.3 Comparison on Security-Efficiency Trade-off

In this section, we discuss the security concerns for the application of confidentiality preserving search of multimedia, demonstrate quantitative results on the protection level achieved by the different techniques, and then specifically discuss the challenges in employing techniques such as homomorphic encryption and cryptography protocols in terms of their computational and communication complexity.

5.3.1 Security Objective for Rank-Ordered Multimedia Search

In the confidentiality preserving multimedia search scenario considered in this dissertation, the server stores only the encrypted images and randomized features, and performs retrieval based on randomized query features. We model the server as a semi-honest adversary, i.e., it follows the execution requirement of the protocol but may use what it sees during the execution to infer additional information. Such a semi-honest model is applicable to such scenarios as web service providers, who would like to learn as much as possible about the users for their own benefits, such as better targeted ads, but would not deliberately break the users' privacy. The user who uses these third-party services wants to utilize the service's computational power for reliable storage, easy access, and better organization of his/her private data set, but wants to reveal as little information as possible to the server beyond what is necessary for the server to provide the necessary services.

Given that the database images are already encrypted using highly secure ciphers, the security objective here will be to minimize information revealed from the randomized features and from the search process. Content-based multimedia retrieval relies on comparison of different types of visual features to capture visual or semantic similarity between images. Visual features can reveal important information about image content and therefore storing raw features without any protection or randomization is never wise. First of all, raw features have fixed structure, from which an adversary can infer certain aspects of image content. For example, each bin in a color histogram reveals proportion of that color in the image. A large proportion of blue color might indicate sky or sea, while a large proportion of green color can suggest trees or grasses. Second, storing raw features allows an adversary to compare them with features of other known images. For example, a close match of salient features such as SIFT can give an adversary high confidence that an encrypted image may contain certain objects such as buildings and landmarks. Both the homomorphic encryption based technique and feature/index randomization techniques will hide the fixed visual feature structure and values, and make it difficult for an adversary to probe the content of encrypted images using known images.

The second source of information leakage is from the search process, where the server will compute distance between the query feature and all the features stored in the database. The result is a list of images ranked by their similarity to the query. The information revealed in this process is the similarity among database images. We will see that such an information leakage is inevitable for the application of rank-ordered search. The first major reason is that the server provides the search functionality and needs to return the similar images. Therefore, the server will know that the returned images are similar to each other. This is different from some secure multi-party computation problem such as binary matching of biometrics or text keywords where only a binary answer is returned and the server can be made oblivious of the matching result. We will show in the following subsections that the utility requirement of returning similar images has some inherent security implications that need to be taken into account when designing secure solutions. The second reason is efficiency. Allowing the server to compare distance between randomized features is necessary to achieve a practical scheme that avoids multiple rounds of communication between the server and the user, as is typically required in secure multi-party computation. This is particularly important for search over large multimedia databases beyond a few hundred or thousand entries, because for each query, the communication bandwidth involved in sending intermediate encrypted values, such as homomorphicly encrypted distance values, back to the user for distance comparison is formidably expensive.

For homomorphic encryption based technique, the server can infer image similarity by observing the search results. For feature/index randomization techniques, the server can directly compute feature distances to infer the similarity information about the encrypted images, and learn the distance distribution between the raw features, because the randomization techniques are approximately distance-preserving. For text documents, the relative frequency of letters or words may reveal its plaintext counterpart, but multimedia content and their signal representations are far more diverse than letters and words. In the following subsections, we design several experiments to study the security implication of revealing distance distribution and demonstrate that the distribution of distances among visual features encodes only a limited amount of information and cannot be easily used to infer the plaintext multimedia content by an adversary.

5.3.2 Protection on Individual Features

As we mentioned earlier, the raw visual features have fixed structure, so that each element in a feature vector has physical meanings that may reveal image content information. Simple permutation of the feature vector is not sufficient because feature values typically have smoothness and correlation property that can be exploited. Homomorphic encryption of each feature value essentially converts the feature vector into a random vector where each component can be considered as an i.i.d. random variable; Feature/index randomization techniques scramble the feature structure and increase randomness of the resulting feature values by jointly using cryptographic primitives and signal processing techniques, while approximately preserving distance between feature vectors.

We use three different metrics to measure the level of protection achieved by the different encryption and randomization techniques. The three metrics are autocorrelation function, entropy, and conditional entropy of the feature vectors. We also generate random feature vectors whose values are drawn from i.i.d. uniform distribution to simulate the results that we can expect from homomorphic encryption of the feature vectors.

The first metric is the autocorrelation function of the feature vector, which measure how correlated the neighbouring feature elements are. The autocorrelation function for the raw color histogram, visual words representation, and randomized features/indexes using different algorithms are shown in Fig. 5.2 and 5.3. We can see that the original color histogram and visual words representation both have non-negligible correlation for lags larger than 0, which means there exist correlation between nearby feature values. For both encrypted and randomized features/indexes the correlation between neighboring feature values or index dimensions have been reduced to close to 0, similar to what we can expect from a sequence of i.i.d. random numbers.

The other two metrics are entropy and conditional entropy of the feature vectors. Given all the feature vectors generated from the 1000 images in the Corel image database, we quantize the entire range of feature values into 256 levels. We then consider the quantized feature value as a random variable and measure its entropy. A higher entropy indicates the feature value has a distribution closer to uniform, thus higher randomness. The conditional entropy $H(X_2|X_1)$ measures randomness of a feature value given its immediate neighbor. The conditional entropy can be approximated by $H(X_2|X_1) = -\sum_{ij} \mu_i P_{ij} \log P_{ij}$, where μ_i is the ensemble distribution of the feature values and P_{ij} is the transition probability.

