ROBUST CONTENT-BASED COPY DETECTION AND INFORMATION THEORETIC INDEXING STRATEGIES

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

AHMET SARACOĞLU

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF DOCTOR OF PHILOSOPHY
IN
ELECTRICAL AND ELECTRONICS ENGINEERING

FEBRUARY 2015

Approval of the thesis:

ROBUST CONTENT-BASED COPY DETECTION AND INFORMATION THEORETIC INDEXING STRATEGIES

submitted by **AHMET SARACOĞLU** in partial fulfillment of the requirements for the degree of **Doctor of Philosophy in Electrical and Electronics Engineering Department, Middle East Technical University** by,

Prof. Dr. Gülbin Dural Ünver
Dean, Graduate School of **Natural and Applied Sciences** _____

Prof. Dr. Gönül Turhan Sayan
Head of Department, **Electrical and Electronics Engineering** _____

Prof. Dr. A. Aydın Alatan
Supervisor, **Electrical and Electronics Engineering, METU** _____

Examining Committee Members:

Prof. Dr. Uğur Halıcı
Department of Electrical and Electronics Engineering, METU _____

Prof. Dr. A. Aydın Alatan
Department of Electrical and Electronics Engineering, METU _____

Prof. Dr. Gözde Bozdağı Akar
Department of Electrical and Electronics Engineering, METU _____

Prof. Dr. Nihan Kesim Çiçekli
Department of Computer Engineering, METU _____

Assoc. Prof. Dr. Selim Aksoy
Computer Engineering Department, Bilkent University _____

Date: _____

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last Name: AHMET SARACOĞLU

Signature :

ABSTRACT

ROBUST CONTENT-BASED COPY DETECTION AND INFORMATION THEORETIC INDEXING STRATEGIES

Saracoğlu, Ahmet

Ph.D., Department of Electrical and Electronics Engineering

Supervisor : Prof. Dr. A. Aydın Alatan

February 2015, 107 pages

Today, 100 hours of video is uploaded *every minute* to YouTube. By the end of 2015, 500 billion hours of video will be viewable from wide range of sources such as on demand video, Internet-based television and social networks. As a result important and unavoidable problems arise; management of the copyrights, numerous duplicates and content discovery. Obviously these problems may generate tremendous loss for content owners and broadcasting/hosting companies while diminishing user satisfaction. Accordingly, efficient duplicate video detection can be utilized for the solution of the aforementioned problems. Content Based Copy Detection (CBCD) emerges as a viable choice against active duplicate detection methodology of watermarking.

In this thesis, building blocks of a content-based copy detection system are investigated. A novel spatio-temporal global representation is initially proposed that exploits visual features independent of the spatial information. This system is improved by a local interest point-based detection pipeline and it is shown to outperform global representation approaches through extensive simulations. On the other hand, it is observed that accuracy of local feature approaches is often limited by the presence of uninformative and redundant features extracted from the frame. Moreover, at large scale index size and corresponding amount of memory becomes a significant bottleneck. In order to decrease the index size while increasing the discriminativeness of the reference feature database, a novel information theoretic indexing method is

proposed and improved further by the introduced entropy estimator. This estimator is shown to yield more robust results compared to naïve frequentist techniques. Furthermore, in comprehensive experiments using the proposed method, it has been shown that only with a fraction of the reference features same detection performance and even for some transformations 0.00 Normalized Detection Cost Rate (NDCR) is achieved, which was not possible previously with full indexing. Extending this foundation, another method to exploit distributions of local features in a temporal volume is also provided. With this temporal approach, for most of the transformations 31% to 83% improvement on NDCR is observed. Finally, in order to capture the dependence of multiple features in a given frame fundamentals of *interaction information* is discussed and a visual phrase representation for content-based copy detection is introduced. Experimental evaluations show that the proposed visual phrase representation and multivariate feature selection approaches are competing with the state-of-the-art.

Keywords: Content-Based Copy Detection, Spatio-Temporal Global Features, Local Features, Bag-of-Visual Words, Information Gain, Entropy Estimation, Visual Phrases, Interaction Information

ÖZ

İÇERİK TABANLI VİDEO KOPYA SEZİMİ VE BİLGİ TEORİSİNE DAYALI DİZİNLEME STRATEJİLERİ

Saracoğlu, Ahmet
Doktora, Elektrik ve Elektronik Mühendisliği Bölümü
Tez Yöneticisi : Prof. Dr. A. Aydın Alatan

Şubat 2015 , 107 sayfa

Günümüzde her bir dakikada 100 saatlik video YouTube'a yüklenmektedir. 2015'in sonu itibariyle 500 milyar saat video, Internet televizyonculuğu, ısmarlama video ve sosyal ağlar gibi çeşitli kaynaklardan seyredilebilir olacaktır. Bunun sonucunda da telif haklarının yönetimi, sayısız kopya ve içerik keşfi gibi önemli ve kaçınılmaz sorunlar ortaya çıkacaktır. Elbette, bu problemler içerik sahipleri ve video içerik veritabanı sağlayıcıları için muazzam kayıplara sebep olurken kullanıcıların memnuniyetini de azaltacaktır. Bu bağlamda, bahsedilen bu problemlerin çözümünde etkili kopya video seziminden yararlanılabilir. İçerik tabanlı kopya sezimi, aktif kopya sezimi yöntemi olan gizli damgalamaya karşı geçerli bir seçenek olarak ortaya çıkmaktadır.

Bu tezde, bir içerik tabanlı kopya sezimi sisteminin yapıtaşları incelenmiştir. Öncelikle, uzamsal bilgiden bağımsız görsel öznitelikleri kullanan özgün uzamsal-zamansal global temsil yöntemi önerilmiştir. Bu yöntem, yerel ilgi noktaları temelli sezim yöntemi ile geliştirilmiş ve global temsil yaklaşımlarına üstünlüğü kapsamlı benzetimler ile gösterilmiştir. Bununla birlikte yerel öznitelik yaklaşımlarının etkinliği video içerisinden çıkarılan ilgisiz ve gereksiz özniteliklerden dolayı sınırlanmaktadır. Ayrıca, büyük ölçekte dizin büyüklüğü ve denk düşen hafıza kayda değer bir darboğaz oluşturmaktadır. Dizin büyüklüğünün azaltılması ve bu sırada referans öznitelik veritabanının ayırt ediciliğinin arttırılması için bilgi teorisine dayalı indeksleme yöntemi öne-

rilmiştir ve sunulan entropi kestirimi ile geliştirilmiştir. Bu kestiricinin sade frekansçı tekniklere göre daha gürbüz sonuçlar ürettiği gösterilmiştir. Ayrıca, kapsamlı deneyler önerilen yöntemin referans özniteliklerin yalnızca bir bölümünün endekslenmesiyle aynı kestirim başarımı ve hatta bazı değişimlerde tam indeksleme ile daha önce elde edilemeyen 0.00 Düzgelenmiş Kestirim Maliyet Oranı (DKMO) elde edilmiştir. Bu temel üzerinde genişletilerek, zamansal hacim içerisindeki yerel özniteliklerin dağılımının kullanılması için bir yöntem geliştirilmiştir. Bu yaklaşım ile çoğu değişim için zamansal bilginin kullanılmadığı yönteme göre DKMO'da %31 ile %83 arasında iyileşme gözlemlenmiştir. Son olarak, bir video karesi içerisindeki birden fazla özniteliğin bağımlılığının işlenebilmesi için *etkileşim bilgisinin* esasları tartışılmış ve içerik tabanlı kopya sezimi için görsel öbek temsili yöntemi getirilmiştir. Deneysel değerlendirmeler önerilen görsel öbek gösteriminin ve çok-değişkenli öznitelik seçimi yaklaşımlarının en gelişkin yöntemler ile rekabet edebildiğini göstermiştir.

Anahtar Kelimeler: İçerik Tabanlı Kopya Sezimi, Uzamsal-Zamansal Global Öznitelikler, Yerel Öznitelikler, Bilgi Kazanımı, Entropi Kestirimi, Görsel Öbekler

To my family and my love Pınar.

ACKNOWLEDGMENTS

This has been a *very long* and stormy journey. While in some rare moments it was enveloped in a disheartening fog of uncertainty, it has been a tremendously satisfying experience to see it develop from the scratchpad to the whole you see now before you. I hereby would like to take the time to thank a few of the many people who helped me to finish this work, apologizing in advance to anyone that I might have forgotten.

First of all, I would like to express my deep and sincere gratitude to my supervisor, Prof. Aydın Alatan. Without his guidance, encouragement and patience I cannot imagine to finalize this work. His support and friendship throughout the years made it all better. It has been an honor to be his Ph.D. student.

It should not be left unmentioned that I would not have enjoyed my professional life as much as I did, without my partners Alphan Es, Banu Acar, Serdar Gedik and Ziya Kadıoğlu from our startup Kuartis. I know it is once in a lifetime experience and I am very glad to share all the joy and pain with them. Also, I am grateful to Prof. Aylin Tarcan and Prof. Ayşe Nur Ecevit for sharing their wisdom and encouragement.

I would also like to thank my former colleagues at TÜBİTAK Space Technologies Research Institute for the friendly environment they created. I have always enjoyed the energy, freedom and the creativity during the unusual times we spent together. It has really been a great privilege to work with such an excellent team. I am also grateful for our fruitful teamwork together with Burak Özkalaycı and Emrah Taşlı during the qualification exam. I would like to also extend my thanks to my friends Bahadır Turhan, Berker Loğoğlu and Serkan Soydan for their faith in me.

Finally, I would like to thank my family and my wife for their love, understanding and support. My mother and father have provided me with a truly exceptional level of intellectual and emotional support during my whole life and I am humbly grateful for that. And most of all, I am indebted to my wife, Pınar; my best friend, our team coach and a genuine guide when I am lost. She made my days with her true affection and I have my deepest appreciation and gratitude as well as my love to her. Nothing in a simple paragraph can express the feelings I have for them. It goes without saying that without them, this thesis would not have been possible.

TABLE OF CONTENTS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

CHAPTER 1

INTRODUCTION

Astounding increase in data transfer rates, skyrocketing capacities in digital data storage, adoption of more efficient multimedia coding standards and flooding of camera phones into everyone's pocket are just a few developments among the myriad technological advances in recent years that have irrevocably changed the lives of millions. New sectors have born, some others have descended from their zenith and many others have evolved. One such sector, in the context of this work, is the broadcasting. "Broadcast Yourself[1]" might be the epitomizing line for the new face of the sector. Mutually, a new catalyzing force has born: "video-hosting services."

At the moment YouTube, Vimeo, LiveLeak and similar services are parts of our daily lives. As the amount of digital media in these sources increase exponentially (in August 2006 YouTube was hosting about 6.1 million videos [1] and as of January 2015, 100 hours of video is uploaded in *every minute* to YouTube [2]) two crucial and unavoidable problems arise; management of the copyrights and numerous duplicates. Obviously both problems may generate tremendous loss for content owner and equally to the hosting companies.

This said, another important issue faced today is that how viewers will find what they desire in vast sea of content. Identically, this is also the problem of content providers. By the end of 2015, 500 billion hours of video will be viewable [3] from wide range of sources such as social networks, Internet-based television and on-demand video. It is clear that it will take a sophisticated multimedia analysis capability discover and lo-

[1] "Youtube – Broadcast Yourself," [Online]. Available: http://www.youtube.com [Accessed: December 21, 2013].

cate the desired content/consumer. Although it would be an exaggeration to state that duplicate detection is the key for meaningful video search, it is not a stretch to imagine that with effective duplicate detection user satisfaction would improve considerably. In this case, not only the text search results but also duplicate videos that might have been changed in someway can be associated and presented with (or removed from) the search results. Moreover, missing metadata of duplicate videos can be completed by tag propagation from associated videos. Furthermore, video-hosting companies can also optimize data storage and bandwidth by collecting duplicate videos to same cluster of servers.

As the television landscape evolve to become smarter and more connected, discovery of content turned to be vital for this domain as well. Moreover, television viewer statistics can be easily collected and analyzed nowadays. However, associating viewer to the content remains an important and valuable hurdle to solve. As the viewed content identified many value propositions can be offered to both viewers and content providers. To viewers/consumers pop-up information and personalized recommendations can be provided on the other hand accurate engagement, retention and conversion rate statistics can be supplied to content providers.

For the solution of the aforementioned dual problems there are two main approaches; passive methods and active methods i.e.: watermarking. However, watermarking has two significant limitations. First, since watermarks must be introduced into the original content before copies/duplicates are made, it cannot be applied to content which is already in circulation. Second, the degree of robustness is not adequate for some of the attacks that we encounter frequently. Passive detection methods, on the other hand, try to directly detect copyright infringements and duplicate videos by comparing questioned data against a database. This approach can be thought as a complementary technology to watermarking which provides a solution to the two problems mentioned above. The primary idea of this approach can be interpreted as the media being the watermark itself. That is, the media (image, video, audio) contains enough unique information to be able to detect copies. The main difficulty of passive detection methods is that the videos are not supposed to be identical. Brightness or contrast enhancement, compression, noise, bandwidth limitation, mixing with unrelated audio, overlay text or geometric transformations can be observed on videos

which yield highly modified duplicate video signals. At this point it is necessary to provide related definitions,

Definition A video V_i is a copy of another video V_j if it is generated after some tolerable transformations on V_j which is called the reference video. Typical transformations include addition, deletion, modification (of aspect, color, contrast, encoding) and camcording. (Figure 1.1)

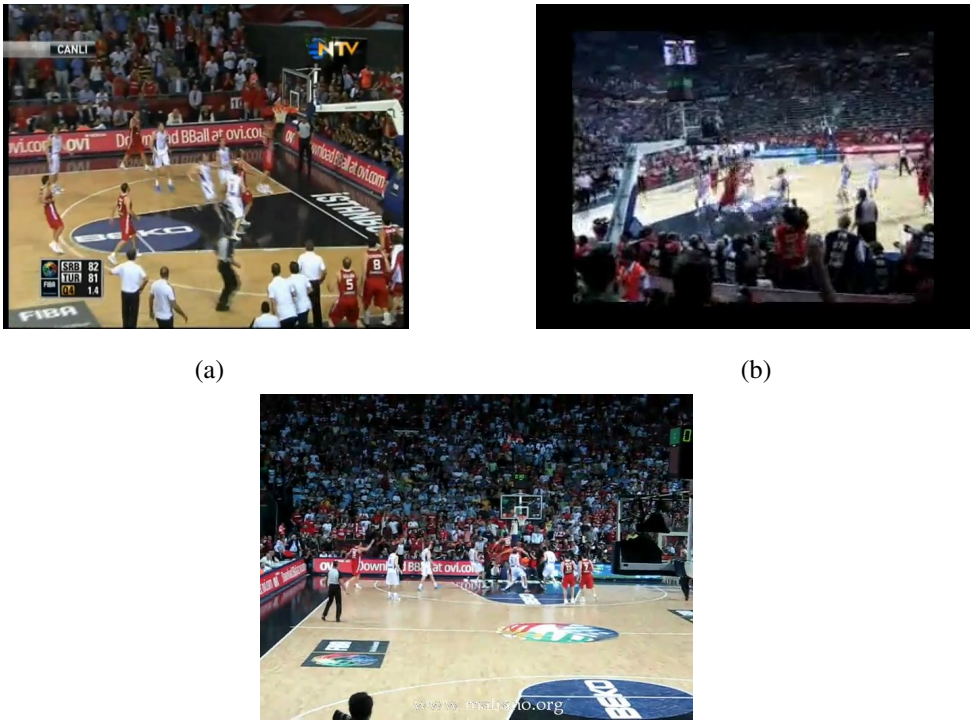
(a) (b)

Figure 1.1: Example frames from copy videos with heavy re-encoding (a) and picture-in-picture (b) transformations.

In the light of this definition, the term Content-Based Copy Detection (CBCD) is coined in the literature to denote the passive duplicate detection. However, it should be noted that multiple terms are referred to the same concept as CBCD. Most frequently used is the Detection of Near Duplicate Video or equivalently Near Duplicate Video Detection.

Although in this study CBCD problem domain is investigated there are related definitions that can be revealing in the concept of duplicate detection such as *near-duplicate* as defined in [4].

Definition A video V_i is a near duplicate of another video V_j if V_i and V_j are highly similar content but appear differently due to acquisitions and transformations. By acquisition, different camera, viewpoint and light conditions are meant and transformations are same as in copy definition. (Figure 1.2)



(a) (b)

(c)

Figure 1.2: Near-duplicate example frames taken from three different YouTube videos (a) Uploaded from TV broadcast. (b) Uploaded from camera phone with heavy encoding and black bars. (c) Uploaded from camera phone with better encoding conditions than (b) but with text insertion and shaky capture.

It can be easily seen from this definition; near duplicate is a superset of copy definition. Although near duplicate definition contains ambiguity on the similarity of content, it has an important, however small, difference which extends the potential of CBCD. As defined, copies must originate from the same reference video. Thus, two videos with different capturing conditions cannot be copies but can be near duplicates. For example, breaking news on the same scene/event taken by different broadcasting agencies or better yet taken by the camera phones of ordinary people at the scene are near duplicates.

Definition A video V_i is a semantically duplicate of another video V_j if V_i and V_j are from the same semantic concept (e.g., a beach scene) with varying viewpoints, sizes, appearances and camera motions. Matching semantic concept can occur under different illumination, appearance, and scene settings. (Figure 1.3)

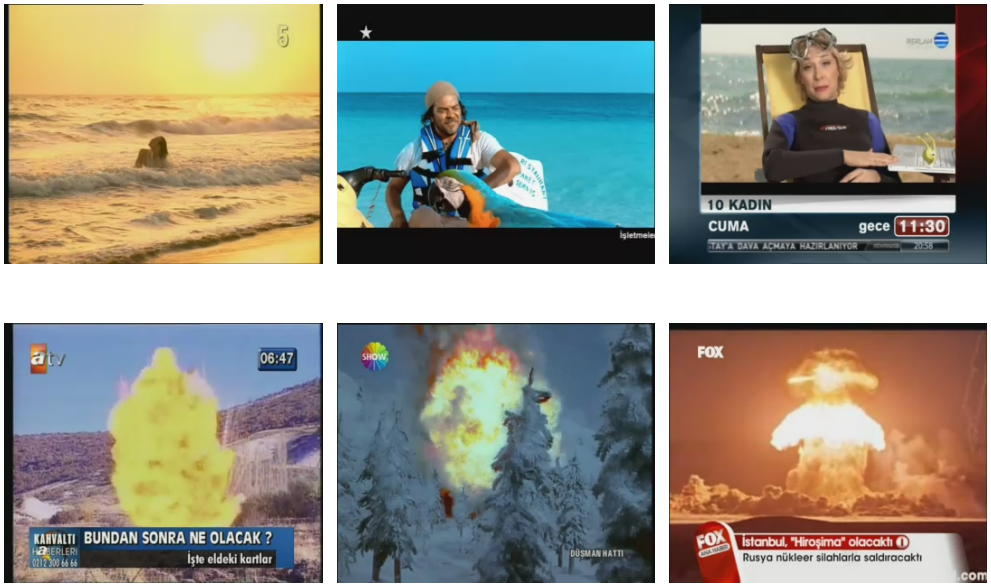


Figure 1.3: Example frames from semantically duplicate (row-by-row) video segments.

Basharat et al. adopt this definition and propose a method to detect semantically duplicate videos in [5]. Semantically duplicate video definition is also closely related to semantic concept detection. However, contrary to concept detection, semantically duplicate detection in general does not utilize a machine learning scheme explicitly; instead a query-based detection is implemented.

It can be seen from the previous definitions that there are considerable differences with duplicate video definitions. In the following table (Table 1.1) a summary of these definitions with related literature cues are given.

Table 1.1: Comparison of copy, near-duplicate and semantically duplicate definitions.

	Definition	**Authors**
Copy	Videos are from the same source but by some kind of transformations differ from each other. Transformations may contain addition, deletion, modification (of aspect, color, contrast, encoding) camcording and etc.	Cherubini et al. [6], TRECVID [6], Joly et al. [7], Law-to et al. [8]
Near-Duplicate	Videos are from the same scene but appear differently due to acquisitions and transformations. Acquisition disparities contain different camera, viewpoint and light condi-tions. Transformations contain addition, deletion, modification (of aspect, color, contrast, encoding) camcording and etc.	Jaimes et al. [9], Rossi et al. [10], Satoh et al. [11]
Semantically Duplicate	Videos that are from semantically same scenes (e.g., an explosion) varying viewpoints, sizes, appearances and camera motions. The same semantic concept can occur under different illumination, appearance, and scene settings.	Cherubini et al. [6], Basharat et al. [5]

1.1 Scope of Thesis

From previous *duplicate/copy* definition and related discussions it is apparent that a robust copy detection method has many important benefits. The building blocks of a conceptual system can be seen in Figure 1.4. At the fingerprint extraction step a huge amount of content is mapped to a lower dimensional space for effective representation of the information conveyed. In this respect, fingerprints can be characterized as compact and descriptive. The compactness of the fingerprints facilitates and accelerates the query process and decreases the required indexing storage. The descriptiveness of the index entails the discriminative power, which enables the discrimination between different contents and robustness under certain transformations. Indexing, on the other hand, enables fast lookup thus accelerating the matching process. Finally by matching the similarity of two videos/fingerprints are inspected.

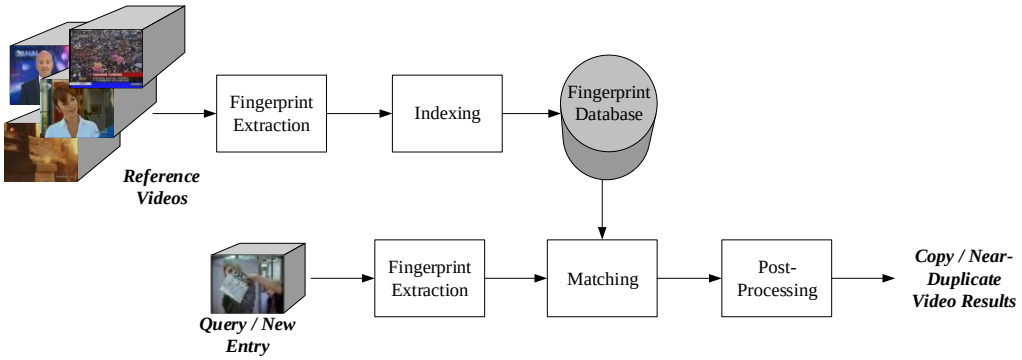


Figure 1.4: Conceptual design of a copy detection system.

In this thesis aforementioned building blocks of a content-based copy detection system are investigated. A novel spatio-temporal global representation is proposed. Also a local interest point-based pipeline is developed based on the literature. Furthermore, in order to increase the discriminativenss of the feature database an information theoretic indexing method is proposed. Extending this foundation a method to exploit temporal distributions of local features and multiple variables are proposed. Finally, visual phrase representation for content-based copy detection is introduced.

1.2 Outline of the Thesis

Chapter 2 is devoted to the state-of-the-art approaches for content-based copy detection and in-depth analysis of related work will be given in Section 2.3. In Chapter 3, two detection methods from different approaches will be proposed for the solution of content based copy detection. Section 3.1 will introduce a novel global spatio-temporal approach whereas in Section 3.1.1 experimental results of this method will be provided. Furthermore, in Section 3.2 a local interest point-based method will be presented. Experimental analysis will be conducted and compared with related work from the literature in 3.2.6.

In Chapter 4, an information theoretic feature selection and indexing method will be introduced by first discussing fundamentals of feature selection. After proposing a mutual information based algorithm, drawbacks of Naive entropy estimator will be explained. And in Section 4.4 a method will be given in order to improve the entropy estimate and merits of this approach will be shown on synthetically generated data in 4.4.1. On the other hand, evaluation of information theoretic feature selection will be conducted on TRECVID 2009 and 2010 CCD dataset in Section 4.5. Lastly, copy detection performance will be tried to be improved by introducing a method to exploit temporal distributions of local features.

Chapter 5 will extend mutual information definition to multiple variables. In order to utilize this perspective a *visual phrase* indexing approach will be proposed in Section 5.1. Moreover, experimental evaluation of visual phrase indexing and corresponding feature selection methods will be investigated in Section 5.2.

Finally, Chapter 6 will summarize the studies conducted in this thesis. Concluding remarks and possible future research paths will be discussed under the light of the contributions.