The entropy and conditional entropy for randomized features/indexes from different algorithms are shown in Table 5.1. The results are averaged over 50 runs of randomized features generated by different secret keys. We can see that both the raw color histogram and visual words representation have relatively low entropy and conditional entropy, which implies that raw features and indexes have limited

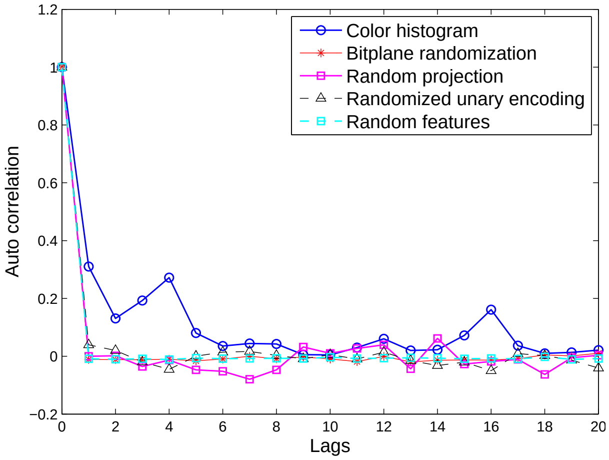


Figure 5.2: Autocorrelation function on randomized features

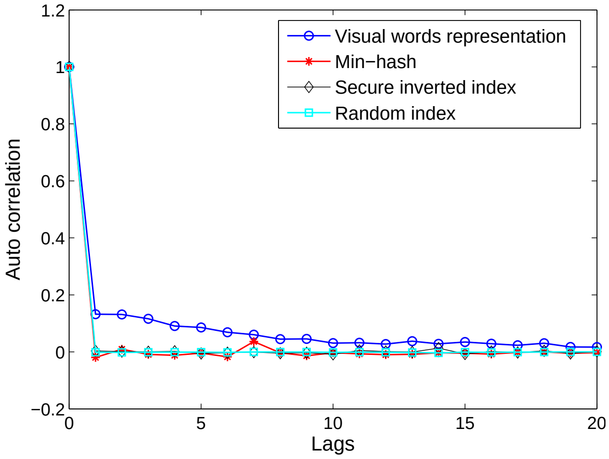


Figure 5.3: Autocorrelation function on randomized indexes

randomness and demonstrate inherent smoothness and correlation among feature values. The features encrypted by homomorphic encryption can be expected to have i.i.d. uniform distribution, so their randomness is measured using uniform random vectors, which achieves the highest entropy and entropy rate. The feature/index randomization techniques also generate protected features with high entropy similar to that of pure random vectors. Since we used 256 levels to quantize the feature values, the maximum possible entropy is 8 bits for a uniform random variable, and lower for a Gaussian random variable. The features from random projection and randomized unary encoding follow Gaussian distribution, and we can see their randomness is close to what can be achieved by a Gaussian random vector; while all the other randomized features/indexes follow uniform distribution, and their entropy are all close to 8 bits.

Table 5.1: Entropy and conditional entropy for randomized features/indexes

Feature type	$H(X)$	$H(X_1\|X_2)$
Color histogram	1.95	1.71
Bitplane randomization	7.72	5.05
Random projection	7.00	6.80
Random unary encoding	6.90	6.80
Gaussian random vectors	6.89	6.72
Index type	$H(X)$	$H(X_1\|X_2)$
Visual words	2.59	2.50
Min-hash	7.93	5.82
Secure inverted index	7.97	7.26
Uniform random vectors	8.00	7.58

The above experiments indicate that the feature/index randomization techniques can generate features and indexes that have similar randomness to a pure random vector or features after the homomorphic encryption. The feature structure is scrambled, and each individual feature values become more independent. The physical meaning in the feature vectors are therefore hidden from the adversaries.

5.3.3 Protection on The Search Process

During the search process, the server will compute distance between randomized features and return a list of encrypted images ranked by their similarity to the query. Therefore, the server will know that the returned images are likely to be similar, and for feature/index randomization techniques, the server will also know the distance distribution among the randomized features. In this subsection, we carry out several experiments to see if revealing such information will be of significant security concern for feature/index randomization techniques.

Clustering on randomized features For homomorphic encryption schemes, the distance between feature vectors are encrypted and thus not directly obtainable by the server. From server's perspective, the encrypted features are the same as a set of i.i.d. uniform random vectors. For feature/index randomization techniques, since the server can compute distance between randomized features, it will be able to perform a clustering of all the features in the database and group encrypted images into clusters where each cluster contains images that are likely to be similar to each other. In the Corel image database that we used here, there are 10 categories each with 100 images. A perfect clustering will generate 10 categories each with the exact 100 images from that category. The better clustering that the server can get using the distances among features, the more information about the database is revealed from the feature distance information.

We carry out K-means clustering on the randomized features/indexes as well as the i.i.d. uniform random vectors that we expect from homomorphic encryption. We assume that the server knows the number of clusters in the database as a prior knowledge. To measure the randomness of the clustering result, we use two metrics. The first one is the average entropy of image categories over the 10 clusters. We consider the image category as a random variable, taking values from 1 to 10. After clustering, each cluster will contain a list of images each with a category number. The entropy of image category can be computed for each cluster and averaged to get a value of average cluster entropy. A perfect clustering will generate an average cluster entropy of 0, and higher entropy indicate that the clustering is more random and more different from the ground-truth. The second metric is the number of unique image categories among the 10 clusters. For each cluster, we consider the dominant image category as the cluster category, then we count how many unique cluster categories are there. A perfect clustering will generate 10 unique categories. The clustering results averaged over 50 runs of K-means clustering with different initial random centroids are shown in Table 5.2.

From the result, we can learn several things. First, clustering on the random feature vectors or vectors from homomorphic encryption achieve the highest entropy and fewest unique cluster categories, indicating the clustering result is most different from the ground-truth. Second, the randomized features and indexes from the randomization techniques achieve similar randomness to that of the raw color histogram. This can be expected from the approximate distance preserving property of the randomization algorithms. Third and most importantly, even the clustering

Table 5.2: K-mean clustering results

Feature type	Average cluster entropy	# of unique cluster categories
Color histogram	1.61	8.72
Random projection	1.50	8.40
Randomized Unary encoding	1.56	8.52
Bitplane randomization	2.49	7.74
Secure Min-hash	1.92	7.96
Secure inverted index	2.16	7.56
Random feature vectors	3.26	7.06

results on raw color histogram are quite different from the ground-truth. We can expect each cluster to contain images from 3 to 6 different categories. This can be mainly attributed to the semantic gap in image search, where low level visual features cannot capture very well high level semantic concepts. In other words, there is a gap from knowing the visual features to knowing the semantic concept of the image, which actually helps add another security layer for multimedia related applications.

Image categories indistinguishability From the previous experiment, we know that the server will not be able to obtain the exact ground-truth clustering from the randomized features. In this subsection, we perform experiments to demonstrate that even if the server can obtain the ground-truth clustering, these clusters of randomized or encrypted features will be highly indistinguishable from the server's point of view.

We assume that the server has the prior knowledge of the category names in the database, but does not know which name corresponds to each of the 10 clusters of encrypted images. In the Corel image database used in this chapter, the 10 categories are “African”, “Beach”, “Architecture”, “Buses”, “Dinosaurs”, “Elephants”, “Flowers”, “Horses”, “Mountain”, and “Food”. The first experiment we carry out here is to see that given a plaintext image from one of the 10 categories, whether the server can successfully associate it with the correct cluster of encrypted images. Since the server does not know the secret key used in randomization, we will randomize the feature of the known plaintext image using a randomly chosen key and use the randomized feature as query to compare with features in the database. The retrieval performance of using every image in the database as query but randomize its feature using a randomly chosen key is shown in Fig. 5.4 and Fig. 5.5.

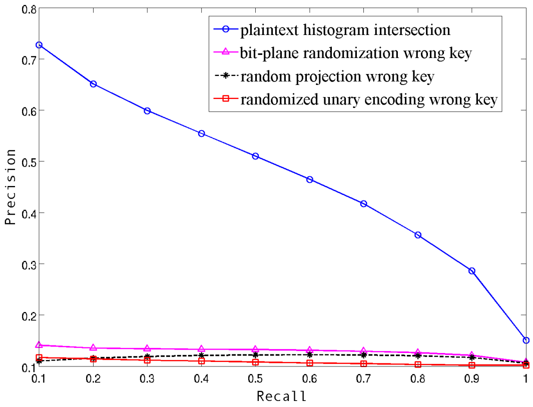


Figure 5.4: Retrieval using a wrong key for feature protection schemes

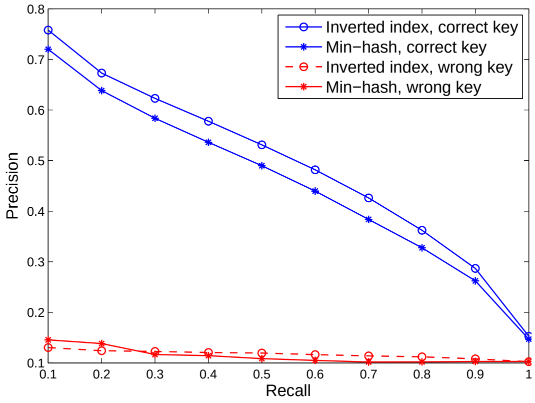


Figure 5.5: Retrieval using a wrong key for secure index schemes

Since the database has 100 images in each of the 10 categories, a random selection from the database would imply a precision value around 0.1 for all recall values. From this figure, we can see that the retrieval precision of feature/index randomization techniques is reduced to around 0.1 if a different secret key is used to randomize the feature or index. In other words, a query index randomized by a different key from the correct one will be equally like to be closest to any randomized feature in the database. Therefore, without knowing the correct secret key, retrieval from an encrypted database is equivalent to picking images randomly from the database. The chance of the server associating a plaintext image of known category to the correct cluster in the encrypted database is no better than random guessing.

Next, we carry out an experiment to see when the server has multiple plaintext images from some image category, whether the distribution of visual features among those images of the same category can be used to differentiate clusters of encrypted images. For each of the 10 categories in the Corel database, we first divide the 100 images in that category equally into two sets $S_{i1}, S_{i2}, i = 1, 2, \cdots, 10$, each with 50 images. The distance distributions from $\{S_{i1}\}$ are used as query to search for closest match in distance distributions from $\{S_{i2}\}$. The purpose of such an experiment is to see whether the distance distribution of visual features has the discriminative power to differentiate different image categories. The less distinctive of distance distributions among image categories, the less information about the image database is revealed. Kullback-Leibler divergence is used as a distance metric for the distributions and the probability of correct match over 100 runs with different secret keys is shown in Table 5.3.

From the table, we can see that for the 10 categories in the Corel database, the

Table 5.3: Search accuracy using distance distribution of different categories

Feature type	Prob. of match	Feature type	Prob. of match
Color histogram	40%	Visual words	40%
Bitplane randomization	29.7%	Secure inverted index	24.8%
Random projection	25.4%	Secure Min-hash	16.6%
Unary encoding	9.8%	Random feature vectors	9.7%

probability of correctly matching two distance distributions from the same category is 40% for both the raw color histogram and visual words representation. This relatively low match accuracy, as compared to the search accuracy using visual features, implies that distance distribution between the visual features is not a very good discriminative feature to differentiate different image categories. After randomization, the match accuracy on the randomized features and indexes are further reduced. Especially for randomized unary encoding, the match accuracy is close to 10%, which is essentially like random guessing and similar to what can be achieved by i.i.d. random vectors. Another thing to notice is that we only have 10 categories in the image database. For a larger image database with more categories, we can expect the match accuracy to further decrease.

All the experiments in this subsection show that due to the extremely diverse representation of multimedia data and the semantic gap between the visual features and semantic concept, it is extremely difficult for an adversary to learn useful information about the image content from the distance distribution of visual features.

Measured by the metrics used in the above experiments, the feature/index randomization techniques can achieve security performance close to that of homomorphic encryption.

5.3.4 Challenges in Employing Homomorphic Encryption Schemes

From the previous comparisons, we have seen that homomorphic encryption schemes achieve the exactly the same search accuracy as that of using plaintext features, and offer the highest amount of randomness in terms of confidentiality protection for the visual features and the search process. The feature/index randomization techniques, although not designed as encryption schemes, come very close to the performance of homomorphic encryption schemes in terms of both search accuracy and confidentiality protection. In this section, we discuss some practical challenges and considerations when employing these two types of techniques in the application of confidentiality preserving rank-ordered multimedia search.

Security benefit of cryptographic approaches The feature/index randomization techniques proposed in [62, 64] are designed with efficiency and distance preserving property in mind, but strictly speaking, they are not encryptions as those commonly used cryptographic ciphers. Secure cryptographic ciphers require semantic security, which demands randomized encryption. Due to the distance preserving requirement, the feature/index randomization schemes are deterministic. Homomorphic encryption, on the other hand, offers randomized encryption of visual features and prevents the server from computing distance between encrypted features directly. In some secure computation problems such as those involve text documents and biometrics, such a security benefit of randomized encryption and hiding the computation results would be important. However, for the problem of rank-ordered image search, the security benefit from homomorphic encryption may not justify its high computational complexity.