CHAPTER 2

LITERATURE REVIEW

In this chapter related work on copy and near-duplicate detection is examined. Unfortunately, there are not many publications on near-duplicate detection as defined in the Chapter 1. One such work addressing near-duplicate detection in consumer photography libraries is published by [9] and another near duplicate detection method on news video is proposed by Satoh et al. in [11] and Takimoto in [12]. And most recently, Revaud et al. [13] proposed an approach to retrieve videos representing the same event. On the other hand, there are quite a number of works addressing content-based copy detection in the literature. First section investigates different representation approaches from the copy-detection literature. Also, related work on indexing and matching approaches is given in the subsequent section. Afterwards, in Section 2.2 detailed information on some of the works from literature is provided. Lastly, different evaluation schemes and publicly available datasets are summarized in Section 2.4.

2.1 Representation

Representation of reference videos and queries – practically queries can be seen as another video – according to the fingerprints used can be categorized as global or local. In global representation methods, smallest unit of a video – which is most of the time a keyframe – is represented by a global descriptor such as ordinal measure [14, 15, 16], color shift and centroids [17], color histograms and etc. Furthermore motion activity of consecutive frames is also utilized for efficient representation of videos. Compar-

ison of global descriptors of the video based on motion, color and spatio-temporal distribution of intensity is given in [18] by Hampapur and Bolle. And combination of different low-level representations [19] is another approach from the literature as a global representation. In general, global features discussed in copy/near-duplicate detection literature have simple computational complexity and robust to simple transformations (e.g., brightness and gamma correction) however; they are more affected by spatial transformations (e.g., frame resizing, insertion of pattern and picture-in-picture) than local representation techniques.

Due to obvious limitations of global representation of a video in the domain of copy detection most of the state-of-the-art methods employ local descriptors such as Scale Invariant Feature Transform [20] and Speeded-up Robust Features [21]. Methods employing local descriptors [7, 8, 22, 23, 24] have shown very high performance in terms of recall and localization precision. Using interest points is mainly motivated by the observation that these points provide detailed representation of a frame while limiting the redundancy between the features. Furthermore, extracting descriptors that are invariant to some distortions (e.g., scale and noise) around interest points improve matching rate of the features enormously.

Since for a given frame many local features, in fact several hundreds of high dimensional vectors, are extracted it is impossible to store and process them in memory and even in disk space. Furthermore, as the number of features increase, the redundancy on the features increases thus decreasing the robustness of the fingerprints. Bag-of-features or bag-of-words model is an established method employed exactly in the same situations for document analysis, object detection and recently on semantic concept detection. With this method, descriptors are quantized into visual words with a clustering algorithm, which is frequently the k-means. Then a frame or any other video unit is represented by the frequency histogram of visual words obtained by assigning each descriptor of the image to the closest visual word. This approach is employed in [15] on both ordinal feature and SIFT descriptors effectively. Furthermore, in [24] bag-of-features is used with Hamming Embedding scheme in order to make the distance between visual word frequency features more significant by using a more informative/finer representation. Moreover, instead of hard voting of previously discussed methods, each descriptor is assigned to several closest visual words in [23]

and [25]. One drawback of this bag-of-features approach is that the visual words/-cluster centers have to be computed offline and in order to span the feature space as effective as possible many cluster centers are needed; e.g. 200000 cluster centers in [23]. Furthermore, such a high number of cluster centers need a much higher number of features from wide variety of data.

Although selection of representation technique is a very important design criterion, there is also the decision of video units where the representation will be implemented. This unit can be keyframes sampled uniformly, sequence of frames on a sliding window, shots and etc. Most of the methods in the literature adopt using keyframes as the basic element of the detection methods and extract necessary features from single frames. Of course, as the number of elements used for representation increases discriminative power of the representation increases, on the other hand computational complexity and storage demand increases accordingly. This said, besides extracting fixed number of frames per second, keyframes around the shot boundaries are also employed in the literature [7, 23, 26]. This way the number of frames extracted for a fixed interval is decreased while also the redundancy between frames is decreased. However, in the presence of strong deformations using keyframes might hinder the quality of shots and hence the features obtained from queries. A simple solution is to sample query videos differently from reference videos. Query video can be sampled densely and uniformly while reference videos are sampled from shot boundaries sparsely. Another strategy for frame sampling can be the selection of frames according to their representative power. One such method has been proposed in [27]. Zhou et al. propose a shot-based near-duplicate video detection method. For each shot, keyframe and an appropriate number of neighboring frames, which are selected based upon similarity to the keyframe, are appointed as the most representative set of the shot. The similarity of two frames is measured by computing the distance between the Color Histogram Descriptors of the frames. In the next step, furthest Voronoi Diagram method is utilized in order to sample the frames. By using such a strategy a diverse distribution is achieved. Finally, from the set of frames local descriptors are extracted and pruned according to their observation frequency.

In another work, [28], authors propose an adaptive frame selection method based on Pearson's correlated coefficient (PCC) which measures the correlation between each

pair of successive frames in the RGB plane. The shot is then partitioned into several fractions at where the correlation coefficients decrease abruptly so that wide diversity is achieved by each fraction. Then reference frame set is produced by selecting the frames from the most representative fractions to ensure wide coverage. In addition to these adaptive/dynamic frame selection methods Chen et al. propose selection of most representative frame of a given interval by computing the number of Hessian points, [29].

Until now one important aspect of video is left out from the representation discussion that is the temporal content of the video. Generally, copy detection methods in the literature benefit from temporal information after the matching features/frames are done [7], [25]. Outliers from initial match set are removed and time localization is improved with methods such as temporal shift-modeling, m-estimate of shifts or simple voting scheme. One interesting work is published by Willems et al. in [22]. A local spatio-temporal feature extraction approach for copy detection problem is proposed. Spatio-temporal features are extracted in the sense much like SURF however, temporal dimension is also incorporated by using a 3x3 Hessian matrix and three-dimensional box filters. Thus, robustness of local features is improved. Furthermore, number of features extracted from unit time is decreased hence; the storage and indexing conditions are improved. However, authors chose to employ this method on the segmentation result of shot detection, which might decrease the localization power and accuracy of the method.

Another work worth mentioning that incorporates temporal information, however indirectly, is [8] by Law-To et al. Method detects local interest points and extracts their descriptors much like [7]. In addition to descriptor vectors, trajectories of interest points are used in the representation. The trajectories enrich the local description with a spatial and temporal behavior of interest points while the redundancy of local description is reduced. Trajectories are computed by using well-known Kanade-Lucas-Tomasi (KLT) feature tracker [30]. Low-level properties of trajectories such as the start and end location/time are then analyzed in order to label the trajectories into behaviors such as persistent, still, moving and etc. Finally point descriptors, low-level properties of trajectories and the labels are used in conjunction for the matching. However, trajectory labels are only incorporated by heuristic voting rules. In

another work, Satoh et al. [11] incorporate the trajectory inconsistencies to match near-duplicate news video shots. Trajectories are also computed from interest points by KLT as in Law-To's approach. However, trajectory inconsistencies are obtained from spatio-temporal patches around initially detected point and its trajectory.

Table 2.1: Review of previously developed representation techniques for video copy detection methods.

Category	**Local**	**Global**
Spatial	Harris [7, 26], SIFT [31, 23], PCA-SIFT [32], F-SIFT [33], SURF [34]	Ordinal [16, 14], Color Shift [17], MPEG-7 Visual Descriptors [35]
Spatio-temporal	Harris Trajectories [8], KLT [11], 3D-SURF [22], STIP [36]	Multi-modal [19, 37], TIRI-DCT [38], Ordinal [39, 40], Motion Direction [18], NMF [41]

2.2 Indexing and Matching

Efficient similarity search in large databases is an important issue in all content-based retrieval schemes. Especially with the local representation approach the size and number of descriptors per frame, the cost of matching and tracking, together with the very large number of videos scalability becomes an important challenge. In its essence, the similarity paradigm is to find similar documents/videos by searching similar features in a database. Generally, the distance between features is used to perform k-nearest neighbor searches on the database. Although there are multi-dimensional index structures such as R-trees or KD-Trees, unfortunately their performance degrades considerably when dimensionality increases and it has been shown [42] that for real-world data they are not more efficient than the brute-force search. Hence, it is obvious that indexing and retrieving of high-dimensional data is very challenging because of the curse of dimensionality. One approach to handle this curse of dimensionality is to employ an approximate similarity search hence trading localization quality for time. The underlying idea in approximate similarity search is to find similar features with a very high probability which is not 1.

Throughout the recent years, several methods have been proposed for approximate

index searches. In that sense, Locality Sensitive Search (LSH) [43] is such an approximate high-dimensional similarity search scheme, which is able to find matches in sub-linear time. It tries to solve the curse of dimensionality problem by hashing the descriptors through a series of projections onto random lines and concatenating the results into a single hash. It has been shown that the collision probability of such hashes is much higher for vectors that are near to each other. And generally several different hash tables are combined in order to improve the probability of finding the correct matches. On the other hand, in [32] Ke et al. were the first to implement LSH as an on-disk database in the context of near-duplicate image retrieval. In order to minimize disk-access, queries are combined in batches. By sorting the query batches and the on-disk database, single sequential scan on the database is enough per batch as first the smallest hash value will be encountered if it exists in the database. As a second step, matched descriptors are checked for outliers since LSH matches with respect to L1 distance. Also Willems et al. [22] employ disk-based LSH indexing scheme with some modifications. First modification is that they prefer p-stable LSH indexing in order to work directly with L_2 norm. Furthermore, the list of buckets in the database is divided into fixed-size blocks and offset to each block. Thus, these offsets are utilized instead of sequential scan in order to compute the next possible block that could contain a matching hash.

On the other hand, Joly et al. in [26] proposed a distortion-based probabilistic similarity search where the feature space is partitioned to relatively small descriptors using a Hilbert space filling curve. It is based on the principle that two points that are close on the Hilbert curve remain close in the original N-dimensional space. Actually, it is a static method in which dynamic insertions or deletions are not possible. The space-partitioning is induced by regular splits of a Hilbert curve and by a simple dichotomic search method the closest point to a derived key is found in the database. When a range query overlaps strong discontinuities of the Hilbert's curve, it is divided in several sub-queries which are not adjacent on the curve. A local sequential scan is then performed for each sub-query. Although the proposed scheme could handle over 40.000 hours of video, an improvement using Z-grid [44] allows for the indexing of 120.000 hours of video. Z-grid based indexing is proposed by the observation that for probabilistic retrieval it is not required to have neighbors in the description space to

also be neighbors in the index. Furthermore, when the partitioning depth exceeds the dimensions of the description space, not all the cells are partitioned along the same dimensions.

Although LSH has been implemented successfully in many applications, methods that consider the distribution of the features such as Vocabulary Tree [45] have been shown to outperform LSH. Consequently, in the literature clustering-based approximate search methods have been proposed with the intention of achieving substantial improvement on both storage and computational time. For example, in [45] Vocabulary Tree (VT) method has been proposed, which is based on hierarchical k-means clustering. And VT has been employed successfully in many applications [29, 15, 14]. However, it should be noted that clustering-based method needs an offline processing in order to compute the cluster centers. And clustering process has two limitations; number of samples and computation time. In order to represent and cluster feature space effectively many features are needed beforehand. Thus, recently quantization-based approaches have been proposed in the literature [23, 25, 46] in order to handle memory and clustering constraints. Barrios et al. [47] introduced a pivot-based indexing method to perform approximate nearest neighbor search for the video segments. In the following section, detailed analysis of the related work is given.

2.3 In-depth Analysis of Related Work

Joly et al. are one of the first researchers adopted local descriptors for copy detection in their work [7]. In their approach first a keyframe is extracted from video stream by intensity of motion. Afterwards, Harris corner detector is employed in order to locate interest points from previously extracted keyframes which is followed by fingerprint computation around the interest points by second order differential decomposition of the graylevel 2D image signal. Later, all of the descriptors are indexed with their time code using a Hilbert space-filling curve. Thus, enabling fast k-nearest neighbor computation. Finally, from the matched keyframes robust time localization is obtained by maximum likelihood estimation. On the experimental dataset (with attacks such as additive noise, resizing, vertical shift and gamma variation) and on real world dataset (with attacks such as resizing, broadcast artifacts, frame encrusting) method performs

with very minimal false alarm and very high recall rate. This is partly due to the mild attacks and partly due to using very high number of descriptors for the data. On the other hand Joly et al. in [26] proposed an approximate similarity search technique in which probabilistic selection of feature space regions on the database is based on the distribution of feature distortion instead of database distribution. By using such an approach search time at constant precision and recall has been improved astonishingly (more than 45 times) when compared to exact range queries. After candidate video sequences are obtained by simple voting scheme spatio-temporal consistency of matched descriptors on the sequences are analyzed. Spatio-temporal model characterizes the tolerated transformations such as resize, rotation, translation in space or time and slow/accelerated motion. Model parameters are estimated by RANSAC.

Kim et al. propose a copy-detection scheme based on the well-known ordinal measure that is robust to the brightness, color and frame resize transformations [40]. Each image frame is first partitioned into 2x2 by intensity averaging, and the partitioned values are stored for indexing and spatial matching. Furthermore, temporal variation of each block of the stream is utilized as a temporal fingerprint. Finally approach combines spatial matching of ordinal signatures of each frame and temporal matching of temporal signatures to detect copies.

Chiu et al. in [14] have applied bag-of-features approach to copy detection problem with both ordinal features and SIFT features while bag-of-features of query is searched over the database by a sliding window scheme. Thus authors adopted histogram pruning to improve the search time. Authors further improved their approach in [14] by incorporating a finer matching method after finding a coarse detection result by the previous method. Fine matching scheme involves a similarity matrix computed between every frame of query and candidate frames by computing the histogram intersection of two frames. The temporal consistency is investigated by line detection on this matrix using Canny edge detection and Hough Transform, which of course increase the computation complexity incredibly. In [48], however, for fine matching authors adopt a graph matching technique. On the hand in a recent work [29] same author plugs an efficient heap manipulation method instead of histogram pruning in order to generate each window's min-hash signature.

Takimoto et al. [12] proposed a robust near-duplicate news video clustering method based on the detection of flash light patterns. The underlying assumption is that camera flashes are often used in impressive scenes such as public speeches of political figures thus; shots containing identical flash patterns are near-duplicates of each other. Method first detects shots containing flash bursts and afterwards flash patterns are compared in order to cluster the scenes. Flash bursts are identified simply by rapid average luminosity changes. In order to decrease the false alarm of flash detection a validation step from frames and temporal occurrences is implemented. Finally, flash patterns are compared frame by frame by considering the temporal offsets. Authors have examined the performance of this method in a very limited dataset with limited content (only news footage). However, it is interesting to see a less travelled road to be taken.

On the other hand in [11], near duplicate shot detection on news video is addressed. Satoh et al. propose a method based on matching temporal pattern of discontinuities obtained from trajectories of interest points. Authors start from the idea that if two shots are taken at the same scene but from different viewpoints a certain fraction of the shots will match with the common temporal offset. Furthermore, as far as the interest points are on a rigid body and the object moves under shaky motion such as nodding head, the temporal discontinuity patterns of two shots from different viewpoints will agree. In the approach interest point extraction is done by SIFT and points are tracked with KLT. Method is evaluated over a small dataset partly because of the difficulties to find near-duplicate video shots. However, as expected, performance of the method is not close to the state-of-the-art copy-detection performance on much larger datasets.

Although it is not directly related to near duplicate detection Fu et al. [49] propose a multiview video summarization method in which, a spatio-temporal shot graph is used for representation of videos. On the other hand, proposed method incorporates low-level and high-level shot importance schemes which consist of computing a score by fusing color histogram and wavelet coefficients as low-level features in addition to faces as high-level features. Furthermore, correlation among multi-view shots considered according to temporal adjacency, visual similarity and semantic correlation. Finally, random walk with multi-objective optimization shot clustering is performed. Authors do not aspire to detect near duplicate videos with numerous transformations

in a large video database, however using content correlations in multiple views are significant. Method coupled with the experimental data is much closer to surveillance applications with loose time constraints.

In another related work Chen et al. [29] proposes a mobile search application which is centered around snapping a photo with a mobile device of a video playing on a TV screen to automatically retrieve and stream the remainder of the video to the mobile device. In their approach a selection method over a short temporal window around a keyframe is employed in order to reduce the temporal redundancy and dynamically select a feature-rich query frames. Although an interesting mobile search application is proposed method, it does not have a smart frame selection method. Method only selects the frame with the highest number of Hessian points in a fixed window. SURF is employed for local feature extraction and features are stored in an inverted Vocabulary Tree. The evaluation dataset contains 2000 YouTube videos and 50 queries, on which high performance is obtained.

Douze et al. proposed a comprehensive image-based copy detection method in [23] which addresses the problem caused by strongly deformed videos. Like most of the recent approaches, method employs local descriptors for the representation of the video. Specifically authors use Hessian-Affine region detector for interest point detector and center symmetric local binary pattern as the descriptor. Bag-of-features approach is used to quantize the descriptors to have a compact representation, which is also refined further by Hamming Embedding method. Reference video representations are then stored in a structure similar to inverted file. After matching query descriptors on the indexed reference database, a spatio-temporal model is estimated from the matching keyframes. Experiments are conducted on TRECVID 2008 CCD dataset.

Non-negative matrix factorization (NMF) based subspace representation is utilized in [41] for the global visual features obtained from the video content. NMF is applied over a sequential block of volumes which are circularly cropped out from video data. Both basis and encoding matrices obtained from temporal video volumes are rearranged and concatenated to obtain a row-based feature vectors. Final floating point value type fingerprint is obtained by an inner product of row vectors with Gaussian

distributed weighting vectors. Afterwards a simple median value thresholding method is utilized to obtain a binary fingerprint. In their experiments temporal granularity of 1 second (25 frames) and a fingerprint size of 1280 bits are used. Although their results are comparable with methods employing global representation, they cannot compete with methods adopting local representation.

Flip-SIFT has been proposed by Zhao et al. in [33] to handle flip or flip-like transformations observed in real-world applications such as Content-based Copy Detection. As pointed out by the authors, interest point detectors are generally flip invariant but descriptors are not tolerant to flip-like transformations. In [33], it is first decided whether a flip has occurred in the detected salient region and if a flip-like transformation is detected invariance is introduced by flipping the region horizontally or vertically before descriptor extraction. The decision of flipping a region is based on the flow or the dominant *curl* along the tangent direction. Afterwards, regions is flipped (or not) and regular SIFT descriptor is extracted. At the indexing stage, a revised inverted file structure is adopted in which flip decision is also retained for each descriptor in the index structure. This extra information is later used for pruning false positives at the geometric consistency checking step. Experimental analysis is carried out on TRECVID 2010 CCD dataset for content-based copy problem domain and for object detection/recognition problem PASCAL VOC 2009 dataset is utilized. Although F-SIFT extraction is on third slower than SIFT it show promising results in both problem domains.

In [37], Kim et al. proposed a multi-modal approach in order to improve video copy detection which combines spatial and temporal features. Two different types of features are employed; a spatial feature obtained from local DCT coefficients of a keyframe and a temporal feature extracted from the temporal variances of consecutive frames. Before extracting temporal and spatial features Kim first downsample the frame to a fixed size in order to handle varying reference and query frame sizes. Furthermore, to overcome flipping attacks downsampled frames are folded to half width over itself and summed. Afterwards TIRI-DCT [38] is applied to obtain the spatial feature. Temporal feature is obtained from a simple variance operation on overlapping sub-blocks of consecutive frames. And finally features are quantized into 256 bits. Authors also propose an adaptive weighting scheme for modality fusion. Over-

all performance of the method is extensively tested on synthetic and TRECVID 2009 CCD dataset.

2.4 Evaluation Methods and Publicly Available Datasets

When the literature on near duplicate video detection is analyzed it becomes apparent that there is no agreed method of evaluating algorithms. Some researchers use precision/recall and their variants while others use measures from information theory and some other utilize measures based on cost functions. In addition to that there are many different datasets used by researchers, which are in some cases publicly available and in others totally inaccessible. Both of these reasons make it harder to compare different approaches in the literature. Nevertheless, in this section different evaluation techniques and publicly available datasets are analyzed for comparison. For a more in depth review of evaluation methods reader is referred to Bailer [50].

2.4.1 Evaluation Methods

Standard recall and precision metrics are utilized in some of the copy/near-duplicate detection methods. In most of them, query or segment based computation scheme is adopted. In Muscle VCD benchmark [51], for example, a query based recall measure is defined as

$$R_q = N_{correct}/N \tag{2.1}$$

where $N_{correct}$ is the number of correct detection and N is the total number of queries. Since metric does not consider the localization precision another measure is introduced;

$$q_f = 1 - N_{miss}/N_{frames} \tag{2.2}$$

where N_{miss} is the total number of mismatched frames due to either non-detected queries or imprecision of localization and N_{frames} is total frame number of the queries.

By using the preceding measures the cost of miss and false alarm are not considered separately. In some applications even a single miss cannot be tolerated and in some others a false alarm thus, a metric based on a cost function would be more revealing. Secondly, with such a hit or miss approach a method that produces a cluster of scenes is not evaluated fairly.

2.4.1.1 Average Precision

Recall and precision values are computed based on the assumption that only a single result returned for a given query. However, a query can be successfully matched with multiple reference videos simply because it is composed of multiple reference videos or as a result of duplicate videos in the database. Hence most of the time a list of detection results are obtained by the detection systems. Furthermore, while evaluating systems that return a ranked list of results, it is important to consider the order in which the results are presented. In the information retrieval community, Average Precision (AP) is extensively employed in order to favor systems that return more relevant detection results earlier. It is the average of precisions computed at each relevant result position in the ranked list;

$$AP = \frac{1}{N_{max}} \sum_{r} \left(P(r) \times \delta(r) \right) \tag{2.3}$$

where r is the rank, N_{max} is the number of relevant results, $P(r)$ is the precision at rank r and $\delta(r)$ is an indicator function equaling 1 if the item at rank r is a relevant result, zero otherwise. Furthermore, by setting a cut-off rank n $AP@n$ can be computed over a truncated list of results.

2.4.1.2 Normalized Detection Cost Rate

Normalized Detection Cost Rate (NDCR) is mainly utilized in the TRECVID Content-Based Copy Detection (CCD) benchmark task. In the CCD task given a reference set of videos and test queries, for each query the place, if any, that some part of the query video occurs with possible transformations is determined. In this context,

NDCR evaluates algorithms corresponding to different application profiles by introducing cost for false alarms C_{FA} and misses C_{miss}, additionally introducing a target false alarm rate R_{Target} per query duration. Application profiles can range from no false alarm to a balanced profile by modifying these parameters. Another factor $\beta = C_{FA}/\left(C_{miss} \times R_{Target}\right)$ is computed which works as a normalization factor across a range of parameters. Finally, metric is defined as

$$NDCR = P_{miss} + \beta R_{FA} \tag{2.4}$$

where the probability of a miss and false alarm rate is defined as

$$P_{miss} = \frac{FN}{N_{Target}} \tag{2.5}$$

$$R_{FA} = \frac{FP}{(T_{refdata} \times T_{query})} \tag{2.6}$$

As it can be seen from NDCR definition the accuracy of finding the exact copy in the reference video is not considered therefore a separate measure is computed using time precision and recall and these two numbers are combined using the $F1$ measure.

2.4.1.3 Normalized Mutual Information

Normalized Mutual Information (NMI) is proposed for clustering repeated takes into scenes in [10]. The resulting set of scene clusters, D, and the set of ground-truth scenes, D', are interpreted as random variables and mutual information between them is computed as;

$$I(D; D') = \sum_i \sum_j \frac{|D_i \cap D'_j|}{N} \log \frac{|ND_i \cap D'_j|}{|D_i||D'_j|} \tag{2.7}$$

where $|D_i \cap D'_j|$ is the number of segments shared between the scenes and N is total number of segments. NMI is then computed by

$$NMI(D, D') = \frac{2 \times I(D; D')}{H(D) + H(D')} \tag{2.8}$$

in which $H(\cdot)$ being the entropy of the scene clusters

$$H(D) = -\sum_{i} \frac{|D_i|}{N} \log \frac{|D_i|}{N} \tag{2.9}$$

2.4.2 Publicly Available Datasets

Most of the publicly available datasets are for copy detection. One of the first copy detection dataset made publicly available is MUSCLE VCD [51], which contains 100 hours of reference data from different sources; web video clips, TV archives and movies. Reference videos cover a large program types including documentaries, movies, sports events, TV shows, cartoons etc. and videos have different bitrates, different resolutions and different video format. Additionally, 18 query videos generated from reference videos with wide variety of transformations are included. Cropping, strong re-encoding, blurring, camcording, resize and zoom are among the transformations. Furthermore, ground-truth information of queries is provided by this dataset.