In the previous comparisons, we have shown that distance information from visual features is not a discriminative feature for differentiating different image categories. Furthermore, the requirement on the server to return a list of encrypted images similar to the query brings some inherent security implication that might diminish the benefits from cryptographic techniques such as homomorphic encryption. The main reason is that the utility requirement of returning similar images inevitably reveals the information that those returned images are similar to each other. Therefore, even if the encryption for the individual features are semantic secure, some information about the ciphertext will be revealed. This is similar to the application of statistical database, where the database is required to return global statistical information about the private data it holds. Such a utility requirement makes semantic security impossible for statistical database, as proved by Dwork [26]. Instead, differential privacy is used to quantify security from different perspective for those applications where ciphertext carries utility to the adversary and semantic security is impossible. Typical technique to achieve differential privacy is to add noise to the returns from the database at the cost of noisy and less useful results. Exploring differential privacy formulation for the problem of image search can be an interesting issue for future research, but may be non-trivial or even impossible given the unique application settings of the problem. Nevertheless, we can see that using homomorphic encryption may not bring significant security benefit over feature/index randomization techniques for the problem of confidentiality-preserving multimedia search.

Efficiency cost of cryptographic approaches In addition to the limited security benefit, the huge computational and communication complexity is another major limitation of homomorphic encryption schemes at this moment. First of all, using encryption such as homomorphic encryption is computationally intensive and causes large amount of ciphertext expansion. Some comparison between the Paillier homomorphic encryption and the proposed randomization algorithms are listed in Table 5.4, where the encryption time and ciphertext size of all the 1000 features/indexes in the database are shown. The homomorphic encryption implementation is based on a C library from http://acsc.cs.utexas.edu/libpaillier/. All implementations are in C/C++ and run on a Linux desktop with 3.0GHz dual core CPU and 4GB RAM. The randomized features are stored in binary format and further compressed using zip. We can see that homomorphic encryption takes far longer time to encrypt the 1000 color histograms from the Corel database and results in largest expansion on the feature size, which also implies that homomorphic encryption will incur high communication cost in order to transfer the encrypted features to the server. Among the feature/index randomization techniques, randomized unary encoding and secure min-hash have relatively longer running time, because they have more randomization steps in their algorithms. Since there is no efficient fully homomorphic encryption implementations available yet, we do not report its complexity here, which can be expected to be much higher than Paillier at this moment.

Table 5.4: Efficiency comparison of feature randomization schemes

	Encryption time	Ciphertext size / expansion factor
Paillier Homomorphic	1778.5s	32005KB / 241.8
Bitplane randomization	0.24s	159KB / 1.2
Random projection	0.38s	462KB / 3.5
Randomized unary encoding	9.64s	457KB / 3.5
Secure inverted index	0.32s	246KB / 2.1
Secure Min-hash	3.04s	296KB / 2.5

The advantages and disadvantages of our proposed randomization techniques and the homomorphic encryption based techniques are summarized in Table 5.5.

Table 5.5: Summary of comparison

	Homomorphic encryption	Feature/index randomization
Advantages	High search accuracy Randomized encryption	High search accuracy Computationally efficient Minimum user involvement
Disadvantages	High computational complexity Frequent user involvement	Deterministic randomization

The advantages of using homomorphic encryption are that it retains the search accuracy of plaintext features and offers randomized encryption so that the server cannot obtain distance between encrypted features directly. The disadvantage is that it is too computation and communication intensive to be practical, requiring frequent user involvement in order to obtain the ranking results. On the other hand, feature/index randomization techniques have the advantage of being highly efficient and requiring minimum user-involvement when computing the search results. The search accuracy and confidentiality protection offered by feature/index randomization are very close to that of homomorphic encryption schemes. The limitation of feature/index randomization is that they are deterministic methods and thus the server can learn distance distribution of randomized features. We provided various experiments to demonstrate that the revealing distance distribution may not be a significant security concern for multimedia data and the utility requirement of rank-ordered search has some inherent security implications that may diminish the security benefit of using homomorphic encryption.

5.4 Chapter Summary

In this chapter, we quantitatively compared the security-efficiency trade-off offered by our proposed techniques and alternative cryptography techniques for the problem of confidentiality-preserving multimedia search. We first discussed how existing cryptography primitives such as homomorphic encryption can be adapted to the rank-ordered search problem. Such adaptation is highly inefficient given the current state of the art. Furthermore, the utility requirement of rank-ordered search limits the amount of security benefits that homomorphic encryption can bring. To justify that our proposed randomization techniques offer a better security-efficiency trade-off, we devised several metrics and experiments for a quantitative comparison. Such a comparative study suggests that a joint signal processing and cryptography point of view may offer better solutions to online multimedia applications that do not require the highest level of security but demand high efficiency and least user involvement.

CHAPTER 6

Conclusions and Future Perspectives

In this dissertation, we have explored two major research problems regarding trustworthiness and confidentiality of online multimedia data. Trustworthiness and confidentiality are two closely related and increasingly important aspects for the emerging technologies of mobile and cloud computing. In the first part, we evaluate trustworthiness of a multimedia document by estimating its processing history using novel forensic techniques. In the second part, we design algorithms to preserve confidentiality of online multimedia while offering efficient search capability.

To evaluate multimedia trustworthiness, this dissertation proposes a new multimedia forensic framework called "Forensic Hash for Multimedia Information Assurance", or FASHION in short. Under this framework, a compact signature called

“Forensic hash” that encodes information of the original multimedia is utilized to provide enhanced forensic analysis for the multimedia document under question. The forensic capability of FASHION goes beyond a binary answer of whether the multimedia document can be trusted or not, but provides an in-depth assessment on the processing history in terms of the types and parameters of the operations that have been applied on the multimedia data. The FASHION framework bridges two related research areas of image hashing and no-reference multimedia forensics, and combines the benefits from both.

The challenge in FASHION is to design compact hash that provides good forensic capability. To avoid the dilemma of one-scheme-fit-all, we suggested a modular design of the forensic hash for the advantage of flexibility and extensibility. For such a modular construction, we designed alignment component and integrity component, which address different forensic questions but complement each other in a synergistic way. Alignment component aims at estimating geometric transforms such as rotation and scaling, and such estimation enables accurate tampering localization which is the main objective of an integrity component. We proposed compact constructions for both components and demonstrated their robustness and accuracy in revealing processing history of digital images.

A good forensic hash design should have good extensibility, i.e., answering a broad scope of forensic questions by introducing new modules or using existing modules. This dissertation demonstrates such an extensibility by using forensic hash to estimate advanced image editing operation such as seam carving. The forensic hash that we originally designed for estimating geometric transform is shown to be very effective in estimating seam carving without changing the hash design, and therefore can be used to evaluate trustworthiness for images that have undergone advanced editing and tampering.