In terms of query/reference videos and transformations, richer copy-detection datasets are provided by TRECVID. In total there are 3 different sets, which are from TRECVID 2008, 2009 and 2010. Each of the dataset contain wide variety of transformations and ample amount of reference and query videos; at least 200 hours of reference videos and 1000 queries, which range in time from 3 seconds to 3 minutes. Each query is created by applying one or more transformations to a randomly selected portion of another video, which may or may not be indexed in the reference video database, and some of the queries are padded with unrelated clips, which are not in the reference videos. Transformations are designed to imitate real life attacks and can be in 10 different forms for video, including but not limited to color transformations, spatial transformations, pattern insertion, re-encoding and different combinations of these [52]. TRECVID 2007, 2008 and 2009 dataset are sourced from Netherlands Institute for Sound and Vision and Internet Archive, which contain news magazine, science news, news reports, documentaries, educational programming, and archival video. On the other hand, TRECVID 2010 CCD dataset is from Internet Archive with Creative Commons licenses in MPEG-4/H.264 with durations between 3.6 and

4.1 minutes and contains user-generated videos, documentaries, movies and etc.

Although it is not generally used for near-duplicate another dataset from TRECVID can be thought as, which is from TRECVID Rushes Task. Rushes are the raw material (extra video, B-rolls footage) used to produce a video. 20 to 40 times as much material may be shot as actually becomes part of the finished product. Rushes contain many sequences of frames that are highly repetitive, e.g., many takes of the same scene redone due to errors (e.g. an actor gets his lines wrong, a plane flies over, etc.), long segments in which the camera is fixed on a given scene or barely moving and etc. Thus it is a good candidate for being a near-duplicate dataset.

Recently Revaud et al. introduced a new dataset for event retrieval in [13]. EVVE (Event Video) dataset is only dedicated to the retrieval of particular events. Dataset contains 166 hours of YouTube video belonging to 13 different but particular events such as "Presidential victory speech of Barack Obama 2008". Some of the videos are captured by professional photographers and cameramen while some are only captured by amateur spectators. Furthermore, a set of 100,000 distractor videos unrelated to the events is provided.

Finally, in Table 2.2 summary of publicly available datasets are given.

Table 2.2: Summary of publicly available datasets

Dataset	Type	Ref. Video	Source	Query Size	Transformations
MUSCLE VCD	Copy	100 hours	Variety of sources	18 clips	Cropping, re-encoding, blurring, camcording, resize, zoom, change of gamma.
TRECVID 2008	Copy	200 hours	Netherlands Institute for Sound and Vision	2000 clips, 3 sec. to 3 min.	Camcording, picture-in-picture, pattern insertion, reencoding, change of gamma.
TRECVID 2009	Copy	400 hours	Netherlands Institute for Sound and Vision	1480 clips, 3 sec. to 3 min.	Picture-in-Picture, pattern insertion, re-encoding, change of gamma, crop, shift, caption, flip, frame dropping, additive noise.
TRECVID 2010	Copy	400 hours	Internet Archive	1608 clips, 3 sec. to 3 min.	Camcording, picture-in-Picture, pattern insertion, re-encoding, change of gamma, crop, shift, caption, flip, frame dropping, additive noise.
TRECVID Rushes	Near-Duplicate	53 hours	BBC	N/A	N/A
EVVE	Near-Duplicate	166 hours	YouTube	13 events	N/A

CHAPTER 3

CONTENT-BASED COPY DETECTION

Conceptual block diagram of a CBCD system can be seen in Figure 3.1. There are two major process branches, one is the generation of the reference index and the other is the querying process. In both processes a fingerprint extraction is carried out in which a huge amount of content is mapped to a lower dimensional space for effective representation of the information conveyed. In this respect, fingerprints can be characterized as compact and descriptive. The compactness of the fingerprints facilitates and accelerates the query process and decreases the required indexing storage. The descriptiveness of the index entails the discriminative power, which enables the discrimination between different contents, and robustness under certain transformations. Indexing, on the other hand, enables fast lookup thus accelerating the matching process. Finally by matching the similarity of two videos/fingerprints are inspected.

In the subsequent two sections, two different Content-Based Copy Detection approaches are discussed. In Section 3.1 a global spatio-temporal feature based approach which is published in [19] is reported and in Section 3.2 an interest point based approach is introduced which is a culmination of a variety of different methods from the literature.

3.1 Long-Coarse Visual Features-Based Copy Detection

In summary this approach is mainly a feature matching between query and the reference videos in which features are extracted from spatio-temporal units of the videos. These aforementioned units are formed by a uniform grid structure, which enables

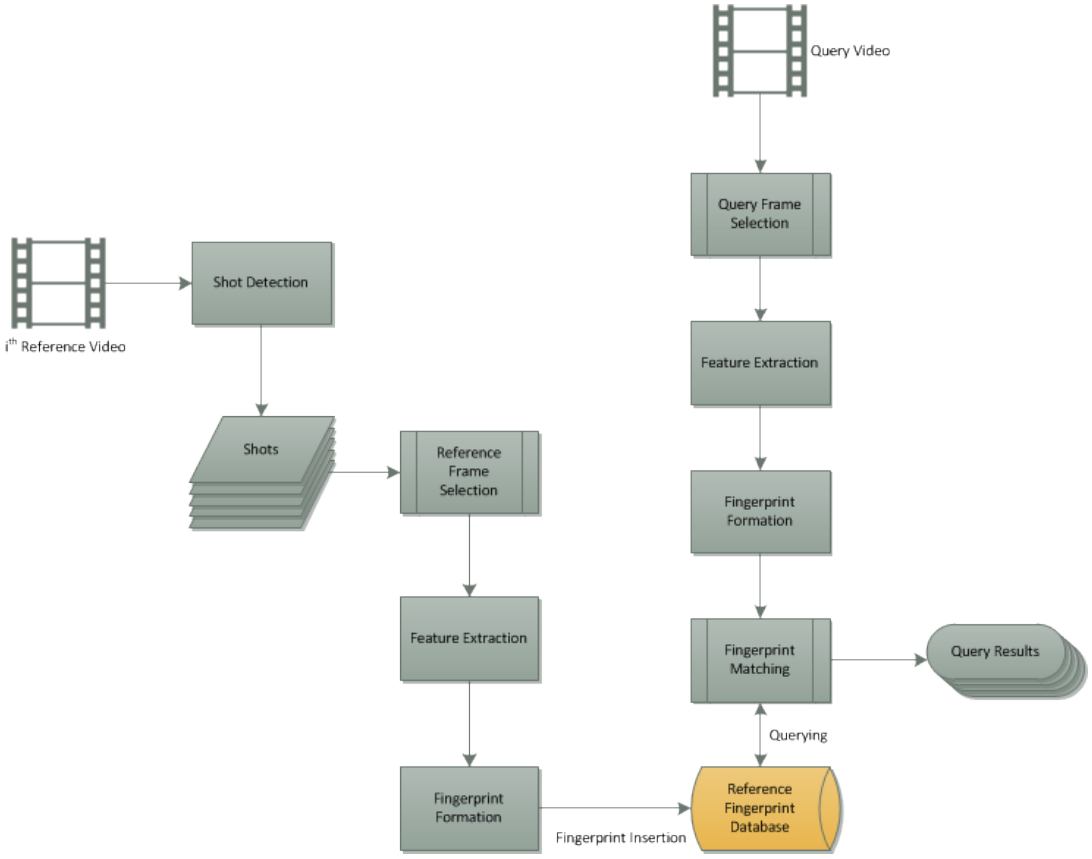


Figure 3.1: Block Diagram of an all-around CBCD System.

spatial and temporal overlapping between separate units. Furthermore, multiresolution property of the whole method is introduced by incorporating downsampled frames into the equation. Additionally, temporality is achieved by extending each spatial grid element in time, yielding a rectangular prism. And for each prism, a feature vector is computed by a set of feature extractors that spans three fundamental visual information sources that are color, texture and motion. In this approach the set of feature extraction methods are derived from some of the MPEG-7 visual descriptors [53] by some modifications. It should be noted that these modifications are introduced in order to decrease the computational complexity of feature extraction and introduce the ability of coarse representation to the descriptors. Moreover, coarse representation is further accentuated by quantization of the feature vectors. Note that, these pseudo-MPEG-7 features are extracted from each grid element and concatenated to form a single long and coarse feature vector for a single prism that extends through time and space on the video. Finally, matching query segments are identified by searching query features on a database that is constructed by the reference video

features.

As the first step of describing a video by the feature vectors, video is segmented into non-overlapping equal time intervals. For each segment, a single long feature vector is computed thus allowing subsets of reference videos to be included and searched in the database. Each segment is also divided by a multi-resolution grid structure in the spatial domain. First level grid structure represents whole frame area whereas second level structure divides frame into 5 uniform regions including a center region overlapping with the corner regions. And the third level grid structure partitions frame into 25 overlapping regions. Pixel values in level i and region j can be represented as $Y_{i,j}(x, y, t)$, $U_{i,j}(x, y, t)$, $V_{i,j}(x, y, t)$, where x, y, t are the spatio-temporal coordinate system variables and Y, U, V are the luminance and color channels of YUV color space. For a given temporal segment, a complete feature vector f_T is obtained by concatenating and quantizing features computed from each aforementioned region.

Although low-level feature extraction methods can be tailored for specific attacks, we have used following features; variants of Dominant Color and Structured Dominant Color for representation of color content, Discrete Cosine Transform and simplistic edge energy for representing texture content and finally motion activity features for representing temporal content. Color features are computed from color histogram (3.1) and structured color histogram (3.2). These histograms are formed from 256 bins and for other color channels, namely U and V, are computed in the same manner.

$$h^Y_{i,j}(c) = \sum_{x,y,t} \delta(Y_{ij}(x, y, t) - c) \tag{3.1}$$

$$sh^Y_{i,j}(c) = \sum_{x,y,t} \delta(Y_{ij}(x, y, t) - c)\alpha^Y(x, y, t) \tag{3.2}$$

In (3.2), binary parameter α^Y takes the value 1 when in the given video volume pixel, neighboring values are in the range that is determined by a threshold otherwise it takes the value 0. Moreover in our work, coarse histograms (3.3) and (3.4) are used which are computed by using pre-determined color levels r_1, r_2, r_3 and r_4.

$$\hat{h}^Y_{i,j}(n) = \sum_{x,y,t} \delta(\hat{Y}_{ij}(x, y, t) - r^Y_n) \tag{3.3}$$

$$\hat{sh}_{i,j}^{Y}(n) = \sum_{x,y,t} \delta(\hat{Y}_{ij}(x,y,t) - r_n^Y)\alpha^Y(x,y,t) \tag{3.4}$$

In (3.3) and (3.4), $\hat{Y}_{i,j}$ represents value of the closest pre-determined color value to the actual value which is described by (3.5). For U and V channels similar computation methods are used.

$$\hat{Y}_{ij}(x,y,t) = \arg\min_{r_m^Y} \|Y_{ij}(x,y,t) - r_m^Y\| \tag{3.5}$$

Edge and motion features are calculated by the help of edge energy $e_{ij}(x,y,t)$ and two-dimensional motion vector components $m_{ij}^X(x,y,t)$ and $m_{ij}^Y(x,y,t)$. Dominant Color Feature (3.6) is calculated as a three dimensional vector on a 3D video volume of YUV color channels. This feature represents the most observed intensity value for each channel in the volume.

$$f_{ij}^{Y,DC} = \arg\max_{c}(h_{ij}(c)) \tag{3.6}$$

Structured Dominant Color Feature (3.7) is defined as the most observed color in the structured color histogram (3.2).

$$f_{ij}^{Y,SDC} = \arg\max_{c}(sh_{ij}(c)) \tag{3.7}$$

Color Frequency Feature (3.8) is computed as the frequency of color values in a given 3D video volume for every color channel. It is determined around the aforementioned pre-determined color values.

$$f_{ij}^{Y,CF} = \begin{bmatrix} \hat{h}_{i,j}^{Y}(1) \\ \hat{h}_{i,j}^{Y}(2) \\ \hat{h}_{i,j}^{Y}(3) \\ \hat{h}_{i,j}^{Y}(4) \end{bmatrix} / \sum_{n=1}^{4} \hat{h}_{i,j}^{Y}(n) \tag{3.8}$$

Structured Color Frequency Feature (3.9) is determined similar to the Color Frequency Feature in which instead of traditional histogram a structured histogram is utilized.

$$f_{ij}^{Y,SCF} = \begin{bmatrix} \hat{sh}_{i,j}^{Y}(1) \\ \hat{sh}_{i,j}^{Y}(2) \\ \hat{sh}_{i,j}^{Y}(3) \\ \hat{sh}_{i,j}^{Y}(4) \end{bmatrix} / \sum_{n=1}^{4} \hat{sh}_{i,j}^{Y}(n) \tag{3.9}$$

Discrete Cosine Transform Feature f_{ij}^{DCT} is computed as the transform coefficients at the lowest four frequencies which are calculated on a 3D luminance video volume.

Edge Energy Feature (3.10) is computed as the average of edge energies calculated on the luminance of the 3D video volume by the help of the 2D spatial Sobel operator.

$$f_{ij}^{EE} = \frac{\sum_{x,y,t} e_{ij}(x,y,t)}{N_x N_y N_t} \tag{3.10}$$

Motion Activity Feature (3.11)is computed as the average of the magnitudes of the motion vectors on a given video prism.

$$f_{ij}^{MA} = \frac{\sum_{x,y,t} \sqrt{m_{ij}^{X}(x,y,t)^2 + m_{ij}^{Y}(x,y,t)^2}}{N_x N_y N_t} \tag{3.11}$$

For a given segment of a video, a feature vector f_T, in other words a fingerprint is obtained by using combinations of features presented on the 3D grid units discussed previously. In this work, two different fingerprints that are depicted in Table 3 are used. The feature values are quantized to four levels by Lloyd's Method [15] resulting with the coarse feature vector d_T. The quantized features are stored in the reference video database as a fingerprint for the given temporal range of the reference video. The overall block diagram of the system is depicted in Figure 3.2.

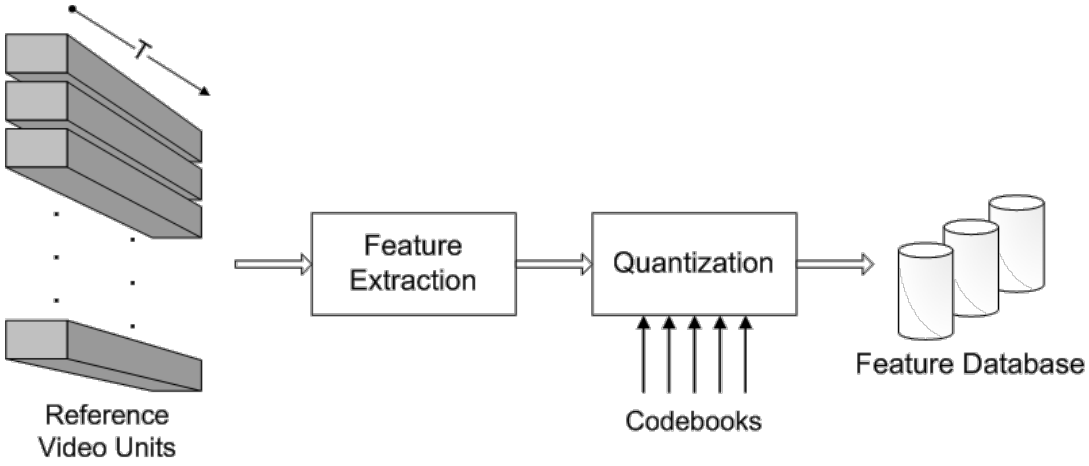


Figure 3.2: Method of constructing visual feature database from reference videos.

For the detection of the query video, firstly query fingerprint, q_T, is computed as described in the previous section. Afterwards, similarities between q_T and every fingerprint, d_T, in the reference database are computed. This comparison is achieved by sliding q_T over entire length of individual reference fingerprints. At this point,

Table 3.1: Fingerprint dimensions and ingredients in a unit interval.

	Fingerprint 1	Fingerprint 2
Dominant Color	3 Levels (93 values)	-
SDC	2 Levels (18 values)	-
Color Frequency	-	3 Levels (372 values)
SCF	-	2 Levels (72 values)
DCT	2 Levels (24 values)	2 Levels (24 values)
Edge Energy	2 Levels (6 values)	2 Levels (6 values)
Motion Activity	2 Levels (6 values)	2 Levels (6 values)
Total	147 Dimensions	480 Dimensions

unique video locations that are exceeding a predetermined similarity value are combined, sorted and presented as the search result. Although similarity measures can be tailored for specific requirements, in this study Euclidean Distance (3.12) and Cosine of the Angle (13) between two fingerprints are used and compared as a measure.

$$s^{EUC} = \frac{1}{1 + \sqrt{\sum_i (d_{t_i} - q_{t_i})^2}} \tag{3.12}$$

$$s^{cos} = \frac{d_t \cdot q_t}{\|d_t\| \|q_t\|} \tag{3.13}$$

3.1.1 Experiments and Discussion

TRECVID 2008 CCD dataset (Section 2.4.2) is utilized for the experimental evaluation of the proposed spatio-temporal method. Non-overlapping temporal window length is selected as 25 frames and a total of $720,000$ global features are extracted from the reference dataset. Similarity measures in (3.13) and (3.12) have been compared to each other. Furthermore, two different coarse fingerprints discussed in the previous section are also tested. Detection performance is reported by the NDCR measure in Figure 3.5 for corresponding attack types. Although proposed global spatio-temporal representation performs better then the median of TRECVID 2008 CCD participants, when compared with the best results obtained on the same dataset the results are not satisfactory especially for attack types like *camcording*, *picture-in-picture* and *post-production*. For such attacks local features should be utilized in order to improve the performance.

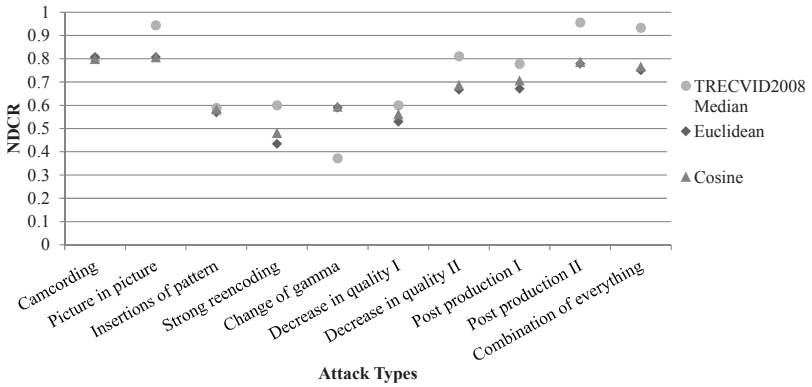


Figure 3.3: NDCR performance of the proposed spatio-temporal global representation (Fingerpint-1) utilizing with different similarity functions.

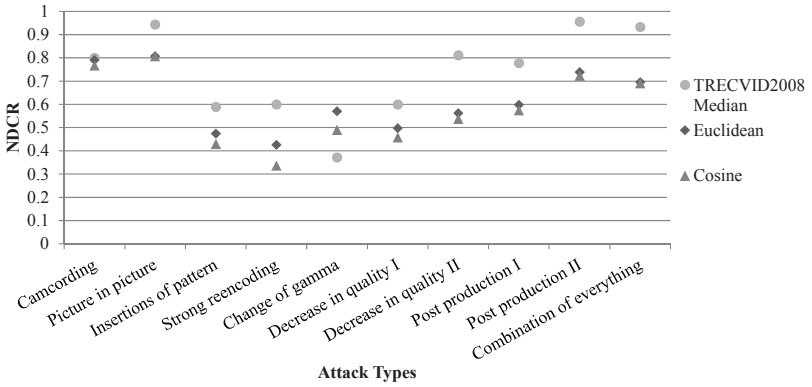


Figure 3.4: NDCR performance of the proposed spatio-temporal global representation (Fingerpint-2) utilizing with different similarity functions.

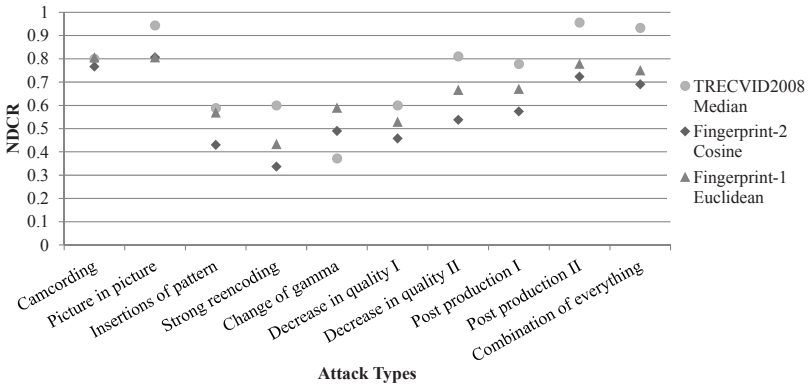


Figure 3.5: Comparison of Fingerprint-1 and Fingerprint-2 in terms of NDCR performance on TRECVID 2008 CCD Dataset.

3.2 Interest-Point Based Copy Detection

For an interest-point based CBCD system, there are many feature extraction choices including but not limited to Harris corner detector coupled with image gradients, Scale Invariant Feature Transform (SIFT) and Speeded-up Robust Features (SURF). Each one of them has been reported as successful as the next one in the CBCD literature. In this study SIFT has been selected as the feature extraction and descriptor method, which has been implemented from scratch for this work. After extracting features one may choose to utilize feature vectors directly, in other words in raw format as the representation of the video segment. On the other hand, as the size of the reference video database increases, using raw feature vectors may increase the processing power, storage and direct access memory requirements of the system. Thus a transformation should be considered in order to meet limitations of any setup. It can be a vector quantization, Bag-of-Visual-Words (BoVW) and etc. but obtaining an optimum representation is crucial. Subsequent to obtaining a representation namely a visual fingerprint of the video segment a fast, efficient and light lookup method in other words an indexing is necessary. Finally, fingerprints are matched by using implicit distance metrics and further improvement is achieved by employing post-processing methods forcing temporal and geometric consistencies on the results. Details of the representation, indexing and matching methods from the literature are discussed in Chapter 2.

As previously mentioned for feature extraction SIFT has been selected in this study. However, in order to handle such a large amount of reference data, a shot-based reference frame selection method is employed instead of dense sampling of frames. Shot boundary frames and a given amount of interior frames are only used for feature extraction. Even using only shot boundaries for reference feature extraction a total of 67953659 vectors are obtained. This amount of vectors requires a total of 32.40 GB of storage and memory space if they are used as it is. So another approach is necessary; as pointed in our previous literature survey BoVW are among the mostly used method in order to decrease the dimensionality and amount of the reference data.

3.2.1 Fundamentals of Bag-of-Visual Words

BoVW is basically a sparse vector representation of occurrence counts of features where features are transformed into the closest visual word over a preferably a large vocabulary. This said BoVW can be interpreted as a voting of individual feature vectors with an approximate nearest neighbor search. From this perspective, BoVW can be unrolled to a voting scheme. Given a query frame represented by its local features y_k and a set of reference frames j represented by its local features $x_{i,j}$ voting score corresponding to the frame j is computed as;

$$s_j = \alpha_j \left(\sum_{k=1...m} \sum_{i=1...m_j} f(x_{i,j}, y_k) \right) \tag{3.14}$$

where matching function f measures the similarity between features. α_j on the other hand is the normalization term which, for example, can be $1/m_j$. For a voting system based on ϵ-search or k-nearest neighbor search matching function can be defined as;

$$f_\epsilon(x, y) = \begin{cases} 1 & \text{if } d(x, y) < \epsilon \\ 0 & \text{otherwise} \end{cases} \tag{3.15}$$

$$f_{kNN}(x, y) = \begin{cases} 1 & \text{if } x \text{ is kNN of } y \\ 0 & \text{otherwise} \end{cases} \tag{3.16}$$

where d is a distance function defined on the feature space. For example on SIFT descriptor space generally Euclidean distance is employed.