In addition to learning the processing history, evaluating quality of a multimedia document after all the processing operations is another important aspect of trustworthiness. This dissertation extends the FASHION idea to the task of quality assessment on images that have undergone retargeting operations. We designed compact side information to capture the structure distortion caused by the retargeting operation and demonstrated positive correlation between the proposed quality score and human subjective ratings. Quality assessment focusing on content structure distortion is not well studied and our work is among the first endeavors in this direction.

In today's online services and more broadly the cloud computing paradigm, confidentiality of data stored online is a critical requirement that needs to be satisfied. In this dissertation, we studied the problem of confidentiality-preserving search for online multimedia. The key challenge in this problem is to achieve content-based search capability over multimedia data that have been encrypted or properly protected for privacy concerns, and in the meanwhile, striking a good balance between security and efficiency for practical applications. In light of the efficiency limitations of purely cryptographic approaches, we address the problem from a practical perspective and proposed distance preserving randomizations by jointly applying techniques from image processing, information retrieval, and cryptography. Efficiency and accuracy of the search is demonstrated on a practical image database.

There are several contributions of this work. To the best of our knowledge, our work is the first endeavor in the community to study the problem of confidentiality-preserving rank-ordered search of multimedia. This problem has unique challenges as compared to problems studied in existing secure computation literature. We provided a quantitative study on the security-efficiency trade-off provided by our proposed randomization techniques and alternative cryptographic approaches. Such a quantitative study brings interesting insight on the design of confidentiality-preserving techniques for multimedia data.

The study of trustworthiness and confidentiality issues for online multimedia data in this dissertation brings up many interesting research questions that are worth further exploration. Following our current forensic hash constructions, it will be interesting to provide a theoretical modelling for the FASHION framework, which models the process of generating the forensic hash and also the process of forensic analysis using the hash. A sound theoretical modelling can help answer some fundamental questions, for example, what the optimal design of a forensic hash is for a particular forensic task, and how the hash length relates to the performance of forensic analysis. Our work in this dissertation has focused on robustly estimating particular operations when the multimedia document has undergone multiple other operations. It would be interesting to explore the possibility of estimating the orders of operations, i.e., which operation occurred earlier in the processing chain. This can provide useful information when evaluating trustworthiness of multimedia data.

In addition to the theoretical formulation of FASHION, continued effort should be made toward designing new hash components for a wider range of operations and forensic tasks. For example, forensic hash that identifies different image enhancement operations and Photoshop editing operations would be valuable to tell people which regions of the image has been artistically enhanced and which regions remain faithful to its original form; forensic hash for video data will be an important extension of the technologies that we studied in this dissertation, and we can imagine compactness and efficiency will be critical aspects in forensic hash design for videos. The FASHION spirit can go beyond multimedia forensic applications, as we have already demonstrated in the study of quality assessment on retargeted images. In addition to revealing structural distortion of retargeting algorithms, it would interesting to see whether we can use the compact side information to reconstruct the original image from a retaregeted one. Depending on how faithful the reconstruction is and how compact the side information is, this can potentially offer a different image compression paradigm.

The search capability that we have studied in this dissertation is an important functionality for secure online multimedia management. A natural next step is to explore a broader range of confidentiality-preserving operations on multimedia data. For example, directly editing multimedia data such as images and videos online without leaking content information can be useful to save users from the computational burdens of such editing operations. Categorization, summarization, and annotation for encrypted multimedia collections are some other desirable functionalities that a secure online multimedia service could provide [63]. As we have demonstrated in this dissertation, designing confidentiality-preserving techniques for multimedia data has unique challenges. A better understanding on what would be the proper security definition for multimedia applications and more effort on seeking the best security-efficiency trade-off will be needed in these future explorations.

Bibliography

[1] Corel test set. `http://wang.ist.psu.edu/~jwang/test1.tar`.

[2] RetargetMe database. `http://people.csail.mit.edu/mrub/retargetme/index.html`.

[3] R. Agrawal, J. Kiernan, R. Srikant, and Y. Xu. Order preserving encryption for numeric data. In *Proc. SIGMOD*, 2004.

[4] S. Avidan and A. Shamir. Seam carving for content-aware image resizing. *ACM Trans. on Graphics*, 26(3), 2007.

[5] H. G. Barrow, J. M. Tenenbaum, R. C. Bolles, and H. C. Wolf. Parametric correspondence and chamfer matching: two new techniques for image matching. In *Proc. of the 5th Int'l Joint Conf. on Artificial intelligence*, pages 659–663, 1977.

[6] S. Bayram, H. T. Sencar, and N. Memon. Source camera identification based on cfa interpolation. In *Proc. of IEEE Int. Conf. on Image Processing*, 2005.

[7] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 24(4):509–522, 2002.

[8] S. Bhattacharjee and M. Kutter. Compression tolerant image authentication. In *Proc. of IEEE Int'l Conf. on Image Processing*, volume 1, pages 435–439, Oct. 1998.

[9] D. Boneh, G. Crescenzo, R. Ostrovsky, and G. Persiano. Public-key encryption with keyword search. In *Proceedings of Eurocrypt*, 2004.

[10] F. L. Bookstein. Principal warps: Thin-plate splines and the decomposition of deformations. *IEEE Trans. Pattern Anal. Mach. Intell.*, 11:567–585, June 1989.

[11] F. L. Boostein. *Morphometric tools for landmark data: geometry and biology.* Cambridge Univ. Press, 1991.

[12] Z. Brakerski and V. Vaikuntanathan. Efficient fully homomorphic encryption from (standard) LWE. In *FOCS*, 2011.

[13] Z. Brakerski and V. Vaikuntanathan. Fully homomorphic encryption from ring-lwe and security for key dependent messages. In *CRYPTO*, 2011.

[14] R. Brinkman, J. M. Doumen, and W. Jonker. Using secret sharing for searching in encrypted data. In *Workshop on Secure Data Management in a Connected World*, pages 18–27, 2004.

[15] A. Broder. On the resemblance and containment of documents. In *Proceedings of Compression and Complexity of Sequences*, pages 21–29, 1997.

[16] A. Broder, M. Charikar, A. Frieze, and M. Mitzenmacher. Min-wise independent permutations. In *Proceedings of the 30th ACM Symposium on Theory of Computing*, pages 327–336, 1998.