On the other hand in BoVW approach, query features are quantized by a quantizer q which is formally defined as;

$$\begin{aligned} q : \mathbb{R}^d &\to [1, k] \\ x &\mapsto q(x) \end{aligned} \tag{3.17}$$

In general quantizer q is obtained by performing k-means clustering on a training set. The quantizer $q(x)$ is the index of the closest centroid to the vector x. And two

vectors x and y are matched if both vectors are quantized to the same centroid.

$$f_q(x, y) = \delta_{q(x)q(y)} \tag{3.18}$$

$$s_j = \alpha_j \left(\sum_{k=1...m} \sum_{i=1...m_j} \delta_{q(x_{i,j})q(y_k)} \right) \tag{3.19}$$

Note that this score also corresponds to the inner product of two BoVW vectors obtained from the query frame and the database frame. Normalization term in that case can be either the L_2 or L_1 norm. For large vocabularies L_2 norm of BoVW vector is very close to the square root of the L_1 norm which can be seen from the voting approach as a compromise between measuring the number and the rate of feature matches.

BoVW based matching combines the advantages of local representation and efficient matching using inverted files however; the quantization reduces the discriminative power of the local features. There is always a tradeoff between the quantization noise and the descriptor noise while choosing the number of quantization levels i.e. the number of centroids k. A low value of k leads to large Voronoi cells thus increasing the probability of a noisy version of a descriptor belonging to the correct cell. But this also increases the number of false assignments. On the other hand a large number of centroids increase the precision however this decreases the probability of a noisy version matching to the same cell.

3.2.2 Hamming Embedding and Product Quantization

One approach to overcome the aforementioned dilemma is to employ refining on the quantization step. That is; after a coarse quantization of the feature vector a detail term refining the quantized index is computed and also stored/indexed. In [24] a binary signature, encoding the location of the vector in the Voronoi cell is extracted. Binary signature is designed so that the Hamming distance between two vectors in

the same cell approximates the Euclidean distance.

$$h(b(x), b(y)) = \sum_{1 \leq i \leq d_b} \left(1 - \delta_{b_i(x) b_j(y)}\right) \tag{3.20}$$

In the aforementioned work a random orthogonal projection matrix and a simple principal component analysis are proposed for computing the binary signature. Additionally in [46], coarse quantization error is encoded as a binary signature which is efficiently obtained by employing a simple product quantization on the difference vector. Product quantization enables to choose the components to be quantized jointly. This also supports better quantization of feature vectors that are structured such as SIFT descriptors, which are built as concatenated orientation histograms with fixed size. Formally it is defined as follows;

$$\begin{gathered} r(y) = x - q_c(x) \\ \underbrace{r_1, \ldots, r_{D^*}}_{u_1(r)}, \ldots, \underbrace{r_{D-D^*+1}, \ldots, r_D}_{u_m(r)} \\ q_p(r(y)) = q_1(u_1(r)) \ldots q_m(u_m(r)) \end{gathered} \tag{3.21}$$

where m is the number of sub-quantizers and $D^* = D/m$ is the dimension of sub-vectors u_i moreover k^* is the number of sub-quantization levels. Finally for such a setting total bit length of the binary signature is $l = m \log(k^*)$. In general when compared with Llyod's algorithm for the same bit length we would need $k = (k^*)^m$ centroids instead of mk^* or just k^*. From the published results best selection of m and k^* are selected as $(8, 256)$ respectively which translates to 64 bit length for the binary signature.

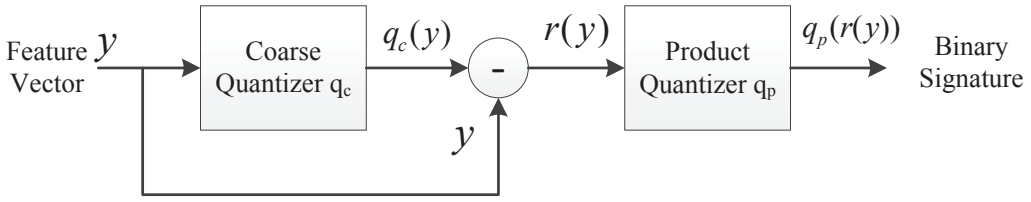


Figure 3.6: Product Quantization based binary signature extraction.

At this point each feature vector is represented by the coarse quantization index and the binary signature. With this methodology overall index/database contains only the frame identifier and the binary signature for each feature vector/descriptor in the implemented inverted list. A simple inverted file indexing method incorporating product quantization is given in Algorithm 1.

Table 3.2: Index structure and index size.

Field	Length (bits)
Frame Identifier	32
Binary Signature	$\lceil m \log k^* \rceil$

input : Given a collection of videos V, a quantization function q
output: L, an inverted feature list of features

Initialize an empty heap structure, $L \leftarrow \emptyset$
for $\forall v \in V$ **do**
    Select frames, $F_v \leftarrow$ `FrameSampler(`v`)`
    for $\forall d \in F_v$ **do**
        Compute local features, F_d
        for $\forall x \in F_d$ **do**
            Compute $w \leftarrow q(x)$
            Compute $r \leftarrow$ `ProductQuantization(`x, w`)`
            Append (d, r) to the list corresponding to w
        end
    end
end
return L

Algorithm 1: Populating inverted feature index.

3.2.3 Feature Matching

At the matching step, much like the previously discussed voting approach, a binary decision can be made by employing a threshold, τ, on the Hamming distance between two binary signatures from the same codeword. However, actually Hamming distances reflect the similarity/closeness of vectors and should be also incorporated to the decision. Since smaller distances correspond to higher confidence on the similarity of the vectors, following Gaussian function is a good choice as a weighting

function;

$$w(h_d) = \exp\left(\frac{-h_d^2}{\sigma^2}\right) \tag{3.22}$$

Figure 3.7 shows different weighting functions obtained by varying σ^2 values. After empirical studies $\sigma^2 = 256$ is selected for further experiments. At this point in order to increase computational efficiency a binary decision can still be made by selecting a threshold τ on Hamming distance. As it can be seen from the figure distances above 28 has very little significance so threshold is selected as $\tau = 28$. A matching score other than 0 is obtained only for distances smaller than τ.

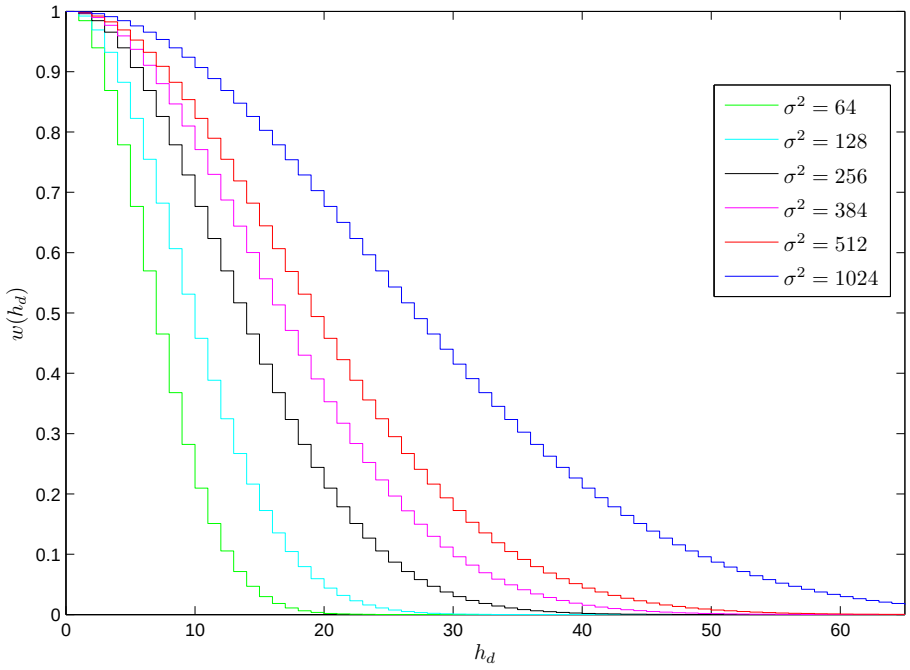


Figure 3.7: Weighting functions for different σ^2 values.

This score still does not take into account the frequency of the visual words over the entire database; which can be easily incorporated by using *tf-idf* weighting; rare visual words are assumed to be more discriminative and should be assigned higher weights. Also as pointed out in [54] squaring the inverse document frequency factor is consistent with the computation of the L_2 distance between BoVW vectors. Inverse document frequency $idf(w, D)$ is computed for each visual word by dividing the total

number of frames, D, by the number of frames containing the word w. In Equation (3.23), 1 is added to denominator not to lead a division-by-zero error.

$$idf(w, D) = log \frac{|D|}{1 + |\{f \in D : w \in f\}|} \tag{3.23}$$

Finally, frame score function formally can be defined as;

$$f_s(x, y) = \begin{cases} idf^2\left(q(x)\right) \times w\left(h_d(q_p(r(x)), q_p(r(y)))\right) & \text{if } q(x) = q(y) \text{ and } h_d < \tau \\ 0 & \text{otherwise} \end{cases} \tag{3.24}$$

Consequently, the total score of a query frame on the reference frame j is computed as follows. In Algorithm 2 matching method is described as a pseudocode in detail.

$$s_j = \alpha_j \left(\sum_{k=1...m} \sum_{i=1...m_j} f_s(x_{i,j}, y_k) \right) \tag{3.25}$$

input : Given an inverted feature index L, a query frame features X, the quantization function q
output: M, a list of matching frames ($|M| = k$)

Allocate accumulator A_d for each frame $d \in L$
Initialize $A_d \leftarrow 0$
for $\forall x \in X$ **do**
    Compute $w \leftarrow q(x)$
    Compute $r \leftarrow$ `ProductQuantization(`x, w`)`
    Fetch inverted list for w, $L_w \leftarrow L(w)$
    for $\forall (d, y) \in L_w$ **do**
        Compute Hamming Score; $s \leftarrow f_s(x, y)$
        Set $A_d \leftarrow A_d + s$
    end
end
for $\forall A_d > 0$ **do**
    Set $S_d \leftarrow A_d / \alpha$
end
Identify the k greatest values S_d values, M
return M

Algorithm 2: Algorithm for frame matching as discussed.

3.2.4 Descriptor and Frame Burstiness Handling

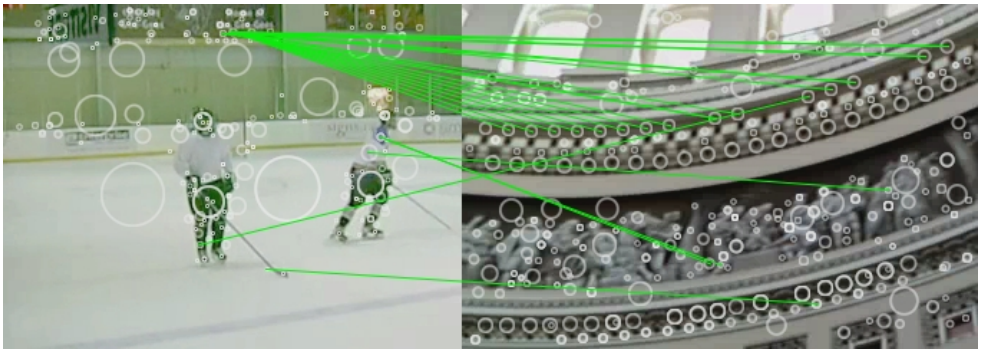
So far, by employing squared *idf* weighting, unbalanced visual word distributions is taken account for. However, *idf* does not improve the effects of burstiness i.e.: if a visual word is observed in an image, it is likely to observe it in multiples. In Figure 3.8, example matching frames showing burstiness is given. As it can be seen from figures, even utilizing a detail fingerprint does not handle it. This phenomenon is first observed in document analysis and different methods to overcome are proposed. In [55], Katz et al. models the in-document distribution of words or phrases using K-mixtures and in [56] a Poisson distribution is utilized to measure the burstiness of a term. Furthermore, Madsen [57] and He [58] have shown that accounting for burstiness improves text classification and topic clustering. On the other hand, a simple way of handling visual term burstiness is proposed in [59] by discarding ambiguous features that occur more than 6 times in a given image. Also, in [60] different normalization techniques are proposed in order to overcome detrimental effects of bursty visual terms without discarding any features. Furthermore, not only the multiple occurrence of a visual element in the same image, which corresponds to *intra-image* burstiness but also the *inter-image* burstiness i.e.: the occurrence rate of visual elements in different images is dealt with a simple penalization method in [60], which have been adopted in this study.

In intra-image normalization, score $f_s(x, y_j)$ of a query descriptor x to a reference descriptor, y_j, of frame j is updated by using (3.26) where $T_j(x)$ is the total score of query descriptor over the reference frame j computed as in (3.27).

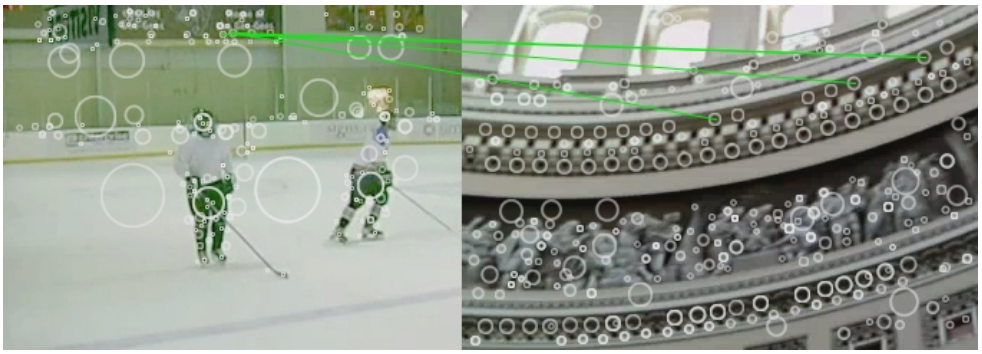
$$f'_s(x, y_j) = f_s(x, y_j)\sqrt{\frac{f_s(x, y_j)}{T_j(x)}} \tag{3.26}$$

$$T_j(x) = \sum_{y_j \in j} f_s(x, y_j) \tag{3.27}$$

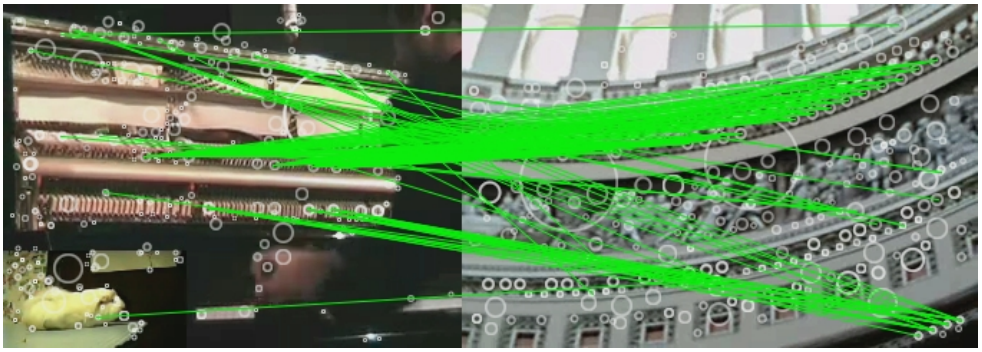
By applying intra-image normalization, if a single descriptor is matched to a single descriptor on a reference frame, score is unchanged. However, if the descriptor matches to multiple features on a reference, frame score is penalized. Inter-image nor-



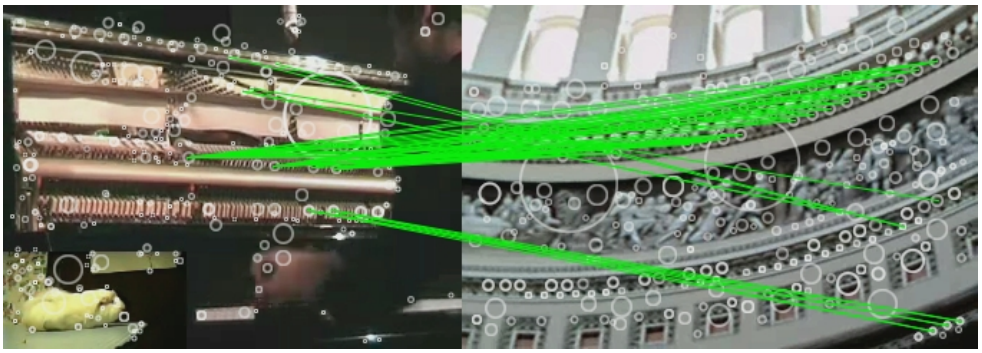
(a)



(b)



(c)



(d)

Figure 3.8: Examples showing burstiness effect with (a),(c) and without (b),(d) detail fingerprints.

malization, on the other hand, addresses bursts across reference images since some visual elements can be frequent across other images. Actually this is tried to be addressed by employing *idf* weighting scheme. However, this approach does only account for the number of features of a given visual word in a reference feature database. Score of a match is updated by (3.28) in which normalization factor $T(x)$ is the sum of matching scores of descriptor x over all reference frames.

$$f'_s(x, y_j) = f_s(x, y_j)\sqrt{\frac{f_s(x, y_j)}{T(x)}} \tag{3.28}$$

$$T(x) = \sum_j \sum_{y_j \in j} f_s(x, y_j) \tag{3.29}$$

After matching and computing scores for query features and reference features the output of the algorithm is a set of tuples (t_q, t_r, r, s_f); where t_q and t_r are the timestamps of the matched query and reference frames, r is the reference video identifier and s_f is the normalized frame match score. Much like the burstiness of a single descriptor, query frames can match well with several reference frames from the reference database increasing the false positives. In order to handle this effect, a frame score normalization is employed. First, the sum, T_f, of all matching reference frame scores, s_f, is computed and the final frame score is obtained by (3.30). The choice of squaring is not arbitrary and empirical analysis can be found for descriptor burstiness in [60].

$$s_f^* = s_f \left(\frac{s_f}{T_f}\right)^2 \tag{3.30}$$

3.2.5 Temporal Alignment

Matching frame set of tuples, (t_q, t_r, r, s_f^*), can be transformed into a reference video sequence in terms of temporal sequence alignment by employing dynamic programming techniques such as dynamic time warping. However, sparsity of the frame scores renders DTW infeasible for content-based copy detection problem. Instead

a simplified approach can be used e.g.: partial alignment [61], Hough voting/transform or RANSAC. Because of its simplicity and prowess to capture small temporal shifts Hough voting is preferred in this work. The temporal Hough transform/voting is performed by computing a soft-assigned 1-D histogram $h_r(\delta t)$ of time shifts $\delta t = t_r - t_q$ for corresponding query and reference frame for reference video, r.

$$h_r(\delta t) = \sum_{(t_r, t_q) \in \mathcal{R}} \delta(t_r - t_q) \tag{3.31}$$

Contrary to bin-voting, in [25] frame scores are also accumulated in Equation (3.31). However, empirical studies showed that preferring only bin-voting decreased false positive matches considerably compared to utilizing frame matching scores. A short list of $(b, \hat{\delta}t)$ hypothesis is obtained from the bins with the highest voting scores. Furthermore, non-maxima suppression around the selected peaks are employed in order to decrease the effects of soft-weighting. Finally, from the matching frame timestamps corresponding to $\hat{\delta}t$, query and reference video alignment can be achieved easily. At this point score, s_b, for the detected sequence is computed as the sum of the frame scores corresponding to the peak for video b. However, once again we may face multiple matching results to a single query sequence so a normalization factor, S_{max}, is applied which is the highes score obtained among all matching sequences.

$$s_b^* = s_b \left(\frac{s_b}{S_{max}} \right)^2 \tag{3.32}$$

3.2.6 Experiments

For the experimental setup TRECVID 2009 and 2010 CCD dataset are utilized as previously discussed in Section 2.4. In the TRECVID 2009 dataset, there are 837 videos with total duration of 400+ hours and from these videos $1,759,980$ frames are selected by sampling uniformly and a total of $276,315,688$ reference feature vectors/descriptors are extracted by in-house implementation of SIFT from reference dataset. Also in the TRECVID 2010 dataset there are 400 hours of data from 11,728 media and $1,921,349$ frames are selected by the same sampling rate i.e.: 2 frames per 3 seconds. And a total of $243,000,059$ reference feature vectors/descriptors are

extracted by in-house implementation of SIFT. On the query side, much denser frame sampling , 5 frames per second, is employed. Such an asymmetric sampling is preferred in order to account for the relatively short length of query sequences and the detrimental effects of transformations on feature extraction so that a higher number of features are extracted from queries. For parameter optimizations a shortlist of 35 queries from the TRECVID 2010 CCD dataset is used as a validation set. A visual codebook containing 100K codewords are computed for efficient coarse quantization by utilizing kMeans++ algorithm [62] on an MPI framework from a different dataset. Furthermore for the binary signature computation 256 centroids are used with subvector dimension of 16. All codeword transformations are carried out by utilizing FLANN [63]. Since SIFT descriptors are not invariant against flip-like transformations, queries and their flipped versions are searched in the reference database concurrently. All things considered index size with additional structures corresponds to a memory of 3.7GB and 3.2 GB for TRECVID 2009, TRECVID 2010 CCD datasets respectively.

Table 3.3: Summary of experimental datasets.

	2009	**2010**
Number of Reference Video	837	11,728
Number of Indexed Frames	1,759,980	1,921,349
Number of Indexed Features	276,315,688	243,000,059
Total Disk Space of Raw Features	52 GB	45 GB
Total Index Memory Size	3.7 GB	3.2 GB
Number of Queries	1,407	1,608
Number of Query Features	93,800,166	98,480,474

In Table 3.4, firstly, precision-recall performance of the baseline method is shown for TRECVID 2009 CCD dataset and in Table 3.5 results are tabulated also for TRECVID 2010 CCD dataset. In Figure 3.9 performance of the baseline method with respect to detection (NDCR) and time localization (Mean F1) in BALANCED profile is shown for TRECVID 2009 CCD dataset. Furthermore, experimental results for TRECVID 2010 CCD dataset are depicted in Figure 3.10. In both figures results from recent literature on the same datasets are also plotted for comparison. It should be noted that for experiments on both datasets same set of parameters optimized in the validation

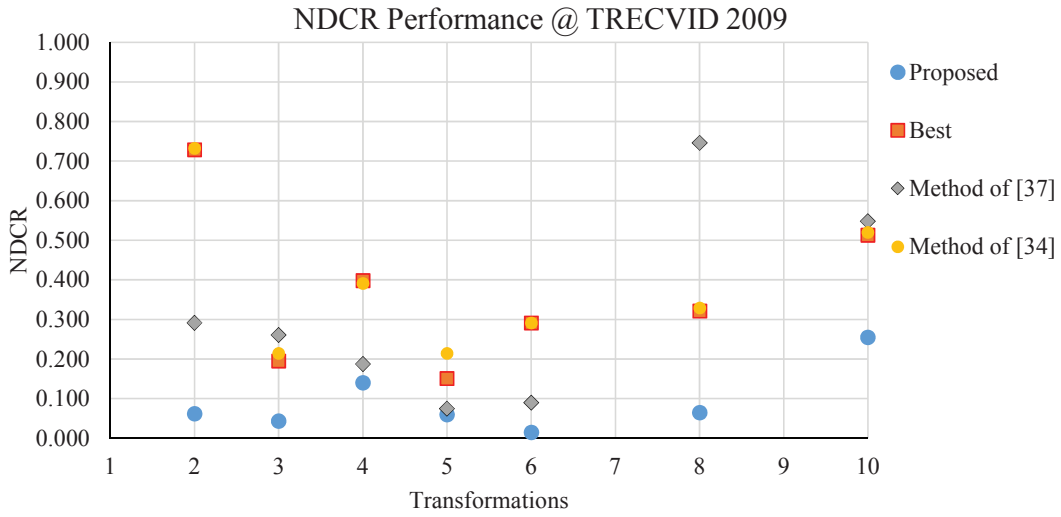
set are used.

Table 3.4: Performance of the baseline method in terms of recall and precision at TRECVID-2009 CCD dataset.