[17] J. Canny. A computational approach to edge detection. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 8:679–714, 1986.

[18] G. Carneiro, A. B. Chan, P. J. Moreno, and N. Vasconcelos. Supervised learning of semantic classes for image annotation and retrieval. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 29(3):394–410, 2007.

[19] E. Chang, M. S. Kankanhalli, X. Guan, Z. Huang, and Y. Wu. Robust image authentication using content based compression. *Multimedia System*, 9:121–130, August 2003.

[20] O. Chum, J. Philbin, M. Isard, and A. Zisserman. Scalable near identical image and shot detection. In *Proceedings of the International Conference on Image and Video Retrieval (CIVR)*, 2007.

[21] O. Chum, J. Philbin, and A. Zisserman. Near duplicate image detection: min-hash and TF-IDF weighting. In *British Machine Vision Conference (BMVC)*, 2008.

[22] I. Damgård and M. Jurik. A generalisation, a simplification and some applications of paillier's probabilistic public-key system. In *Proceedings of the 4th International Workshop on Practice and Theory in Public Key Cryptography: Public Key Cryptography*, pages 119–136. Springer-Verlag, 2001.

[23] R. Datta, D. Joshi, J. Li, and J. Z. Wang. Image retrieval: ideas, influences, and trends of the new age. *ACM Computing Surveys*, 2008.

[24] E. Delp, N. Memon, and M. Wu (eds). *Special Issue on Forensics Analysis of Digital Evidence, IEEE Signal Processing Magazine*, 26(2), March 2009.

[25] J. Duchon. Splines minimizing rotation-invariant semi-norms in sobolev spaces. In *Constructive Theory of Functions of Several Variables*, volume 571 of *Lecture Notes in Mathematics*, pages 85–100. Springer Berlin / Heidelberg, 1977.

[26] C. Dwork. Differential privacy. In *33rd International colloquium on Automata, Languages, and Programming (ICALP)*, 2006.

[27] U. Engelke, V. X. Nguyen, and H.-J. Zepernick. Regional attention to structural degradations for perceptual image quality metric design. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 869 –872, April 2008.

[28] Z. Erkin, M. Franz, J. Guajardo, S. Katzenbeisser, I. Lagendijk, and T. Toft. Privacy-preserving face recognition. *Privacy Preserving Technologies, LNCS*, 5672:235–253, 2009.

[29] Z. Erkin, A. Piva, S. Katzenbeisser, R. L. Lagendijk, J. Shokrollahi, G. Neven, and M. Barni. Protection and retrieval of encrypted multimedia content: when cryptography meets signal processing. *EURASIP Journal on Information Security*, 7(2):1–20, 2007.

[30] H. Farid. Exposing digital forgeries from JPEG ghosts. *IEEE Trans. on Information Forensics and Security*, 4:154–160, 2009.

[31] H. Farid. A survey of image forgery detection. *IEEE Signal Processing Magazine*, 2(26):16–25, 2009.

[32] A. M. Ferman, A. M. Tekalp, and R. Mehrotra. Robust color histogram descriptors for video segment retrieval and identification. *IEEE Trans. Image Processing*, 11(5):497–508, May 2002.

[33] C. Fillion and G. Sharma. Detecting content adaptive scaling of images for forensic applications. In *Proc. of SPIE: Media Forensics and Security*, volume 7541, pages 7541–36, 2010.

[34] M. A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun. of ACM*, 24(6):381–395, 1981.

[35] L. Frédéric and M. Benoit. Rash: Radon soft hash algorithm. In *European Signal Processing Conf. (EUSIPCO)*, 2002.

[36] J. Fridrich and M. Goljan. Robust hash functions for digital watermarking. In *IEEE Proc. Int'l Conf. on Information Technology: Coding and Computing*, pages 178–183, March 2000.

[37] T. El Gamal. A public key cryptosystem and a signature scheme based on discrete logarithms. In *Proceedings of CRYPTO 84 on Advances in cryptology*, pages 10–18. Springer-Verlag New York, Inc., 1985.

[38] C. Gentry. Fully homomorphic encryption using ideal lattices. In *Proceedings of the 41st annual ACM symposium on Theory of computing (STOC)*, pages 169–178. ACM, 2009.

[39] A. Gionis, P. Indyk, and R. Motwani. Similarity search in high dimensions via hashing. In *Proceedings of the International Conference on Very Large Data Bases*, 1999.

[40] S. Goldwasser and S. Micali. Probabilistic encryption. *Journal of Computer and System Sciences*, 28(2):270–299, 1984.

[41] M. Grangetto, E. Magli, and G. Olmo. Multimedia selective encryption by means of randomized arithmetic coding. *IEEE Transactions on Multimedia*, 8(5):905–917, 2006.

[42] J. Han and M. Kamber. *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 2000.

[43] J. Han, M. Kamber, and A. K. H. Tung. *Geographic Data Mining and Knowledge Discovery*, chapter Spatial Clustering Methods in Data Mining: A Survey, pages 1–29. Taylor and Francis, 2001.

[44] C. Harris and M. Stephens. A combined corner and edge detector. In *Alvey Vision Conference*, pages 147–151, 1988.

[45] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254 –1259, 1998.

[46] S. Jeong, C. Won, and R. Gray. Image retrieval using color histograms generated by Gauss mixture vector quantization. *Computer Vision and Image Understanding*, 94:44–66, 2004.

[47] W. Jiang, M. Murugesan, C. Clifton, and L. Si. Similar document detection with limited information disclosure. In *IEEE 24th International Conference on Data Engineering*, 2008.

[48] F. Jing, M. Li, H.-J. Zhang, and B. Zhang. An efficient and effective region-based image retrieval framework. *IEEE Trans. on Image Processing*, 13(5):699–709, May 2004.

[49] M. K. Johnson and H. Farid. Exposing digital forgeries by detecting inconsistencies in lighting. In *Proc. of ACM Multimedia Security Workshop*, 2005.

[50] Z. Karni, D. Freedman, and C. Gotsman. Energy-based image deformation. In *Proceedings of the Symposium on Geometry Processing*, SGP '09, pages 1257–1268, 2009.

[51] J. Katz and Y. Lindell. *Introduction to Modern Cryptography: Principles and Protocols*. Chapman & Hall/CRC, 2007.

[52] H. Kim, J. Wen, and J. D. Villasenor. Secure arithmetic coding. *IEEE Transactions on Signal Processing*, 55(5):2263–2272, 2007.

[53] S. S. Kozat, K. Mihcak, and R. Venkatesan. Robust perceptual image hashing via matrix invariances. In *Proc. of IEEE Int'l Conf. on Image Processing*, pages 3443–3446, Oct. 2004.

[54] P. Krähenbühl, M. Lang, A. Hornung, and M. Gross. A system for retargeting of streaming video. *ACM Trans. Graph.*, 28:126:1–126:10, December 2009.

[55] E. C. Larson and D. M. Chandler. Unveiling relations between regions of interest and image fidelity metrics. In *Proc. SPIE*, page 6822, 2008.

[56] E. C. Larson, C. Vu, and D. M. Chandler. Can visual fixation patterns improve image fidelity assessment? In *IEEE International Conference on Image Processing*, pages 2572 –2575, Oct. 2008.

[57] K. Lauter, M. Naehrig, and V. Vaikuntanathan. Can homomorphic encryption be practical? Cryptology ePrint Archive, Report 2011/405, 2011.

[58] C.-Y. Lin and S.-F. Chang. A robust image authentication method distinguishing jpeg compression from malicious manipulation. *IEEE Trans. on Circuits and Systems for Video Technology*, 11(2):153–168, 2001.

[59] Y.-C. Lin, D. Varodayan, and B. Girod. Distributed source coding authentication of images with affine warping. In *Proc. of IEEE Int. Conf. on Acoustic, Speech, and Signal Processing (ICASSP)*, 2009.

[60] T. Lindeberg. *Scale-Space Theory in Computer Vision.* Kluwer Academic Publishers, 1994.

[61] D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[62] W. Lu, A. Swaminathan, A. L. Varna, and M. Wu. Enabling search over encrypted multimedia databases. In *SPIE/IS&T Media Forensics and Security*, pages 7254–18. Proc. of SPIE, vol. 7254, January 2009.

[63] W. Lu, A. Varna, and M. Wu. Secure video processing: Problems and challenges. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5856–5859, May 2011.

[64] W. Lu, A. L. Varna, A. Swaminathan, and M. Wu. Secure image retrieval through feature protection. In *IEEE Conference on Acoustics, Speech and Signal Processing*, pages 1533–1536, April 2009.

[65] W. Lu, A. L. Varna, and M. Wu. Forensic hash for multimedia information. In *SPIE Media Forensics and Security*, pages 7541–0Y, January 2010.

[66] W. Lu, A. L. Varna, and M. Wu. Confidentiality-preserving search of online multimedia. *under revision*, 2011.

[67] W. Lu and M. Wu. Multimedia forensic hash based on visual words. In *Proc. of IEEE Int'l Conf. on Image Processing (ICIP)*, pages 989–992, Sept. 2010.

[68] W. Lu and M. Wu. Forensic hash for multimedia information assurance. *under review in IEEE Trans. on Information Forensics and Security*, 2011.

[69] W. Lu and M. Wu. Seam carving estimation using forensic hash. In *Proceedings of the 13th ACM multimedia workshop on Multimedia and security*, MM&Sec'11, pages 9–14, New York, NY, USA, 2011. ACM.

[70] J. Lukas, J. Fridrich, and M. Goljan. Digital camera identification from sensor pattern noise. *IEEE Trans. on Information Forensics and Security*, 1(2):205–214, 2006.

[71] Y. Mao and M. Wu. A joint signal processing and cryptographic approach to multimedia encryption. *IEEE Transactions on Image Processing*, 15(7):2061–2075, 2006.

[72] J. Meinguet. Multivariate interpolation at arbitrary points made simple. *Zeitschrift fr Angewandte Mathematik und Physik (ZAMP)*, 30:292–304, 1979.

[73] V. Monga and B. L. Evans. Robust perceptual image hashing using feature points. In *Proc. of IEEE Int'l Conf. on Image Processing*, pages 677–680, Oct. 2004.

[74] V. Monga and M. Mihcak. Robust and secure image hashing via non-negative matrix factorization. *IEEE Trans. on Information forensics and security*, 2, September 2007.

[75] R.T. Ng and Jiawei Han. Clarans: a method for clustering objects for spatial data mining. *Knowledge and Data Engineering, IEEE Transactions on*, 14(5):1003–1016, 2002.

[76] D. Nistér and H. Stewénius. Scalable recognition with a vocabulary tree. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, 2006.

[77] D. Nistér and H. Stewénius. Scalable recognition with a vocabulary tree. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2006.

[78] F. Ono, W. Rucklidge, R. Arps, and C. Constantinescu. Jbig2-the ultimate bi-level image coding standard. In *IEEE International Conference on Image Processing*, volume 1, pages 140–143, 2000.

[79] M. Osadchy, B. Pinkas, A. Jarrous, and B. Moskovich. Scifi - a system for secure face identification. In *2010 IEEE Symposium on Security and Privacy*, pages 239–254, 2010.

[80] P. Paillier. Public-key cryptosystems based on composite degree residuosity classes. In *Proceedings of the 17th international conference on Theory and application of cryptographic techniques*, EUROCRYPT'99, pages 223–238. Springer-Verlag, 1999.

[81] C. Papadimitriou and K. Stieglitz. *Combinatorial Optimization: Algorithms and Complexity*. Prentice Hall, 1982.

[82] O. Pele and M. Werman. Fast and robust earth mover's distances. In *IEEE 12th International Conference on Computer Vision*, pages 460–467, 2009.

[83] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

[84] A. C. Popescu and H. Farid. Exposing digital forgeries by detecting traces of re-sampling. *IEEE Trans. on Signal Processing*, 53(2):758–767, 2005.

[85] Y. Pritch, E. Kav-Venaki, and S. Peleg. Shift-map image editing. In *ICCV*, pages 151–158, Sept 2009.

[86] M. P. Queluz. Towards robust, content based techniques for image authentication. In *IEEE Second Workshop on Multimedia Signal Processing*, pages 297 –302, Dec. 1998.

[87] R. L. Rivest, A. Shamir, and L. Adleman. A method for obtaining digital signatures and public-key cryptosystems. *Commun. ACM*, 21:120–126, February 1978.

[88] S. Roy and Q. Sun. Robust hash for detecting and localizing image tampering. In *Proc. of IEEE Int'l Conf. on Image Processing*, volume 6, pages 117–120, 2007.