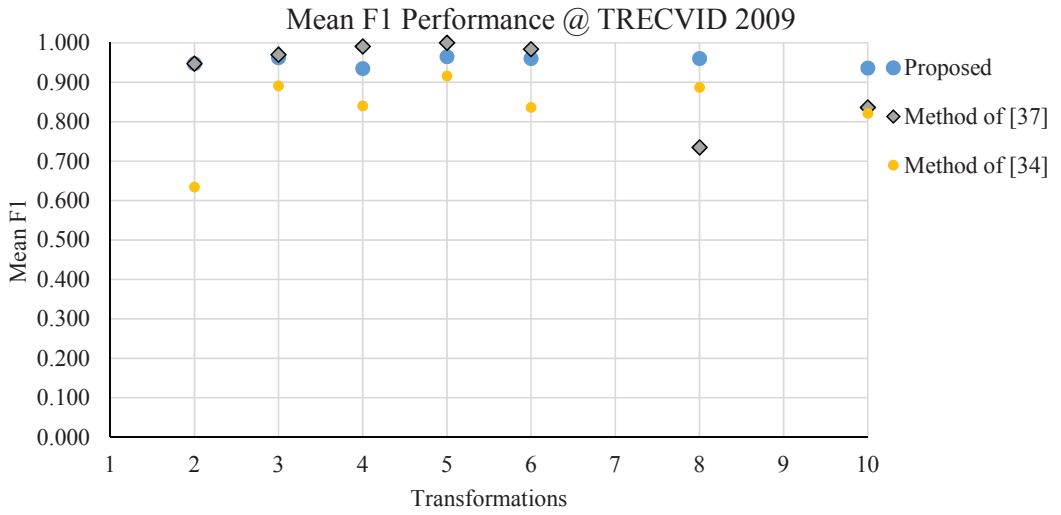
TRECVID2009	**Proposed**		**Method of [64]**	
	Recall	**Precision**	**Recall**	**Precision**
T1	NA	NA	NA	NA
T2	0.985	0.964	0.731	1.000
T3	0.993	0.985	0.955	1.000
T4	0.933	0.926	0.843	0.991
T5	0.993	0.978	0.955	0.992
T6	0.978	1.000	0.985	0.992
T7	NA	NA	NA	NA
T8	0.978	1.000	0.806	1.000
T9	NA	NA	NA	NA
T10	0.828	0.941	0.687	1.000
Average	0.955	0.970	0.852	0.997

Table 3.5: Performance of the baseline method in terms of recall and precision at TRECVID-2010 CCD dataset.

TRECVID2010	**Proposed**		**Method of [33]**	
	Recall	**Precision**	**Recall**	**Precision**
T1	0.716	0.800	0.403	-
T2	0.865	0.943	0.679	-
T3	0.948	0.977	0.940	-
T4	0.672	0.947	0.582	-
T5	0.925	0.954	0.948	-
T6	0.709	0.941	0.634	-
T7	NA	NA	NA	NA
T8	0.948	0.962	0.873	-
T9	NA	NA	NA	NA
T10	0.672	0.968	0.604	-
Average	0.807	0.936	0.708	0.719

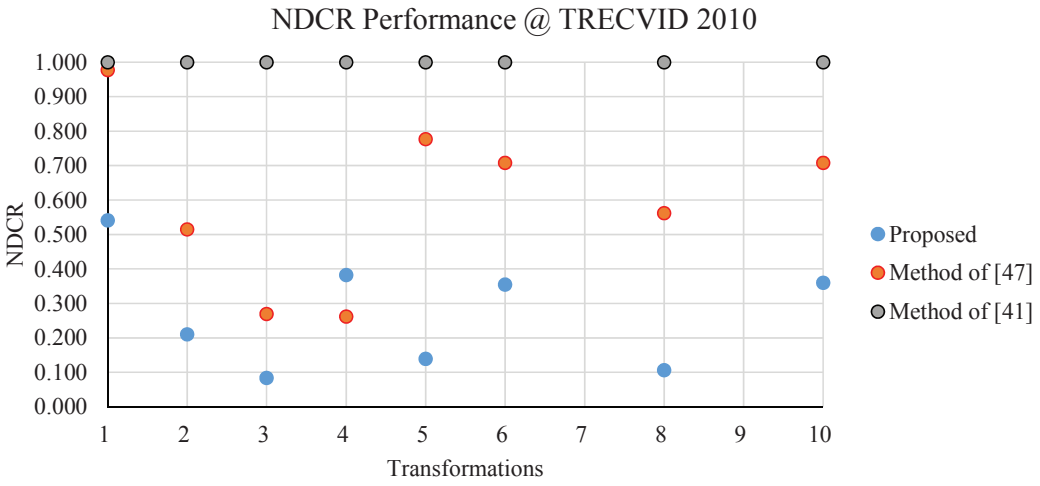


(a)

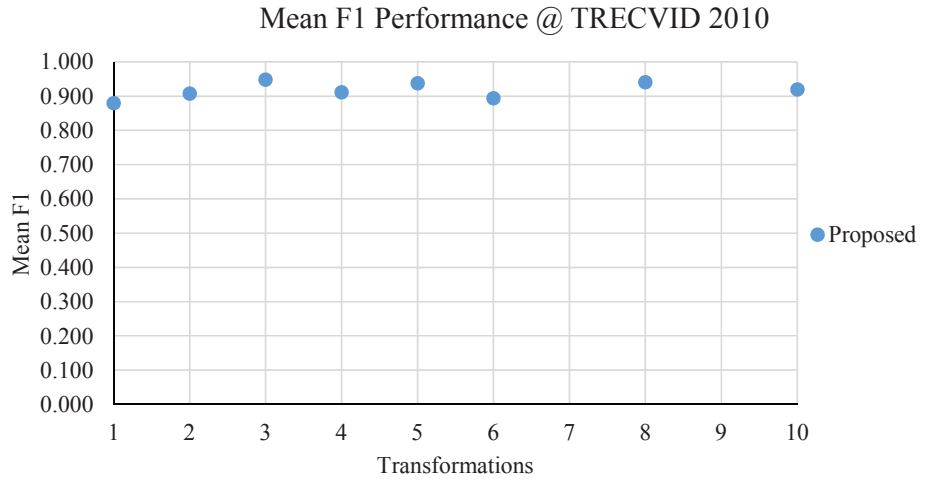


(b)

Figure 3.9: Detection (a) and time localization performance (b) of the baseline method for different attack types at TRECVID 2009 CCD dataset.



(a)



(b)

Figure 3.10: Detection (a) and time localization performance (b) of the baseline method for different attack types at TRECVID 2010 CCD dataset.

3.2.7 Discussions

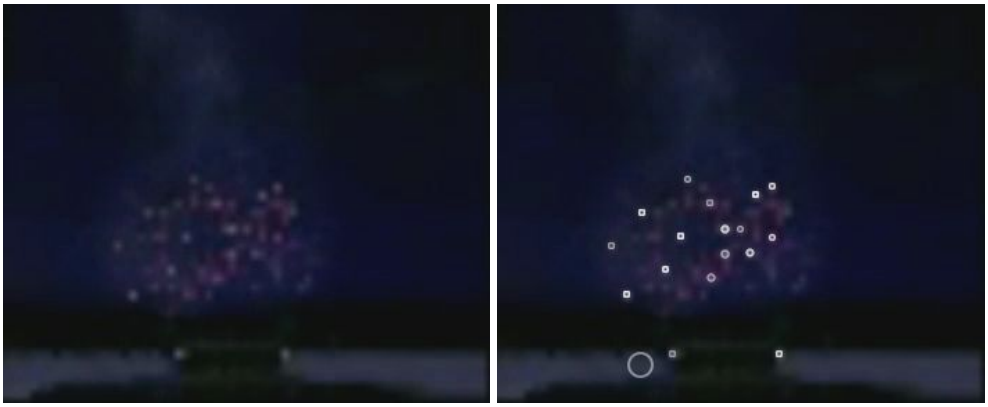
It can be seen from the results in Section 3.2.6 baseline method proposed in Section 3.2 has superior detection and time localization performance in both TRECVID 2009 and 2010 CCD datasets when compared with related work from the recent literature and alike. In Table 3.5, results from [33] are also depicted. Although Zhao et al. introduced Flip Invariant SIFT, their results are lower even for the transformations

containing flip-like attacks (T8 and T10). Many details left missing in their work such as the size of the codebook, presence and length of a detail fingerprint, which affects a fair comparison.

When results from two datasets are compared, one can easily observe that there is a performance loss on TRECVID 2010 CCD dataset. Main reason for this loss is because of the disparity between the quality of productions and distribution of the content in TRECVID 2010 dataset. While TRECVID 2009 CCD dataset contains mostly professionally produced news magazine, science news, news reports, documentaries, educational programming, and archival video from Netherlands Institute for Sound and Vision and Internet Archive, 2010 dataset consists of unstructured user-generated internet videos, that tend to have shorter shots, faster scene changes, computer animations and unexpected effects. Furthermore, in 2010 dataset there are videos consisting of a single shot from a single fixed camera and even a video showing a single page of a document for the full extend of the video. Some extremities can be seen in Figure 3.11 which also depicts corresponding queries from the dataset. Figure 3.11a-d exemplifies a sample case in which feature extraction produces very few features. On the other hand sample frames for a fixed camera is shown in 3.11e-h. One way to overcome these problems might be employing a dense grid feature extraction coupled with denser frame sampling. In this study for each frame an average of 158 and 126 features are extracted and indexed for TRECVID 2009 and 2010 dataset respectively. As it can be seen for TRECVID 2010 considerably less features are indexed and this causes a performance drop compared with TRECVID 2009 dataset.

Furthermore, while comparing baseline method with other approaches from the literature it is observed that most of the related local feature indexing approaches extract higher number of features per frame e.g. 1 keyframe per 1.6 seconds with 300 feature per frame in [33] and 2 frames per second with an average of 420 features per frame in [23]. Thus increasing the average number of features per frame can further improve baseline performance on TRECVID 2010 dataset and even on TRECVID 2009. However, it should be noted that as the number of features in the reference database increases necessity for false positive descriptor and frame pruning methods. At this point one can utilize a longer detail fingerprint and even a geometric consistency check to overcome the burden of growing index size.

Although in the literature there are many approaches containing a post-processing stage for geometric consistency check we have not applied any such method because of the empirical findings we have observed in the validation dataset with our setup. When a variation of weak geometric consistency as discussed in [23] is adopted, a small improvement in detection rate for some of the attacks is observed however, overall computational complexity increases drastically because of the need for relative locations of features and extra memory accesses. Moreover on contrary to most of the related work, no specific method for Picture-in-Picture (PiP) is employed. There are methods that also index downsampled versions of the reference database to handle PiP attacks. But as it is evident from Figure 3.9 and 3.10 and Table 3.4 and 3.5 there is no need for that too. On the other hand, for queries involving camcording attack (Transformation-1) encountered in TRECVID 2010 CCD dataset a higher number of features from queries are most probably necessary. On top of that, since SIFT descriptors' matching power and reproducibility decrease as the viewpoint changes, a higher threshold, τ, in (3.24) and a wider tail (in other words higher σ^2) in (3.22) for fingerprint matching should be considered. Both of these changes, of course, increases the rate false positive descriptor matches.



(a) (b)

(c) (d)

(e) (f)

(g) (h)

Figure 3.11: Frame samples from problematic query and reference video sequences. Sequence (a)-(d) are from

CHAPTER 4

INFORMATION THEORETIC FEATURE INDEXING

Considering the previously discussed approaches in Chapter 3, given a fixed reference database size, the amount of index size depends on the number of features per video. So as the number of reference videos or the temporal and spatial sampling rate increases the index structure grows considerably. Causing the number of collisions surge for a fixed hash length and as a result both detection accuracy and computational complexity suffers. One approach to remedy the database growth is to select a subset of features to be indexed. In other words instead of ingesting all features blindly, selecting most important and informative features could be strategically effective to overcome the data size.

4.1 Fundamentals

Selecting a subset of features is an important pre-processing step for many applications and pattern recognition problems. Formally speaking for a given set of features Y, let $\mathcal{X}_d$ be the set of all possible subsets of size d and $J(X)$ a criterion function that evaluates feature subsets $X \in \mathcal{X}_d$ then feature selection problem can be defined as in Equation 4.1 assuming higher value of J indicates a *better* feature subset. At this point it should be noted that feature selection can mean selecting individual dimension(s) of feature vector-space or features corresponding to a class.

$$\tilde{X}_d = \arg\max_{X \in \mathcal{X}_d} J(X) \tag{4.1}$$

Feature selection methods according to their evaluation criteria can be categorized as *filter*, *wrapper*, *embedded* or *hybrid*. Filter methods are based on statistical, information theoretic or distance measures calculated from subset of features. Commonly used measures are χ^2 test, Euclidean distance and information gain. On the other hand, wrappers utilize performance score of a predefined classifier, examples include genetic algorithms, simulated annealing and etc. Lastly in embedded methods search for an optimal feature subset is built into the classifier itself. A very well known example for embedded methods is the decision trees.

Furthermore, according to subset search strategies, methods can be classified as *exhaustive*, *sequential* or *random*. In exhaustive search all possible subsets are traversed while evaluating the criterion function. Search is complete but computationally intensive and most of the time intractable. Conversely in sequential approaches completeness is traded for simplicity. However, many variations have shown acceptable performance in the literature such as sequential forward feature selection and bi-directional selection. A more complete review of the feature selection literature can be found in [65] and [66].

4.2 Related Work

As an example problem domain, most of the methods proposed for the text document classification problem adopts Best Individual Feature (BIF) selection approach. First an evaluation criterion to be applied to each single word is selected. Afterwards, all words independently evaluated and sorted according to the assigned criterion. Finally a predefined number of words are selected as the best representing feature subset of a document. Another example would be visual object recognition problem where Principal Component Analysis can be employed for whitening and dimension reduction before training and classification on visual features such as SIFT or SURF descriptors.

It is evident that feature selection, whether for dimension reduction or relevant training subset determination, is exhaustively utilized in classification problems. On the other hand feature selection approach has very few precedents in the BoVW-based

retrieval problem domain and none in content-based copy detection. Most of the literature for visual feature selection is implemented as an application-dependent task and the earliest approaches have been applied to geo-localization problem. Given a dense street-view geo-tagged database, Schindler et al. [67] select informative features, i.e. features occurring mostly in images of some specific location. Similarly, in [68] an information content probability is obtained for each feature with respect to their specific location. In [69] *useful features* are selected by an unsupervised geometric verification method. At the useful feature extraction images are queried and only the futures identified as inliers during a RANSAC-based spatial verification step are selected. This is only applicable if object or location already exists in the image database prior to selecting useful features.

Naikal et al. in [70] proposed an offline informative feature selection for object recognition problem and suggest utilizing Sparse PCA (SPCA). To select informative visual words corresponding to a specific category/object an empirical covariance matrix is first computed for each object category in the database and SPCA is applied on this category dependent covariance matrix. Authors observe that first two principal vectors are enough for selecting informative features for foreground dominant objects. So, non-zero entries in the principal vectors are selected as the informative visual words. With this scheme in their experiments 405 visual words out of 1000 are identified as informative words in a dataset of 33 categories. In the experiments a 5% increase in the recognition rate is observed compared with blind training of the categories. Moreover, authors only work with small vocabularies, making the solution unsuitable for large datasets. On the other Wang et al. proposes to rank visual features according to the *tf-idf* scores computed from the same class [71]. However, much information is left readers imagination. Nonetheless, all mentioned methods till now, require multiple instances of the same object, scene, location or category.

In [72], Tolias et al. proposed to select features that display self-similarity properties within a single image. Authors first apply self-matching methods from the literature between an image and either itself or its reflection, where tentative feature correspondences are found in the descriptor space, without quantization. As a result, repeating patterns and local symmetries are detected and indexed. This is because of the observation that features repeating within a single image are likely to repeat across different

views too, so repeating features are good candidates for a selection without sacrificing recognition performance. The same detection performance is obtained with only a small fraction of its index size compared to the full feature index on a building/urban dataset. It should be noted that at feature selection step a considerable computation overhead is introduced. Furthermore, authors' main concern is the reproducibility of the indexed features on the other hand in large visual datasets discriminative abilities of index features are more important.

4.3 Informative Feature Selection

Features occurring in a reference frame that is rarely observed in other frames can be intuitively thought as informative features. And informativeness of a visual word for a given frame can be measured by estimating the mutual information $I(f_k; w_i)$ between the visual word w_i and the frame f_k. In general, mutual information is a measure of the dependence between two random variables, in this case the frame f_k and the visual word w_i. It expresses the quantity of information one has obtained on f_k by observing w_i. Although Figure 4.1 depicts the relation between mutual information and entropy of variables as an information diagram.

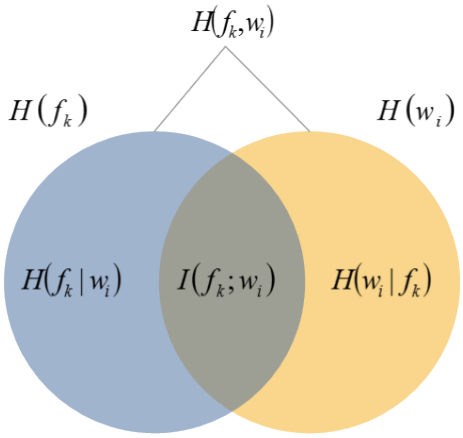


Figure 4.1: Mutual information diagram.

In Equation (4.3) formal definition of mutual information is given and visual words

maximizing this mutual information can be selected as the informative features.

$$\mathcal{L} = \arg\max_{w_i} I(f_k; w_i) \tag{4.2}$$

$$I(f_k; w_i) = H(f_k) - H(f_k|w_i) \tag{4.3}$$

where $H = -\sum_{i=1}^{K} p_i \log p_i$ is the discrete Shannon entropy. Since in (4.3) $H(f_k)$ is constant across all codewords we obtain;

$$\begin{aligned} I(f_k; w_i) &= \mathcal{C} - H(f_k|w_i) \\ I(f_k; w_i) &\propto -H(f_k|w_i) \end{aligned} \tag{4.4}$$

Finally;

$$\mathcal{L} = \arg\max_{w_i} I(f_k; w_i) \equiv \arg\min_{w_i} H(f_k|w_i) \tag{4.5}$$

Hence for a given frame and its corresponding visual words, informative features are the ones which minimizes the conditional entropy. From the definition of conditional entropy equations (4.6) are obtained in which $w_i = 0, 1$ represents absence or occurrence of visual word w_i, respectively.

$$\begin{aligned} H(f_k|w_i) &= \sum_{k=0,1} P(w_i = k) H(f_k|w_i = k) \\ &= P(w_i = 1) H(f_k|w_i = 1) + P(w_i = 0) H(f_k|w_i = 0) \end{aligned} \tag{4.6}$$

Considering the previous discussion, it can be seen that informativeness of a visual word can be measured by a simple function. One approach is to estimate observation probabilities and thus the conditional entropy in an offline manner after every feature of the reference database is collected. On the other hand in Algorithm 3 a sequential (online) method is provided in which features are selected iteratively for each frame.

This way an inverted index can be populated online.

input : Given a frame f_k and corresponding set of visual words $W_k = \{w_i|_{i=1}^{K}\ w_i \in V\}$, visual vocabulary V, an entropy estimator $\hat{H}$, subset size S
output: L, A list of informative codewords of f_k ($|L| = S$)

1 $L \leftarrow \emptyset$
2 **while** $|L| < S$ **do**
3 $\quad g^* \leftarrow MAXFLT$
4 $\quad$ **for** $\forall w_i \in W_k$ **do**
5 $\quad\quad g \leftarrow$ `ComputeConditionalEntropy(`f_k, w_i`)`
6 $\quad\quad$ **if** $g < g^*$ **then**
7 $\quad\quad\quad (g^*, w^*) \leftarrow (g, w_i)$
8 $\quad\quad$ **end**
9 $\quad$ **end**
10 $\quad L \leftarrow L \cup w^*$
11 $\quad W_k \leftarrow L \setminus w^*$
12 **end**
13 **return** L

Algorithm 3: Information-Gain Based Sequential Feature Selection

In line 5 of the Algorithm 3, conditional entropy as in (4.6) is computed for each word extracted from the frame f_k. Since $H(f_k|w_i)$ is unknown it is typically estimated from the observed samples. The most commonly used entropy estimation method is derived from the maximum likelihood estimates of the discrete probabilities. More specifically each of the unknown probabilities in equations (4.6) can be estimated by employing a frequentist approach. Consequently, (4.7) is obtained in which the total number of features in the database is N_{total} whereas N_{w_i} is the number of features belonging to codeword w_i. On the other hand the number of times codeword w_i observed in frame f_k is represented by $N_{f_k w_i}$. And finally, N_{f_k} is the total number of features in frame f_k.

$$
\begin{aligned}
\hat{P}(w_i = 1) &= \frac{N_{w_i}}{N_{total}} \\
\hat{P}(w_i = 0) &= 1 - \hat{P}(w_i = 1)
\end{aligned}
$$

$$
\begin{aligned}
\hat{P}(f_k|w_i = 1) &= \frac{N_{f_k w_i}}{N_{w_i}} \\
\hat{P}(f'_k|w_i = 1) &= 1 - \hat{P}(f_k|w_i = 1)
\end{aligned}
\tag{4.7}
$$

$$
\begin{aligned}
\hat{P}(f_k|w_i = 0) &= \frac{N_{f_k} - N_{f_k w_i}}{N_{total} - N_{w_i}} \\
\hat{P}(f'_k|w_i = 0) &= 1 - \hat{P}(f_k|w_i = 0)
\end{aligned}
$$

4.4 Improved Mutual Information-based Feature Selection

The method introduced in the previous section for entropy estimation is called a plug-in estimator, where a function is evaluated on an estimated probability distribution. Generally, when we consider the total number of observations N and the number of occurrences of i in the ensemble be n_i, then with the choice of $\hat{p}_i = \frac{n_i}{N}$ we obtain the naive estimate

$$
\begin{aligned}
\hat{H} &= -\sum_{i=1}^{K} \hat{p}_i \log \hat{p}_i \\
&= \log N - \frac{1}{N} \sum_{i=1}^{K} n_i \log n_i
\end{aligned}
\tag{4.8}
$$

However, (4.8) is biased and leads to a systematic underestimation of the entropy H [73]. A detailed computation of the expectation value of $\hat{H}$ with respect to the multinomial distribution

$$
p(n_1, ..., n_K, p_1, ..., p_K, N) = N! \prod_{i=1}^{K} \frac{p_i^{n_i}}{n_i} \tag{4.9}
$$

up to the second order in N was given first by Harris [74].

$$\mathbb{E}[\hat{H}] = H - \frac{K-1}{2N} + \frac{1}{12N^2}\left(1 - \sum_{i=1}^{K}\frac{1}{p_k}\right) + \mathcal{O}(N^{-3}) \tag{4.10}$$

The first correction term $\mathcal{O}(1/N)$ can be evaluated fairly easily and was first obtained by Miller [75]. On the other hand second and higher-terms of the bias depend on the unknown true probabilities p_i and can not be estimated reliably. By employing the first term in (4.10) as a correction, we obtain the Miller-adjusted entropy estimate $\hat{H}_M = \hat{H} + \frac{K-1}{N}$. However, when we use it to evaluate (4.6) as in (4.11) we obtain a constant correction term on the naive estimate that effects all of the words in the same way. So, although we may improve the entropy estimate, the Miller correction has no effect on the selection of words.

$$\begin{aligned}
\hat{H}_M(f_k|w_i) &= \sum_{k=0,1}\hat{P}(w_i=k)\hat{H}_M(f_k|w_i=k) \\
&= \hat{P}(w_i=1)\hat{H}_M(f_k|w_i=1) + \hat{P}(w_i=0)\hat{H}_M(f_k|w_i=0) \\
&= \frac{N_{w_i}}{N_{total}}\left(\hat{H}(f_k|w_i=1) + \frac{1}{2N_{w_i}}\right) \\
&+ \frac{N_{total}-N_{w_i}}{N_{total}}\left(\hat{H}((f_k|w_i=0) + \frac{1}{2(N_{total}-N_{w_i})}\right) \\
&= \frac{N_{w_i}}{N_{total}}\hat{H}(f_k|w_i=1) + \frac{N_{total}-N_{w_i}}{N_{total}}\hat{H}((f_k|w_i=0) + \frac{1}{N_{total}} \\
&= \hat{H}(f_k|w_i) + \mathcal{C}
\end{aligned} \tag{4.11}$$

In the literature, there are other methods proposed to improve the entropy estimate. A detailed analysis of different estimators has been reported in Schürmann's [73]. From that analysis it can be seen that Grassberger estimate has a superior performance in both of the absolute bias and the statistical error. Grassberger estimates is a family of discrete entropy estimators derived from the assumption of Poisson distributed frequencies. The corresponding estimator is given as;

$$\hat{H}_G = \log N - \frac{1}{N}\sum_{i=1}^{K} n_i G(n_i) \tag{4.12}$$

where the function $G(n)$ is given in closed-form as;

$$G(n_i) = \psi(n_i) + \frac{1}{2}(-1)^{n_i}\left(\psi(\frac{n_i+1}{2}) - \psi(\frac{n_i}{2})\right) \tag{4.13}$$

In equation (4.13), the digamma function $\psi(n)$ is the logarithmic derivative of the Γ-function, i.e.: $\psi(x) = \frac{d}{dx}\ln\Gamma(x)$. For large values of n $(n \to \infty)$ $G(n)$ converges to the logarithm. On the other hand, for small values of n $G(n)$ behaves differently and hence the $\hat{H}_G$ expected to be more accurate then the naive estimate $\hat{H}$. And this improvement comes without an additional computational complexity. Thus line 5 in Algorithm 3 is replaced with Grassberger entropy estimator to obtain Algorithm 4.

input : Given a frame f_k and corresponding set of visual words $W_k = \{w_i|_{i=1}^{K}\ w_i \in V\}$, visual vocabulary V, an entropy estimator $\hat{H}$, subset size S
output: L, A list of informative codewords of f_k ($|L| = S$)

```
1  L ← ∅
2  while |L| < S do
3      g* ← MAXFLT
4      for ∀w_i ∈ W_k do
5          g ← ComputeGrassbergerConditionalEntropy(f_k, w_i)
6          if g < g* then
7              (g*, w*) ← (g, w_i)
8          end
9      end
10     L ← L ∪ w*
11     W_k ← L \ w*
12 end
13 return L
```

Algorithm 4: Improved feature selection algorithm by using Grassberger entropy estimator.

4.4.1 Evaluation of Entropy Estimation Methods

In order to compare both estimation methods, random samples are generated from various known distributions. From these samples, entropies are estimated and compared with true entropies. Two distributions are of interest in this experiment because of their relation with bag-of-words approach, namely binomial and Poisson distributions. If we consider assignment of a feature to a particular word as a coin toss with probability p than for n features the bag-of-words histogram follows a binomial

distribution, $B(n, p)$, under the i.i.d assumptions. On the other hand binomial distribution can be accurately approximated with the Poisson distribution, $Pois(\lambda)$ for faster computation if n is large and p is sufficiently small. This can be easily seen from probability mass functions of binomial (4.15) and Poisson distributions (4.14). Because of this property Poisson distribution assumption is frequently made in document analysis problems.

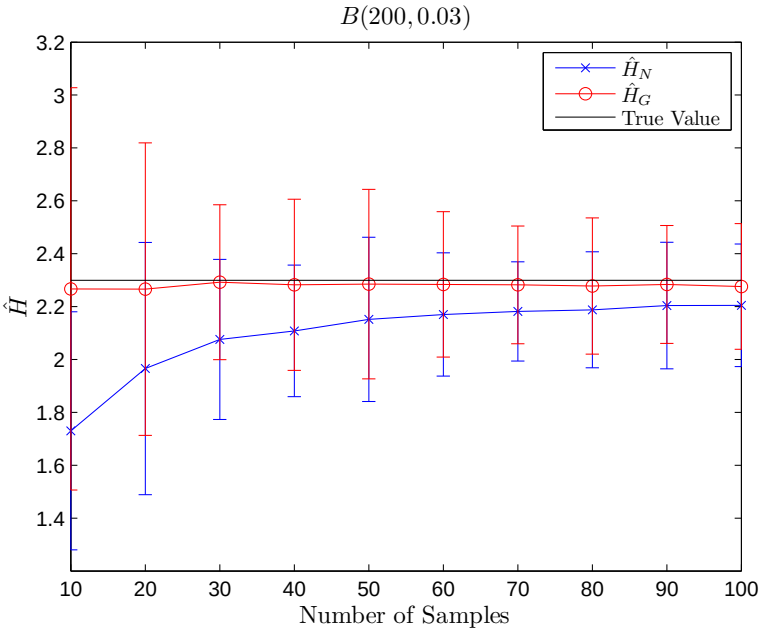
$$Pr(X = k) = \frac{\lambda^k \mathrm{e}^{-\lambda}}{k!} \tag{4.14}$$

$$Pr(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \tag{4.15}$$

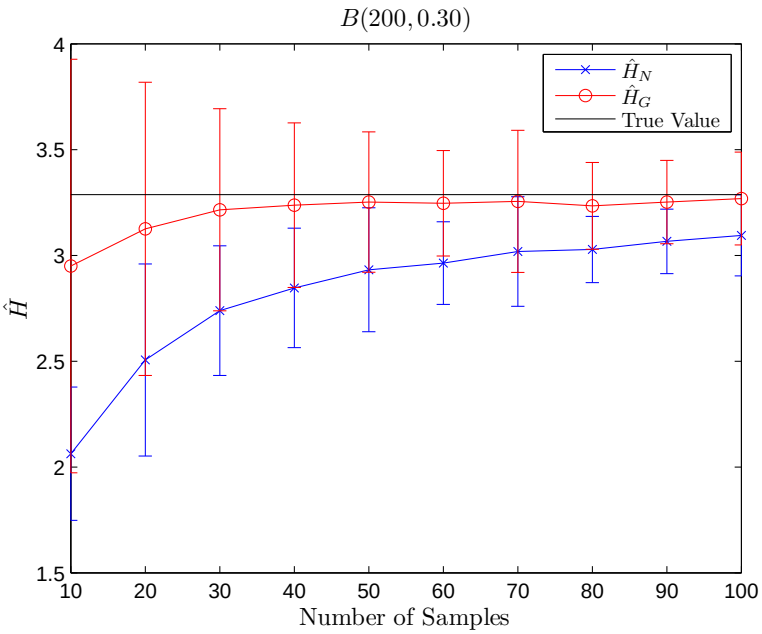
In Figure 4.2 and Figure 4.4 entropy estimations are depicted for varying sample sizes and distribution parameters. Furthermore, mean square error of estimators are shown in Figure 4.3 and Figure 4.5. As it can be seen from both set of figures Grassberger estimate is superior to the naive approach. It should be noted that true entropy associated with binomial and Poisson distributions contain closed form infinite sums. However, there are fairly accurate approximations of these entropies [76] as given in (4.16) and (4.17) which are employed here to obtain true entropies and corresponding mean square error values. Another point worth mentioning is the behavior of estimates when p changes for binomial distribution. It can be seen from Figure 4.2a and Figure 4.2b that for sufficiently small p both estimators work better. Actually this is the case for many BoVW applications when a large codebook is exploited. Conversely, the performance of the estimators decrease as probability, p increases. This observation is due to the fact that as p increases the Poisson approximation of Binomial distribution worsens causing large fluctuations. Moreover, as the p value decreases performance of the Grassberger estimate $\hat{H}_G$ increases.

$$H_B \sim \frac{1}{2} log(2\pi enp(1-p)) + O(1/n) \tag{4.16}$$

$$H_P \sim \frac{1}{2} log(2\pi e\lambda) - \frac{1}{12\lambda} \tag{4.17}$$

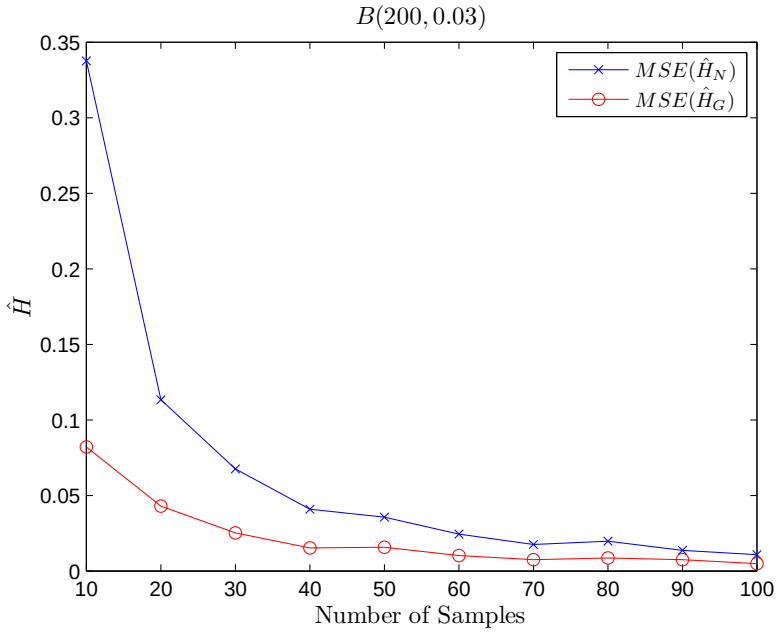


(a)

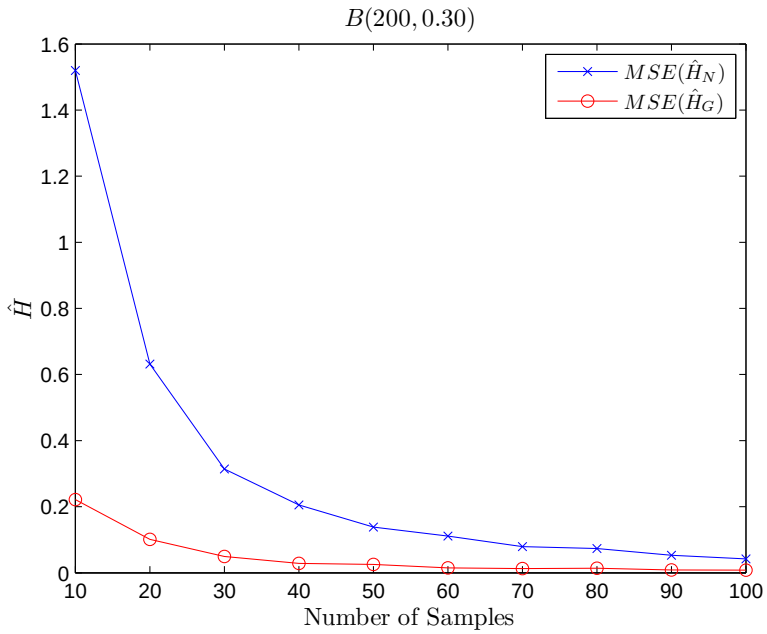


(b)

Figure 4.2: Entropy estimates of binomially distributed random variables (a) $B(200, 0.03)$ and (b) $B(200, 0.3)$ for different number of samples.

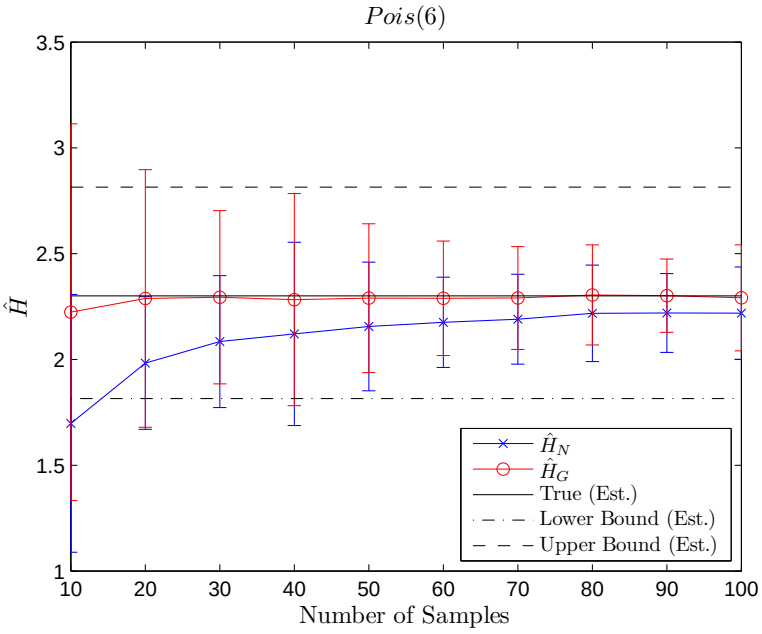


(a)

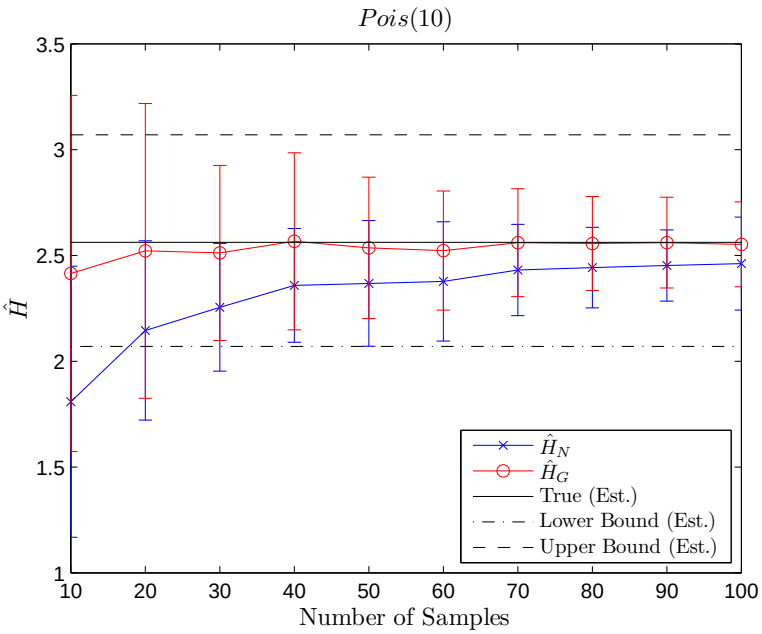


(b)

Figure 4.3: Mean square error for entropy estimates of (a) $B(200, 0.03)$ and (b) $B(200, 0.3)$ for different number of samples.

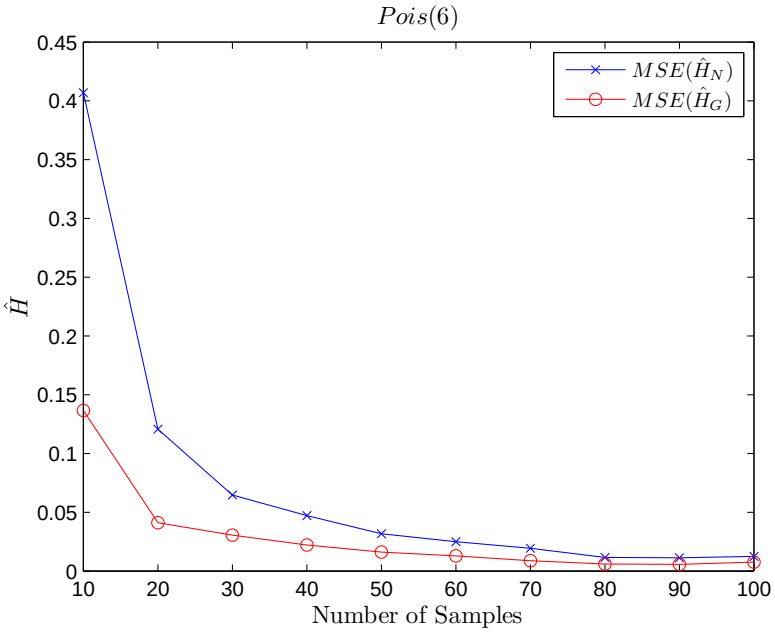


(a)

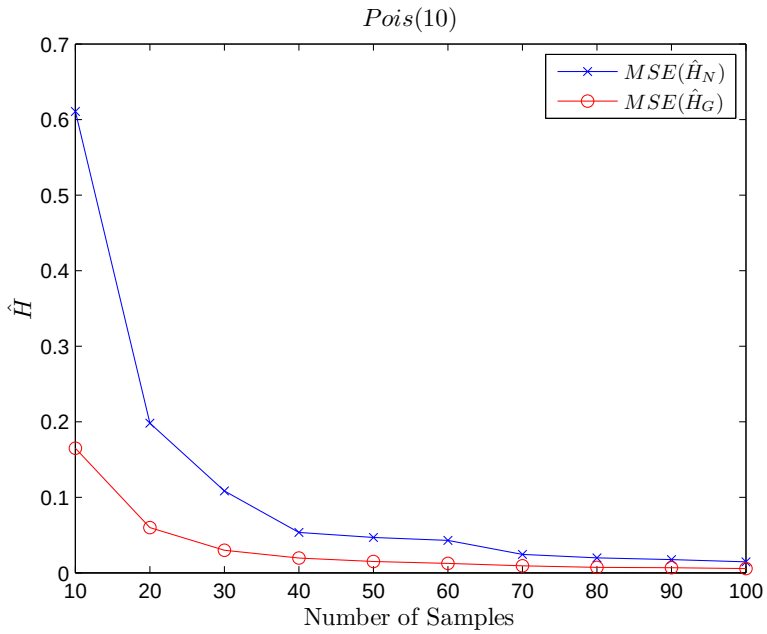


(b)

Figure 4.4: Entropy estimates of Poisson distributed random variables (a) $Pois(6)$ and (b) $Pois(10)$ for different number of samples.



(a)



(b)

Figure 4.5: Mean square error for entropy estimates of (a) $Pois(6)$ and (b) $Pois(10)$ for different number of samples.

4.5 Experimental Evaluation of Informative Feature-Based Indexing

Experimental setup is prepared much like in Section 3.2.6 with the same datasets. Uniform sampling is performed on reference and query videos with the asymmetric sampling rates, 1.5 fps and 5fps respectively. SIFT descriptor is utilized with a codebook of 100K words for visual representation. As in 3.2.6, product quantization is adopted with 256 codewords to obtain 64 bit detail fingerprint for each feature. Queries are also flipped and searched in the reference database concurrently in order to address flip-like transformations. At the indexing stage, Algorithm 3 is adopted with different cutoff values and entropy estimation methods on TRECVID 2009 and 2010 CCD dataset. Table 4.1 and 4.2 give key points about the experiments. Entropy estimators introduced previously are also compared in Table 4.3-4.8 by reporting NDCR differences (Δ) observed. The timings are given as a speedup factor w.r.t. detection time in *full index*.

Table 4.1: Summary of experiments in TRECVID 2009 CCD dataset.

	TRECVID2009			
Raw Feature Count	$276,315,688$			
Cut-off	$T_c = 0.5$		$T_c = 0.8$	
Entropy Est.	Naïve	Proposed	Naïve	Proposed
Feature Count	$165,341,524$	$159,970,011$	$230,535,498$	$229,408,422$
Compaction Ratio	59.84%	57.89%	83.43%	83.02%
Timings	$2.40\times$	$2.49\times$	$1.67\times$	$1.68\times$
Table References	Table 4.3, 4.4, 4.5		Table 4.6, 4.7, 4.8	

Table 4.2: Summary of experiments in TRECVID 2010 CCD dataset.

	TRECVID2010	
Raw Feature Count	$243,000,059$	
Cut-off	$T_c = 0.5$	
Entropy Est.	Naïve	Proposed
Feature Count	$141,425,795$	$138,320,913$
Compaction Ratio	58.20%	56.92%
Timings	$2.00\times$	$2.04\times$
Table References	Table 4.9, 4.10, 4.11	

It can be seen from the results that by using less number of features same and even

for most transformations better performance is achieved compared to full indexing. Especially for *Pattern Insertion*, *Change of Gamma*, *Decrease of Quality* and *Post-Production* transformations, better performance is consistently achieved for all profiles on both datasets by informative feature-based indexing. As expected the relevance of the matching descriptors/frames is increased by using only informative features thus the precision performance of the copy detection algorithm is improved. This could be easily observed at Table 4.8 and 4.11 in which no-false-alarm profile results are reported. Although when results on TRECVID 2010 CCD dataset are examined a performance degradation is observed notably for *Camcording*, *Heavy Re-encoding* and *Picture-in-Picture* transformations. As discussed in Section 3.2.7, on the average 126 SIFT features are extracted from every frame which is considerably low compared with related work. Unfortunately this *under-representation* has adverse effects on the performance of copy detection algorithm, specifically on TRECVID 2010 CCD dataset. Since as also examined in Section 2.4 dataset has many unstructured content furthermore, mentioned transformations have severe effects on the local feature reproducibility.

Additionally Grassberger estimate dominantly beats the Naive estimate in terms of NDCR performances achieved. From the results it can be seen that as the number of indexed features decrease ($T_c = 0.8$ vs. $T_c = 0.5$) performance gain of using Grassberger estimate is more evident. One interesting point observed is that with Grassberger entropy estimation always fewer features are selected as informative compared to Naive entropy estimation. Lastly, since a fraction of features are indexed a shorter inverted list is queried. Thus memory usage efficiency is improved and coupled with having less distractor features, matching is achieved much faster.

Table 4.3: Evaluation results obtained with $C_M = 10$, $C_{FA} = 1$ and $T_c = 0.5$ at TRECVID 2009 CCD dataset. Total number of selected features in the case of Naive and Grassberger entropy estimators are $165,341,524$ and $159,970,011$ respectively.

NDCR	**2009**	**2009 – $T_c = 0.5$**		
	Original	**Naïve**	**Proposed**	Δ
T1 (Camcording)	NA	NA	NA	NA
T2 (PiP-Type1)	0.062	0.140	0.160	-0.020
T3 (Pat. Ins.)	0.043	0.018	0.007	0.011
T4 (Strong Re.)	0.140	0.235	0.232	0.003
T5 (Ch. Gamma)	0.060	0.085	0.088	-0.003
T6 (Dec.Q.3)	0.015	0.018	0.000	0.018
T7 (Dec.Q.5)	NA	NA	NA	NA
T8 (PP.3)	0.065	0.060	0.050	0.010
T9 (PP.5)	NA	NA	NA	NA
T10 (Rnd.5)	0.255	0.347	0.344	0.003

Table 4.4: Evaluation results obtained with $C_M = 1$, $C_{FA} = 1$ and $T_c = 0.5$ at TRECVID 2009 CCD dataset. Total number of selected features in the case of Naive and Grassberger entropy estimators are $165,341,524$ and $159,970,011$ respectively.

NDCR	**2009**	**2009 – $T_c = 0.5$**		
	Original	**Naïve**	**Proposed**	Δ
T1 (Camcording)	NA	NA	NA	NA
T2 (PiP-Type1)	0.153	0.149	0.172	-0.023
T3 (Pat. Ins.)	0.052	0.022	0.007	0.015
T4 (Strong Re.)	0.373	0.459	0.343	0.116
T5 (Ch. Gamma)	0.060	0.243	0.231	0.012
T6 (Dec.Q.3)	0.015	0.037	0.000	0.037
T7 (Dec.Q.5)	NA	NA	NA	NA
T8 (PP.3)	0.161	0.112	0.112	0.000
T9 (PP.5)	NA	NA	NA	NA
T10 (Rnd.5)	0.418	0.519	0.526	-0.007

Table 4.5: Evaluation results obtained with $C_M = 1$, $C_{FA} = 1000$ and $T_c = 0.5$ at TRECVID 2009 CCD dataset. Total number of selected features in the case of Naive and Grassberger entropy estimators are $165,341,524$ and $159,970,011$ respectively.

NDCR	**2009**	**2009 – $T_c = 0.5$**		
	Original	**Naïve**	**Proposed**	Δ
T1 (Camcording)	NA	NA	NA	NA
T2 (PiP-Type1)	0.224	0.149	0.172	-0.023
T3 (Pat. Ins.)	0.052	0.022	0.007	0.015
T4 (Strong Re.)	0.373	0.597	0.343	0.254
T5 (Ch. Gamma)	0.060	0.306	0.231	0.075
T6 (Dec.Q.3)	0.015	0.037	0.000	0.037
T7 (Dec.Q.5)	NA	NA	NA	NA
T8 (PP.3)	0.224	0.112	0.112	0.00
T9 (PP.5)	NA	NA	NA	NA
T10 (Rnd.5)	0.418	0.552	0.545	0.007

Table 4.6: Evaluation results obtained with $C_M = 10$, $C_{FA} = 1$ and $T_c = 0.8$ at TRECVID 2009 CCD dataset. Total number of selected features in the case of Naive and Grassberger entropy estimators are $230,535,498$ and $229,408,422$ respectively.