[89] M. Rubinstein, D. Gutierrez, O. Sorkine, and A. Shamir. A comparative study of image retargeting. *ACM Trans. on Graphics, Proceedings Siggraph Asia*, 29(5), 2010.

[90] M. Rubinstein, A. Shamir, and S. Avidan. Improved seam carving for video retargeting. *ACM Trans. Graph.*, 27:16:1–16:9, August 2008.

[91] M. Rubinstein, A. Shamir, and S. Avidan. Multi-operator media retargeting. *ACM Trans. Graph.*, 28:23:1–23:11, July 2009.

[92] A.-R. Sadeghi, T. Schneider, and I. Wehrenberg. Efficient privacy-preserving face recognition. In *12th International Conference on Information Security and Cryptology*, 2009.

[93] G. Salton and M. J. McGill. *Introduction to Modern Information Retrieval.* McGraw-Hill, 1983.

[94] A. Sarkar, L. Nataraj, and B. S. Manjunath. Detection of seam carving and localization of seam insertions in digital images. In *Proc. of the 11th ACM workshop on Multimedia and security*, pages 107–116, 2009.

[95] C. Schmid, R. Mohr, and C. Bauckhage. Evaluation of interest point detectors. *Int. J. Comput. Vision*, 37:151–172, June 2000.

[96] H. T. Sencar and N. Memon. Overview of state-of-the-art in digital image forensics. *World Scientific Press*, 2008.

[97] A. Sharmir and O. Sorkine. *Visual media retargeting.* ACM SIGGRAPH Asia Courses, 2009.

[98] J. Shashank, P. Kowshik, K. Srinathan, and C.V. Jawahar. Private content based image retrieval. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[99] D. Simakov, Y. Caspi, E. Shechtman, and M. Irani. Summarizing visual data using bidirectional similarity. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008.

[100] M. Slaney and M. Casey. Locality-sensitive hashing for finding nearest neighbors. *IEEE Signal Processing Magazine*, 25(2):128–131, 2008.

[101] N. P. Smart and F. Vercauteren. Fully homomorphic encryption with relatively small key and ciphertext sizes. In *Public Key Cryptography, Lecture Notes in Computer Sciences*, volume 6056, pages 420–443. Springer, 2010.

[102] D. Song, D. Wagner, and A. Perrig. Practical techniques for searches in encrypted data. In *IEEE Symposium on Research in Security and Privacy*, pages 44–55, 2000.

[103] D. Stehl and R. Steinfeld. Faster fully homomorphic encryption. In *Advances in Cryptology - ASIACRYPT 2010*, volume 6477 of *Lecture Notes in Computer Science*, pages 377–394. Springer Berlin / Heidelberg, 2010.

[104] A. Swaminathan, Y. Mao, G-M. Su, H. Gou, A. L. Varna, S. He, M. Wu, and D. W. Oard. Confidentiality preserving rank-ordered search. In *Proceedings of the ACM Workshop on Storage, Security, and Survivability*, pages 7–12, Oct. 2007.

[105] A. Swaminathan, Y. Mao, and M. Wu. Robust and secure image hashing. *IEEE Trans. on Information Forensics and Security*, 1(2):215–230, June 2006.

[106] A. Swaminathan, M. Wu, and K. J. Ray Liu. Non-intrusive component forensics of visual sensors using output images. *IEEE Trans. on Information Forensics and Security*, 2(1):91–106, March 2007.

[107] P. C. Teo and D. J. Heeger. Perceptual image distortion. In *IEEE International Conference on Image Processing*, volume 2, pages 982–986, Nov. 1994.

[108] M. van Dijk, C. Gentry, S. Halevi, and V. Vaikuntanathan. Fully homomorphic encryption over the integers. In *Advances in Cryptology EUROCRYPT 2010*, volume 6110 of *Lecture Notes in Computer Science*, pages 24–43. Springer Berlin / Heidelberg, 2010.

[109] R. Venkatesan, S. M. Koon, M. H. Jakubowski, and P. Moulin. Robust image hashing. In *Proc. of IEEE Int'l Conf. on Image Processing (ICIP)*, volume 3, pages 664–666, Sept. 2000.

[110] Y.-S. Wang, C.-L. Tai, O. Sorkine, and T.-Y. Lee. Optimized scale-and-stretch for image resizing. *ACM Trans. Graph.*, 27:118:1–118:8, December 2008.

[111] Z. Wang and A. C. Bovik. A universal image quality index. *Signal Processing Letters, IEEE*, 9(3):81–84, Mar. 2002.

[112] Z. Wang and A. C. Bovik. *Modern image quality assessment.* Morgan & Claypool Publishers, March 2006.

[113] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. on Image Processing*, 13(4):600–612, 2004.

[114] Z. Wang and X. Shang. Spatial pooling strategies for perceptual image quality assessment. In *IEEE International Conference on Image Processing*, pages 2945–2948, Oct. 2006.

[115] D. Wangrattanapranee and A. Nishihara. Rigid image registration by using corner and edge contents with application to super-resolution. In *Proc. of the Conf. on Digital Image Computing*, 2008.

[116] A. B. Watson. Dctune: A technique for visual optimization of dct quantization matrices for individual images. In *Society for Information Display Digest of Technical Papers*, volume XXIV, pages 946–949, 1993.

[117] A. B. Watson, G. Y. Yang, J. A. Solomon, and J. Villasenor. Visibility of wavelet quantization noise. *Image Processing, IEEE Transactions on*, 6(8):1164 –1175, 1997.

[118] L. Wolf, M. Guttmann, and D. Cohen-Or. Non-homogeneous content-driven video retargeting. In *IEEE 11th International Conference on Computer Vision*, pages 1 –6, Oct. 2007.

[119] W. K. Wong, David W.-L. C., B. Kao, and N. Mamoulis. Secure kNN computation on encrypted databases. In *Proceedings of the 35th SIGMOD International Conference on Management of Data*, pages 139–152, 2009.

[120] L. Xie, G. R. Arce, and R. F. Graveman. Approximate image message authentication codes. *IEEE Trans. on Multimedia*, 3(2):242 –252, June 2001.

[121] M.-L. Yiu, G. Ghinita, C. S. Jensen, and P. Kalnis. Outsourcing search services on private spatial data. In *Proceedings of the 2009 IEEE International Conference on Data Engineering*, pages 1140–1143, 2009.

[122] J. Zobel and A. Moffat. Inverted files versus signature files for text indexing. *ACM Transactions on Database Systems*, 23(4):453–490, 1998.