NDCR	**2009**	**2009 – 80%**		
	Original	**Naïve**	**Proposed**	Δ
T1 (Camcording)	NA	NA	NA	NA
T2 (PiP-Type1)	0.062	0.080	0.108	-0.028
T3 (Pat. Ins.)	0.043	0.025	0.025	0.000
T4 (Strong Re.)	0.140	0.183	0.175	0.008
T5 (Ch. Gamma)	0.060	0.037	0.047	-0.010
T6 (Dec.Q.3)	0.015	0.007	0.025	-0.018
T7 (Dec.Q.5)	NA	NA	NA	NA
T8 (PP.3)	0.065	0.037	0.055	-0.018
T9 (PP.5)	NA	NA	NA	NA
T10 (Rnd.5)	0.255	0.304	0.294	0.010

Table 4.7: Evaluation results obtained with $C_M = 10$, $C_{FA} = 1$ and $T_c = 0.8$ at TRECVID 2009 CCD dataset. Total number of features in the case of Naive and Grassberger entropy estimators are $230,535,498$ and $229,408,422$ respectively.

NDCR	**2009**	**2009 – 80%**		
	Original	**Naïve**	**Proposed**	Δ
T1 (Camcording)	NA	NA	NA	NA
T2 (PiP-Type1)	0.153	0.209	0.220	-0.011
T3 (Pat. Ins.)	0.052	0.075	0.075	0.000
T4 (Strong Re.)	0.373	0.381	0.444	-0.063
T5 (Ch. Gamma)	0.060	0.037	0.060	-0.023
T6 (Dec.Q.3)	0.015	0.007	0.030	-0.023
T7 (Dec.Q.5)	NA	NA	NA	NA
T8 (PP.3)	0.161	0.037	0.146	-0.109
T9 (PP.5)	NA	NA	NA	NA
T10 (Rnd.5)	0.418	0.538	0.403	0.135

Table 4.8: Evaluation results obtained with $C_M = 1$, $C_{FA} = 1000$ and $T_c = 0.8$ at TRECVID 2009 CCD dataset. Total number of features in the case of Naive and Grassberger entropy estimators are $230,535,498$ and $229,408,422$ respectively.

NDCR	**2009**	**2009 – 80%**		
	Original	**Naïve**	**Proposed**	Δ
T1 (Camcording)	NA	NA	NA	NA
T2 (PiP-Type1)	0.224	0.209	0.231	-0.022
T3 (Pat. Ins.)	0.052	0.075	0.075	0.000
T4 (Strong Re.)	0.373	0.381	0.522	-0.141
T5 (Ch. Gamma)	0.060	0.037	0.060	-0.023
T6 (Dec.Q.3)	0.015	0.007	0.030	-0.023
T7 (Dec.Q.5)	NA	NA	NA	NA
T8 (PP.3)	0.224	0.037	0.261	-0.224
T9 (PP.5)	NA	NA	NA	NA
T10 (Rnd.5)	0.418	0.552	0.403	0.149

Table 4.9: Evaluation results obtained with $C_M = 10$, $C_{FA} = 1$ and $T_c = 0.5$ at TRECVID 2010 CCD dataset. Total number of selected features in the case of Naive and Grassberger entropy estimators are $141,425,795$ and $138,320,913$ respectively.

NDCR	**2010**	**2010 - $T_c = 0.5$**		
	Original	**Naïve**	**Proposed**	Δ
T1 (Camcording)	0.518	0.778	0.792	-0.014
T2 (PiP-Type1)	0.257	0.379	0.368	0.011
T3 (Pat. Ins.)	0.126	0.099	0.077	0.022
T4 (Strong Re.)	0.270	0.430	0.435	-0.005
T5 (Ch. Gamma)	0.140	0.171	0.168	0.003
T6 (Dec.Q.3)	0.304	0.416	0.402	0.014
T7 (Dec.Q.5)	NA	NA	NA	NA
T8 (PP.3)	0.143	0.150	0.150	0.000
T9 (PP.5)	NA	NA	NA	NA
T10 (Rnd.5)	0.399	0.453	0.412	0.041

Table 4.10: Evaluation results obtained with $C_M = 1$, $C_{FA} = 1$ and $T_c = 0.5$ at TRECVID 2010 CCD dataset. Total number of selected features in the case of Naive and Grassberger entropy estimators are $141,425,795$ and $138,320,913$ respectively.

NDCR	**2010**	**2010 - $T_c = 0.5$**		
	Original	**Naïve**	**Proposed**	Δ
T1 (Camcording)	0.846	0.933	0.921	0.012
T2 (PiP-Type1)	0.541	0.733	0.718	0.015
T3 (Pat. Ins.)	0.187	0.264	0.164	0.100
T4 (Strong Re.)	0.660	0.771	0.682	0.089
T5 (Ch. Gamma)	0.336	0.358	0.336	0.022
T6 (Dec.Q.3)	0.694	0.836	0.712	0.124
T7 (Dec.Q.5)	NA	NA	NA	NA
T8 (PP.3)	0.286	0.331	0.328	0.003
T9 (PP.5)	NA	NA	NA	NA
T10 (Rnd.5)	0.642	0.607	0.592	0.015

Table 4.11: Evaluation results obtained with $C_M = 1$, $C_{FA} = 1000$ and $T_c = 0.5$ at TRECVID 2010 CCD dataset. Total number of selected features in the case of Naive and Grassberger entropy estimators are $141,425,795$ and $138,320,913$ respectively.

NDCR	**2010**	**2010 - $T_c = 0.5$**		
	Original	**Naïve**	**Proposed**	Δ
T1 (Camcording)	0.866	0.933	0.933	0.000
T2 (PiP-Type1)	0.902	0.887	0.887	0.000
T3 (Pat. Ins.)	0.187	0.299	0.164	0.135
T4 (Strong Re.)	0.888	0.881	0.873	0.008
T5 (Ch. Gamma)	0.336	0.358	0.336	0.022
T6 (Dec.Q.3)	0.694	0.836	0.799	0.037
T7 (Dec.Q.5)	NA	NA	NA	NA
T8 (PP.3)	0.343	0.336	0.328	0.008
T9 (PP.5)	NA	NA	NA	NA
T10 (Rnd.5)	0.642	0.634	0.612	0.022

4.6 Improving Information Theoretic Indexing by Exploitation of Temporal Dependencies

In the previously discussed approach, frame features are selected according to their informativeness by measuring the information gain of corresponding visual words. However, in aforementioned method frames are considered independent from each other. Actually it is known that most of the time i.i.d. assumption does not hold for BoVW, especially for the collective distribution of words from consecutive frames. Furthermore, when experiments involving random insertion of reference frames conducted it has been observed that higher performance is achieved with successive frame insertion compared to random insertion. One way to exploit this is to select features from joint distribution of consecutive frame features in a temporal video volume. Thus instead of finding a subset of features by (4.2), the mutual information over consecutive frames is utilized.

$$\mathcal{L}_T = \arg\max_{w_i} I(\mathbf{f}_T; w_i) \tag{4.18}$$

where $f_T = (f_k, f_{k+1}, ..., f_{k+T-1})$. In this approach for each visual word, w_i discriminative/informative features for a given temporal volume are selected. In a way temporally consistent informative features over multiple frames is captured without any additional computational overhead.

In the literature there are different approaches utilizing temporal information. Willems et al. [22] extract spatio-temporal local features. On the other hand Kim et al. utilize a global feature extracted from a temporal volume [37] much like [38]. Also Law-To et al. in [8] exploit temporal information by labeling the trajectory of individual local features. Contrary to those approaches aforementioned method utilizes temporal information according to the consistency of local features' informativeness over a temporal volume.

4.6.1 Experiments and Discussions

Following experimental results are obtained on TRECVID 2009 and 2010 CCD dataset. Informative features are selected in a temporal volume for varying temporal lengths (T_V) and throughout the experiments compaction ratio, T_c, is fixed at $1/2$. Results are depicted in Table 4.12-4.14 for TRECVID 2009 dataset and in Table 4.15-4.17 for TRECVID 2010 dataset. As it can be seen from results for most of the transformations temporal indexing does not improve performance considerably compared to full indexing. However, when compared with informative feature indexing from individual frames, merits of temporal indexing are obvious as depicted in Table 4.18 and 4.19. Furthermore, as the temporal window length increases the detection performance improves since *discriminative* consistency among local features are better captured from larger temporal volume. It should be also noted that expected detection results are not observed in TRECVID 2010 dataset. This is mostly because of the reference video characteristics such as fast changing scenes and single shot videos.

Table 4.12: Evaluation results of informative feature selection in a temporal volume obtained with $T_c = 0.5$ and $T_V = 2$ at TRECVID 2009 CCD dataset.

NDCR	**2009**	**2009 – $T_c = 0.5$ – $T_V = 2$**		
	Original	**Naïve**	**Proposed**	Δ
T1 (Camcording)	NA	NA	NA	NA
T2 (PiP-Type1)	0.062	0.085	0.113	-0.028
T3 (Pat. Ins.)	0.043	0.065	0.053	0.012
T4 (Strong Re.)	0.140	0.160	0.148	0.012
T5 (Ch. Gamma)	0.060	0.058	0.055	0.003
T6 (Dec.Q.3)	0.015	0.032	0.025	0.007
T7 (Dec.Q.5)	NA	NA	NA	NA
T8 (PP.3)	0.065	0.078	0.065	0.013
T9 (PP.5)	NA	NA	NA	NA
T10 (Rnd.5)	0.255	0.302	0.306	-0.004

Table 4.13: Evaluation results of informative feature selection in a temporal volume obtained with $T_c = 0.5$ and $T_V = 3$ at TRECVID 2009 CCD dataset.

NDCR	**2009**	**2009 – $T_c = 0.5$ – $T_V = 3$**		
	Original	**Naïve**	**Proposed**	**Δ**
T1 (Camcording)	NA	NA	NA	NA
T2 (PiP-Type1)	0.062	0.093	0.093	0.000
T3 (Pat. Ins.)	0.043	0.058	0.065	-0.007
T4 (Strong Re.)	0.140	0.160	0.170	-0.010
T5 (Ch. Gamma)	0.060	0.045	0.045	0.000
T6 (Dec.Q.3)	0.015	0.045	0.025	0.020
T7 (Dec.Q.5)	NA	NA	NA	NA
T8 (PP.3)	0.065	0.058	0.058	0.000
T9 (PP.5)	NA	NA	NA	NA
T10 (Rnd.5)	0.255	0.307	0.314	-0.007

Table 4.14: Evaluation results of informative feature selection in a temporal volume obtained with $T_c = 0.5$ and $T_V = 6$ at TRECVID 2009 CCD dataset.

NDCR	**2009**	**2009 – $T_c = 0.5$ – $T_V = 6$**		
	Original	**Naïve**	**Proposed**	**Δ**
T1 (Camcording)	NA	NA	NA	NA
T2 (PiP-Type1)	0.062	0.080	0.087	-0.007
T3 (Pat. Ins.)	0.043	0.058	0.058	0.000
T4 (Strong Re.)	0.140	0.188	0.166	0.022
T5 (Ch. Gamma)	0.060	0.058	0.058	0.000
T6 (Dec.Q.3)	0.015	0.050	0.045	0.005
T7 (Dec.Q.5)	NA	NA	NA	NA
T8 (PP.3)	0.065	0.050	0.050	0.000
T9 (PP.5)	NA	NA	NA	NA
T10 (Rnd.5)	0.255	0.269	0.262	0.007

Table 4.15: Evaluation results of informative feature selection in a temporal volume obtained with $T_c = 0.5$ and $T_V = 2$ at TRECVID 2010 CCD dataset.

NDCR	**2010**	**2010 – $T_c = 0.5$ – $T_V = 2$**		
	Original	**Naïve**	**Proposed**	**Δ**
T1 (Camcording)	0.518	0.711	0.709	0.002
T2 (PiP-Type1)	0.257	0.307	0.323	-0.016
T3 (Pat. Ins.)	0.126	0.123	0.104	0.019
T4 (Strong Re.)	0.270	0.365	0.366	-0.001
T5 (Ch. Gamma)	0.140	0.191	0.216	-0.025
T6 (Dec.Q.3)	0.304	0.396	0.391	0.005
T7 (Dec.Q.5)	NA	NA	NA	NA
T8 (PP.3)	0.143	0.182	0.161	0.021
T9 (PP.50	NA	NA	NA	NA
T10 (Rnd.5)	0.399	0.424	0.406	0.018

Table 4.16: Evaluation results of informative feature selection in a temporal volume obtained with $T_c = 0.5$ and $T_V = 3$ at TRECVID 2010 CCD dataset.

NDCR	**2010**	**2010 – $T_c = 0.5$ – $T_V = 3$**		
$C_M = 10, C_{FA} = 1$	**Original**	**Naïve**	**Proposed**	**Δ**
T1 (Camcording)	0.518	0.713	0.709	0.004
T2 (PiP-Type1)	0.257	0.301	0.311	-0.010
T3 (Pat. Ins.)	0.126	0.123	0.133	-0.010
T4 (Strong Re.)	0.270	0.357	0.347	0.010
T5 (Ch. Gamma)	0.140	0.221	0.224	-0.003
T6 (Dec.Q.3)	0.304	0.387	0.377	0.010
T7 (Dec.Q.5)	NA	NA	NA	NA
T8 (PP.3)	0.143	0.152	0.155	-0.003
T9 (PP.50	NA	NA	NA	NA
T10 (Rnd.5)	0.399	0.406	0.431	-0.025

Table 4.17: Evaluation results of informative feature selection in a temporal volume obtained with $T_c = 0.5$ and $T_V = 6$ at TRECVID 2010 CCD dataset.

NDCR	**2010**	**2010 – $T_c = 0.5$ – $T_V = 6$**		
	Original	**Naïve**	**Proposed**	**Δ**
T1 (Camcording)	0.518	0.639	0.600	0.039
T2 (PiP-Type1)	0.257	0.272	0.289	-0.017
T3 (Pat. Ins.)	0.126	0.123	0.123	0.000
T4 (Strong Re.)	0.270	0.339	0.344	-0.005
T5 (Ch. Gamma)	0.140	0.206	0.209	-0.003
T6 (Dec.Q.3)	0.304	0.351	0.366	-0.015
T7 (Dec.Q.5)	NA	NA	NA	NA
T8 (PP.3)	0.143	0.160	0.161	-0.001
T9 (PP.50	NA	NA	NA	NA
T10 (Rnd.5)	0.399	0.435	0.432	0.003

Table 4.18: Performance of temporal feature indexing on TRECVID 2009 CCD dataset for varying T_V compared with individual frame indexing while $T_c = 0.5$. Proposed Grassberger entropy estimator is utilized.

	Single Frame	**Volume**		
		$T_V = 2$	$T_V = 3$	$T_V = 6$
T1 (Camcording)	NA	NA	NA	NA
T2 (PiP-Type1)	0.160	0.113	0.093	0.087
T3 (Pat. Ins.)	0.007	0.053	0.065	0.058
T4 (Strong Re.)	0.232	0.148	0.170	0.166
T5 (Ch. Gamma)	0.088	0.055	0.045	0.058
T6 (Dec.Q.3)	0.000	0.025	0.025	0.045
T7 (Dec.Q.5)	NA	NA	NA	NA
T8 (PP.3)	0.050	0.065	0.058	0.050
T9 (PP.5)	NA	NA	NA	NA
T10 (Rnd.5)	0.344	0.306	0.314	0.262

Table 4.19: Performance of temporal feature indexing on TRECVID 2010 CCD dataset for varying T_V compared with individual frame indexing while $T_c = 0.5$. Proposed Grassberger entropy estimator is utilized.

NDCR	**Single Frame**	**Volume**		
		$T_V = 2$	$T_V = 3$	$T_V = 6$
T1 (Camcording)	0.518	0.709	0.709	0.600
T2 (PiP-Type1)	0.257	0.323	0.311	0.289
T3 (Pat. Ins.)	0.126	0.104	0.133	0.123
T4 (Strong Re.)	0.270	0.366	0.347	0.344
T5 (Ch. Gamma)	0.140	0.216	0.224	0.209
T6 (Dec.Q.3)	0.304	0.391	0.377	0.366
T7 (Dec.Q.5)	NA	NA	NA	NA
T8 (PP.3)	0.143	0.161	0.155	0.161
T9 (PP.50	NA	NA	NA	NA
T10 (Rnd.5)	0.399	0.406	0.431	0.432

CHAPTER 5

INFORMATION AND INTERACTION AMONG FEATURES

In the chapter informative features are selected by measuring how much information one visual word conveys about the reference frame. Basically, to effectively *discriminate* f_k, whose visual words are $\{w_i\}_{i=1}^{N}$, is to select most informative features by measuring mutual information, $I(f_k; w_i)$.

Generally speaking, mutual information is a multivariate measure that can easily capture the dependence of multiple variables. When two variable/codeowrd case is considered $\mathbf{W} = (w_i, w_j)$, the information diagram of Figure 5.1 suggests that there are different ways to capture this dependence.

1. **Multivariable Mutual Information:** An obvious extension of single variable case is to compute mutual information between f_k and random variable $\mathbf{W} = (w_i, w_j)$ by Equation (5.1). This solution is depicted in Figure 5.1a.

$$I(f_k; \mathbf{W}) = H(f_k) - H(f_k|\mathbf{W}) \tag{5.1}$$

2. **Multivariate Mutual Information:** Another way of capturing dependence of multiple variable is to incorporate the *interaction* between variables w_i and w_j. This is easily achieved by computing $I(f_k; w_i; w_j)$ as represented in Figure 5.1b. By using basic identities a recognizable form of $I(f_k; w_i; w_j)$ is obtained

in Equation (5.2). This measure is also called *interaction information*.

$$\begin{aligned} I(f_k; w_i; w_j) &= I(f_k; w_i|w_j) - I(f_k; w_i) \\ &= I(f_k; w_i) + I(f_k; w_j) - I(f_k; w_i, w_j) \\ &= H(f_k) - H(f_k|w_i) + H(f_k) - H(f_k|w_j) - H(f_k|w_i, w_j) \\ &= H(f_k) + H(f_k|w_i, w_j) - H(f_k|w_i) - H(f_k|w_j) \end{aligned} \tag{5.2}$$

3. **Total Correlation:** Lastly, there is the *total correlation* $C(f_k, w_i, w_j)$ which describes the total amount of dependence between all variables including f_k as shown in Figure 5.1c.

$$C(f_k, w_i, w_j) = H(f_k) + H(w_i) + H(w_j) - H(f_k, w_i, w_j) \tag{5.3}$$

Interaction information can be viewed as the amount of information that is common between all variables but not present in any subset as depicted in Figure 5.1b. Unlike univariate mutual information it can be negative or positive but it is always symmetric i.e.: $I(f_k; w_i; w_j) = I(f_k; w_j; w_i)$. On the other hand total correlation is nonnegative and equals to zero if and only if all variables are independent. However, it will be nonzero if only a pair of variables are dependent. For example, if $P(f_k, w_i, w_j) = P(f_k, w_i)P(w_j)$ then total correlation will be non-zero so it can not be claimed that there is an interaction among all three variable. But, for such a case, interaction information will be zero.

Since both $I(f_k; w_i|w_j)$ and $I(f_k; w_i)$ are nonnegative, $I(f_k; w_i; w_j)$ is positive if $I(f_k; w_i|w_j) > I(f_k; w_i)$ and negative when the inequality is the other way. It is called *positive interaction* if $I(f_k; w_i; w_j) > 0$ and *negative interaction* if $I(f_k; w_i; w_j) < 0$.

Assume that there is an uncertainty about f_k, which is most of the case. But we have information about w_i and w_j. So by w_i we eliminate $I(f_k; w_i)$ bits of uncertainty from f_k and w_j also eliminates $I(f_k; w_j)$ bits. But w_i and w_j together eliminates $I(f_k; w_i, w_j)$ bits of uncertainty. Thus, if interaction information is positive we improve our guess about f_k On the other hand, if the interaction is negative it can be understood that the information in w_j about f_k is *redundant* given w_i. Of course what has been done for two variables can be extended for multiple variables.

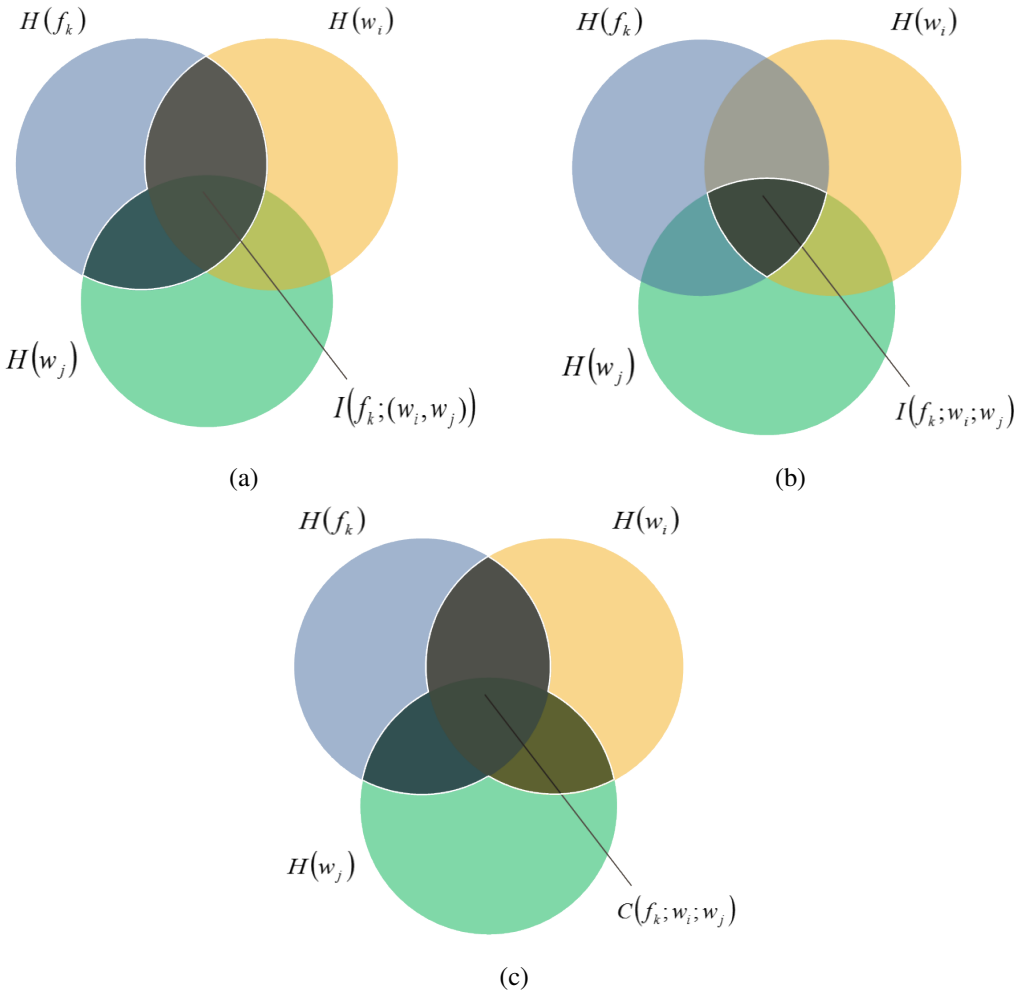


Figure 5.1: Information diagram for three variables. Diagram depicts (a) multivariable, (b) multivariate mutual information and (c) total correlation.

5.1 Indexing Visual Phrases for Content-based Copy Detection

In previous section, mutual information is extended to multiple variables and fundamentals of interaction information is discussed. It has been also shown in Section 4.5 that, it is preferable to index features which have *non-redundant* information about the video frame. Actually the connection of this approach to the interaction information is easily recognizable. Since there are multiple features to be indexed, their interaction with each other could be utilized in order to improve the feature subset selection.

A greedy and exhaustive approach would be to investigate every possible combina-

tion of features/codewords in a given frame. However, this would be infeasible for very large datasets and vocabularies. Instead, methods from previous studies extract *visual phrases* or *collocations* consisting of multiple features. In the literature, visual phrases are extracted from either the entire image or from local neighborhoods. In [77], local feature co-occurrences are extracted from the entire image however an off-line and complex method is proposed for phrase pool generation. On the other hand, local co-occurrence based representation is adopted in [78, 79]. However, aforementioned approaches employ sophisticated phrase extraction methods. Furthermore, these methods address object or scene recognition problem. Contrary to these domains in content-based copy detection do not need a complex parts-based representation thus instead of employing a learning based phrase extraction method, a more simplistic approach is preferred in this work. On the other hand Zhang et al. proposed geometry preserving visual phrases by encoding quantized locations of local features [80]. Features are indexed with quantized spatial information into an inverted file, instead of encoding phrases jointly. And similarity of two images is computed from a 2-D Hough Voting histogram on spatial offset values between matching descriptors. However this approach only captures the translation invariance.

5.1.1 Visual Phrase Extraction

In this work, visual phrase of length k is defined as k local features that are spatially closest to each other in a given frame and radius. And to tolerate transformations observed in CCD datasets, k-nearest-neighbors of each local feature is paired to obtain multiple visual phrases corresponding to a feature. In Figure 5.2, extracted visual phrases are shown for length-2, in such a way that a constellation map is obtained on the frame.

5.1.2 Representation and Indexing

A BoVW approach is adopted for visual phrase representation. Each feature of the visual phrase is represented with a visual word however, on contrary to disjointly indexing, visual phrases are indexed with a single hash value obtained from the visual

(a) (b)

Figure 5.2: Visual phrases of length-2 are extracted from (a) reference and (b) query frames.

words by a simple function (5.4).

$$h_N(w_i, w_j) = \min(w_i, w_j) + \max(w_i, w_j) \times N, \tag{5.4}$$

where N is the codebook size. This way, a much larger index size is attained with smaller visual vocabularies. With this mapping, index size can be computed by $N + N \times (N-1)/2$. For example for a visual vocabulary of size $1,024$, total index size is $524,800$.

Also, a detail fingerprint for each feature comprising the visual phrase is computed with Product Quantization and stored in the inverted file structure as discussed in Section 3.2.2. At the descriptor matching step, much like the previously introduced method, Hamming distance between fingerprints are weighted by a Gaussian function. However, since there are multiple fingerprints in a visual phrase there can be different thresholding approaches. In a small validation set for length-2 visual phrases following thresholding methods (5.5) with different τ values are investigated and empirically $h_{d1}, h_{d2} \leq \tau$ achieved better results.

$$\begin{aligned} h_{d1}, h_{d2} &\leq \tau \\ h_{d1} + h_{d2} &\leq \tau \\ h_{d1}^2 + h_{d2}^2 &\leq \tau^2 \end{aligned} \tag{5.5}$$

Furthermore, intra and inter-frame normalization techniques are also applied to obtain

frame matching score. Finally, temporal alignment is accomplished by utilizing 1-D Hough Transform on time shifts $\delta t = t_r - t_q$ for corresponding query and reference frames.

5.2 Experiments and Discussions

The experiments are performed on TRECVID 2009 and 2010 CCD datasets. Uniform sampling is performed on reference and query videos with the asymmetric sampling rates, 1.5 fps and 5fps respectively. For the coarse representation two different codebooks of sizes 512 and 1024 codewords are compared with each other. Visual phrase length of 2 is selected and furthermore, visual phrases are extracted with 1-NN and 2-NN for comparison. Detail fingerprints are computed by product quantization with a codebook of 256 codewords hence, a total of 128-bit is encoded with each phrase. It should be noted that product quantization codebooks are generated for each coarse codebook independently.

In the first part of the experiments, visual phrase indexing performance for different codebooks and assignments are obtained on both datasets. Table 5.1 and Table 5.3 depicts results obtained on TRECVID 2009 and 2010 CCD datasets respectively. In the remaining parts of the experiments informative features are selected by employing both multivariable and multivariate mutual information. For these experiments the cutoff rate is chosen to be 0.5.

Table 5.1: Performance evaluation of visual phrase-based indexing at TRECVID 2009 CCD dataset. Evaluation profile parameters are set as $C_M = 10$, $C_{FA} = 1$. Total number of features in the reference database are $181,338,757$ for 1-NN and 2-NN cases respectively.

NDCR	**Original**	**VP(1,2)**
	(100K,256)	**(1024,256)**
T1 (Camcording)	NA	NA
T2 (PiP-Type1)	0.062	0.266
T3 (Pat. Ins.)	0.043	0.015
T4 (Strong Re.)	0.140	0.385
T5 (Ch. Gamma)	0.060	0.078
T6 (Dec.Q.3)	0.015	0.007
T7 (Dec.Q.5)	NA	NA
T8 (PP.3)	0.065	0.030
T9 (PP.50	NA	NA
T10 (Rnd.5)	0.255	0.393

Table 5.2: Performance evaluation of compressed reference feature database with multivariate and multivariable information gain of visual phrases at TRECVID 2009 CCD dataset. Entropies are estimated by proposed Grassberger estimator. Evaluation profile parameters are set as $C_M = 10$, $C_{FA} = 1$. Total number of features in the compressed reference databases are $189,299,172$ and $92,183,999$ for 2-NN and 1-NN cases respectively while original database has $205,460,930$ features.

NDCR	**VP(1,2)**	**VP(2,2) – $T_c = 0.5$ Multivariate**		**VP(2,2) – $T_c = 0.5$ Multivariable**	
	(1024,256)	**(1024,256)**	**(512,256)**	**(1024,256)**	**(512,256)**
T1 (Camcording)	NA	NA	NA	NA	NA
T2 (PiP-Type1)	0.266	0.270	0.329	0.357	0.356
T3 (Pat. Ins.)	0.015	0.015	0.000	0.007	0.000
T4 (Strong Re.)	0.385	0.454	0.526	0.392	0.528
T5 (Ch. Gamma)	0.078	0.101	0.058	0.112	0.108
T6 (Dec.Q.3)	0.007	0.018	0.030	0.025	0.007
T7 (Dec.Q.5)	NA	NA	NA	NA	NA
T8 (PP.3)	0.030	0.043	0.060	0.030	0.047
T9 (PP.50	NA	NA	NA	NA	NA
T10 (Rnd.5)	0.393	0.423	0.481	0.478	0.484

Table 5.3: Performance evaluation of visual phrase-based feature indexing at TRECVID 2010 CCD dataset. Evaluation profile parameters are set as $C_M = 10$, $C_{FA} = 1$. Total number of features in the reference database are $181,338,757$ and $323,939,569$ for 1-NN and 2-NN cases respectively.

NDCR	**Baseline**	**VP(2,2)**		**VP(1,2)**	
	(100K,256)	**(512,256)**	**(1024,256)**	**(512,256)**	**(1024,256)**
T1 (Camcording)	0.518	0.718	0.656	0.767	0.770
T2 (PiP-Type1)	0.257	0.526	0.462	0.677	0.592
T3 (Pat. Ins.)	0.126	0.114	0.142	0.107	0.111
T4 (Strong Re.)	0.270	0.530	0.499	0.606	0.615
T5 (Ch. Gamma)	0.140	0.139	0.128	0.136	0.150
T6 (Dec.Q.3)	0.304	0.552	0.489	0.605	0.578
T7 (Dec.Q.5)	NA	NA	NA	NA	NA
T8 (PP.3)	0.143	0.122	0.151	0.117	0.111
T9 (PP.5)	NA	NA	NA	NA	NA
T10 (Rnd.5)	0.399	0.467	0.514	0.575	0.542

Table 5.4: Performance evaluation of compressed reference feature database with *multivariate* information gain of pairwise features at TRECVID 2010 CCD dataset. Entropies are estimated by proposed Grassberger estimator. Evaluation profile parameters are set as $C_M = 10$, $C_{FA} = 1$. Total number of features in the compressed reference databases are $166,533,842$ and $92,183,999$ for 2-NN and 1-NN cases respectively.

NDCR	**VP(2,2)**	**VP(2,2) – $T_c = 0.5$**		**VP(1,2) – $T_c = 0.5$**	
	(1024,256)	**(1024,256)**	**(512,256)**	**(1024,256)**	**(512,256)**
T1 (Camcording)	0.656	0.855	0.826	0.881	0.836
T2 (PiP-Type1)	0.462	0.608	0.681	0.757	0.757
T3 (Pat. Ins.)	0.142	0.129	0.117	0.125	0.106
T4 (Strong Re.)	0.499	0.707	0.747	0.795	0.774
T5 (Ch. Gamma)	0.128	0.132	0.158	0.190	0.211
T6 (Dec.Q.3)	0.489	0.626	0.621	0.656	0.664
T7 (Dec.Q.5)	NA	NA	NA	NA	NA
T8 (PP.3)	0.151	0.132	0.140	0.142	0.122
T9 (PP.50	NA	NA	NA	NA	NA
T10 (Rnd.5)	0.514	0.522	0.600	0.573	0.608

Table 5.5: Performance evaluation of compressed reference feature database with *multivariable* information gain of pairwise features at TRECVID 2010 CCD dataset. Entropies are estimated by proposed Grassberger estimator. Evaluation profile parameters are set as $C_M = 10$, $C_{FA} = 1$. Total number of features in the compressed reference databases are $182,821,730$ and $100,501,416$ for 2-NN and 1-NN cases respectively.

NDCR	**VP(2,2)**	**VP(2,2) – $T_c = 0.5$**		**VP(1,2) – $T_c = 0.5$**	
	(1024,256)	**(1024,256)**	**(512,256)**	**(1024,256)**	**(512,256)**
T1 (Camcording)	0.656	0.884	0.938	0.962	0.936
T2 (PiP-Type1)	0.462	0.641	0.719	0.751	0.748
T3 (Pat. Ins.)	0.142	0.099	0.113	0.088	0.092
T4 (Strong Re.)	0.499	0.648	0.727	0.807	0.812
T5 (Ch. Gamma)	0.128	0.160	0.129	0.209	0.197
T6 (Dec.Q.3)	0.489	0.593	0.671	0.719	0.708
T7 (Dec.Q.5)	NA	NA	NA	NA	NA
T8 (PP.3)	0.151	0.160	0.164	0.144	0.141
T9 (PP.50	NA	NA	NA	NA	NA
T10 (Rnd.5)	0.514	0.528	0.589	0.594	0.643

CHAPTER 6

CONCLUSION

In the light of the previous discussions, benefits and challenges of a robust video duplicate detection system is apparent. Although there might be different video duplicate notions, in this study *copy video* definition is adopted. Copy videos are defined as videos from the same source but by some kind of tolerable transformations differ from each other. Transformations may contain addition, deletion, modification (of aspect, color, contrast, encoding) camcording and etc. In this respect, throughout the studies building blocks of a content-based copy detection system are investigated comprehensively. In the next section, summary and concluding remarks of this thesis is provided. Finally, possible future research paths are discussed in Section 6.3.

6.1 Summary

After presenting the problem statement and scope of the thesis in Chapter 1, related work from literature is analyzed for content-based copy detection problem in Chapter 2. Moreover, publicly available datasets and corresponding performance evaluation metrics are given in the same chapter.

In Chapter 3, two different approaches are investigated for the solution of content based copy detection. A novel global spatio-temporal approach is proposed in Section 3.1 whereas in Section 3.1.1 experimental results of this method is provided. Aforementioned method utilizes temporal volume of frames in a given grid structure for fingerprint extraction in which, multiple feature extractors that spans three fundamental visual information sources namely color, texture and motion are employed.

Furthermore, in Section 3.2 a local interest point-based method is presented. Experimental analysis is conducted on two large datasets, TRECVID 2009 and TRECVID 2010 CCD datasets. Results are also compared with related work from the literature in 3.2.6.

In Chapter 4, an information theoretic feature selection and indexing method is introduced by first discussing fundamentals of feature selection. Afterwards a mutual information based algorithm is proposed. And for the estimation of joint and conditional probability distributions frequentist histogram based approach is adopted. However, drawbacks of Naive estimator is shown and in Section 4.4 a better estimator is given in order to improve the entropy estimate. Advantages of this approach is first shown on synthetically generated data in terms of MSE and bias on estimates in 4.4.1. Afterwards, content-based copy detection performance of the informative feature selection-based indexing is reported in TRECVID 2009 and 2010 CCD datasets. Also, in Section 4.6, a method to exploit temporal distributions of local features of successive frames is introduced and experimental results are reported for the same datasets.

In Chapter 5 mutual information definition is extended to multiple variables. And *interaction information* concept is introduced for local feature selection. In order to utilize this perspective a *visual phrase* indexing approach is proposed in Section 5.1. Moreover, experimental evaluation of visual phrase indexing and corresponding feature selection methods is further investigated in Section 5.2.

6.2 Concluding Remarks

The proposed global spatio-temporal content-based copy detection approach in Section 3.1 showed promising results experimentally when compared with other approaches from the literature. However, local interest point-based pipeline as presented in Section 3.2 performed better in terms of detection and false alarm rate on larger datasets. Furthermore, time localization performance of interest point-based approach is much higher when compared to proposed spatio-temporal method. When compared with other works in the literature, presented interest-point based method

showed superior results. For example in TRECVID 2009 CCD dataset for all transformations/attacks on the average presented method achieved 0.091 NDCR score which is nearly a four-fold improvement compared to recently published results.

As discussed in Chapter 4, index size effects both detection accuracy and computational complexity. Furthermore, because of the increase in the number of reference videos or the temporal/spatial sampling rate, the index structure might grow considerably hence hindering the performance of the interest-point based methods. In this thesis to remedy the database growth informative feature selection based approach is proposed to overcome the index size. It has been shown that in content-based copy detection datasets, by only indexing a fraction of features same and even better detection performance is achieved by mutual information based feature selection. Especially for *Pattern Insertion*, *Change of Gamma*, *Decrease of Quality* and *Post-Production* transformations, better performance is consistently achieved for all profiles. Even 0.00 NDCR score is achieved for some transformations which was not possible previously with full indexing. Moreover, proposed entropy estimator based informative feature selection method, as expected, showed superior detection performance compared to Naive entropy estimation. Finally, since a fraction of features are indexed a shorter inverted list is queried. Thus memory usage efficiency is improved and coupled with having less distractor features, matching is achieved at least $2.0\times$ much faster when half of the features are selected.

Also, from randomized informative feature indexing it has been observed that higher detection performance is achieved with successive frame insertion compared to random insertion. Thus, a method to exploit temporal distributions of local features of successive frames is introduced. For most of the transformations temporal indexing does not improve performance considerably compared to full indexing. However, when compared with informative feature indexing from individual frames, merits of temporal indexing are better seen. For most of the transformations 31% to 83% improvement on NDCR is observed. Furthermore, as the temporal window length increases the detection performance improves since *discriminative* consistency among local features are better captured from larger temporal volume. It should be also noted that expected detection results are not observed in TRECVID 2010 dataset. This is mostly because of the reference video characteristics such as fast changing scenes

and single shot videos.

Since there are multiple features to be indexed in a frame, in order to improve the feature subset selection it is proposed to utilize feature interactions in a given frame. Although proposed visual phrase extraction method is not well-suited for the content-based copy detection problem, it has been shown by the experimental analysis that employing multivariate modeling improves feature selection performance when compared with multivariable approach. That is to say, dependencies between features are better captured by multivariate mutual information.

6.3 Future Work

In this thesis, it has been shown that local interest point based approaches are more efficient in content-based copy detection problems. Although in the literature there are methods employing geometric consistency filtering, in the proposed interest point based method no geometric constraints or spatial information is utilized. Even so, when compared with others, proposed method showed considerably better results. However, it would be interesting to utilize spatial information in the matching.

During the investigation of the information theoretic indexing strategies, a frequentist approach is adopted for the estimation of conditional and joint probability distributions. However, as a future work performance of the feature selection can be analyzed for Kernel density estimation (KDE).

Throughout the studies informativeness of a feature is measured by the distribution of codewords in the reference index and corresponding frame. As discussed in [72], similarities and repeating patterns are revealing to find the informative/discriminative features in a frame. Although in the proposed method feature similarities have been *somewhat* captured, since feature selection is carried out after codeword transformation. In addition, context dependent feature similarities can be exploited in order to improve feature selection strategies. Furthermore, prior information on the known attacks and transformations can be utilized to improve the robustness of selection to the severe transformations.

Although the scope of this thesis is limited to the detection of *copy* videos, methods introduced in this study can be further applied for *near-duplicate* detection. Especially, investigating local feature informativeness in a multiple viewpoint setup would be an interesting research path. Likewise, information theoretic indexing can be applied for different visual analysis problems such as image retrieval, object recognition and etc.

REFERENCES

[1] Lee Gomes. Will All of Us Get Our 15 Minutes On a YouTube Video?, 2006.

[2] `http://www.youtube.com/yt/press/statistics.html`. [Online; accessed 03-February-2014].

[3] S. Poehlein, V. Saxena, G.T. Willis, Jeff Fedders, and Martin Guttmann. Viewpoint paper: Moving media in the cloud. Technical report, Intel, 2010.

[4] Zi Huang, Heng Tao Shen, Jie Shao, Bin Cui, and Xiaofang Zhou. Practical Online Near-Duplicate Subsequence Detection for Continuous Video Streams. *IEEE Transactions on Multimedia*, 12(5):386–398, August 2010.

[5] Arslan Basharat, Yun Zhai, and Mubarak Shah. Content based video matching using spatiotemporal volumes. *Computer Vision and Image Understanding*, 110(3):360–377, June 2008.

[6] Mauro Cherubini, Rodrigo de Oliveira, and Nuria Oliver. Understanding near-duplicate videos. In *Proceedings of the seventeen ACM international conference on Multimedia - MM '09*, page 35, New York, New York, USA, October 2009. ACM Press.

[7] Alexis Joly, C. Frelicot, and O. Buisson. Robust Content-Based Video Copy Identification in a Large Reference Database. *Image and Video Retrieval*, 2728:511–516, 2003.

[8] J. Law-To, V. Gouet-Brunet, O. Buisson, and N. Boujemaa. Local Behaviours Labelling for Content Based Video Copy Detection. In *18th International Conference on Pattern Recognition (ICPR'06)*, pages 232–235. Ieee, 2006.

[9] A. Jaimes and A.C. Lou. Detection of non-identical duplicate consumer photographs. In *Fourth International Conference on Information, Communications and Signal Processing and the Fourth Pacific Rim Conference on Multimedia*, pages 16–20. IEEE, 2003.

[10] E. Rossi, S. Benini, R. Leonardi, B. Mansencal, and J. Benois-Pineau. Clustering of scene repeats for essential rushes preview. In *2009 10th Workshop on Image Analysis for Multimedia Interactive Services*, pages 234–237. IEEE, May 2009.

[11] Shin'ichi Satoh, Masao Takimoto, and Jun Adachi. Scene duplicate detection from videos based on trajectories of feature points. In *Proceedings of the international workshop on Workshop on multimedia information retrieval - MIR '07*, page 237, New York, New York, USA, 2007. ACM Press.

[12] M. Takimoto and M.S. Shin'ichi Satoh. Identification and detection of the same scene based on flash light patterns. In *2006 IEEE International Conference on Multimedia and Expo*, pages 9–12. IEEE, 2006.

[13] Jerome Revaud, Matthijs Douze, Cordelia Schmid, and Herve Jegou. Event Retrieval in Large Video Collections with Circulant Temporal Encoding. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2459–2466. Ieee, June 2013.

[14] Chih-Yi Chiu, Jenq-Haur Wang, and Hung-Chi Chang. Efficient Histogram-Based Indexing for Video Copy Detection. In *Ninth IEEE International Symposium on Multimedia Workshops (ISMW 2007)*, pages 265–270. IEEE, December 2007.

[15] Chih-Yi Chiu, Cheng-Chih Yang, and Chu-Song Chen. Efficient and Effective Video Copy Detection Based on Spatiotemporal Analysis. In *Ninth IEEE International Symposium on Multimedia (ISM 2007)*, pages 202–209. IEEE, December 2007.

[16] D.N. Bhat and S.K. Nayar. Ordinal measures for image correspondence. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(4):415–423, April 1998.

[17] J. Zobel. Detection of video sequences using compact signatures. *ACM Transactions on Information Systems*, 24(1):1–50, January 2006.

[18] Arun Hampapur, Kiho Hyun, and Ruud M. Bolle. Comparison of sequence matching techniques for video copy detection. In *Proceedings of SPIE*, pages 194–201. SPIE, 2001.

[19] Ahmet Saracoglu, Ersin Esen, Tugrul K. Ates, Banu Oskay Acar, Unal Zubari, Ezgi C. Ozan, Egemen Ozalp, a. Aydin Alatan, and Tolga Ciloglu. Content Based Copy Detection with Coarse Audio-Visual Fingerprints. *2009 Seventh International Workshop on Content-Based Multimedia Indexing*, pages 213–218, June 2009.

[20] David G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 60(2):91–110, November 2004.

[21] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. SURF: Speeded Up Robust Features. In *Computer Vision–ECCV 2006*, 2006.

[22] Geert Willems, Tinne Tuytelaars, and Luc Van Gool. Spatio-temporal features for robust content-based video copy detection. In *Proceeding of the 1st ACM international conference on Multimedia information retrieval - MIR '08*, page 283, New York, New York, USA, 2008. ACM Press.

[23] Matthijs Douze, Hervé Jegou, and Cordelia Schmid. An Image-Based Approach to Video Copy Detection With Spatio-Temporal Post-Filtering. *IEEE Transactions on Multimedia*, 12(4):257–266, June 2010.

[24] Herve Jegou, Matthijs Douze, and Cordelia Schmid. Hamming embedding and weak geometric consistency for large scale image search. In *Computer Vision–ECCV 2008*, pages 304–317. Springer, 2008.

[25] Matthijs Douze, H Jégou, and C Schmid. Compact video description for copy detection with precise temporal alignment. In *Computer Vision–ECCV 2010*, 2010.

[26] Alexis Joly, Olivier Buisson, and Carl Frelicot. Content-Based Copy Retrieval Using Distortion-Based Probabilistic Similarity Search. *IEEE Transactions on Multimedia*, 9(2):293–306, February 2007.

[27] Xiangmin Zhou, Xiaofang Zhou, Lei Chen, Athman Bouguettaya, Nong Xiao, and John a. Taylor. An Efficient Near-Duplicate Video Shot Detection Method Using Shot-Based Interest Points. *IEEE Transactions on Multimedia*, 11(5):879–891, August 2009.

[28] Shiyang Lu, Zhiyong Wang, Meng Wang, Max Ott, and Dagan Feng. Adaptive reference frame selection for near-duplicate video shot detection. In *17th IEEE International Conference on Image Processing (ICIP)*, pages 2341–2344, 2010.

[29] David Chen, Ngai-Man Cheung, Sam Tsai, Vijay Chandrasekhar, Gabriel Takacs, Ramakrishna Vedantham, Radek Grzeszczuk, and Bernd Girod. Dynamic selection of a feature-rich query frame for mobile video retrieval. In *17th IEEE International Conference on Image Processing*, pages 1017–1020, 2010.

[30] Carlo Tomasi and T Kanade. Detection and Tracking of Point Features Technical Report CMU-CS-91-132. Technical Report 7597, 1991.

[31] TK Ates, E Esen, A Saracoglu, M Soysal, Y Turgut, O Oktay, and AA Alatan. Content based video copy detection with local descriptors. In *Signal Processing and Communications Applications Conference (SIU), 2010 IEEE 18th*, pages 49–52. IEEE, 2010.

[32] Yan Ke, Rahul Sukthankar, and Larry Huston. An efficient parts-based near-duplicate and sub-image retrieval system. In *12th annual ACM international conference on Multimedia - MULTIMEDIA '04*, page 869, New York, New York, USA, October 2004. ACM Press.

[33] WL Zhao and CW Ngo. Flip-invariant SIFT for copy and object detection. *Image Processing, IEEE Transactions on*, 22(3):980–991, 2013.

[34] Zhu Liu, Tao Liu, David C. Gibbon, and Behzad Shahraray. Effective and scalable video copy detection. In *Proceedings of the international conference on Multimedia information retrieval - MIR '10*, page 119, New York, New York, USA, 2010. ACM Press.

[35] Onur Küçüktunç, Muhammet Baştan, Uğur Güdükbay, and Özgür Ulusoy. Video copy detection using multiple visual cues and MPEG-7 descriptors. *Journal of Visual Communication and Image Representation*, 21(8):838–849, November 2010.

[36] J. Law-To, L. Chen, Alexis Joly, Ivan Laptev, Olivier Buisson, V. Gouet-Brunet, N. Boujemaa, and Fred Stentiford. Video copy detection: a comparative study. In *Proceedings of the 6th ACM international conference on Image and video retrieval*, pages 371–378. ACM, 2007.

[37] Semin Kim, Jae Young Choi, Seungwan Han, and Yong Man Ro. Adaptive weighted fusion with new spatial and temporal fingerprints for improved video copy detection. *Signal Processing: Image Communication*, 29(7):788–806, August 2014.

[38] Mani Malek Esmaeili, Mehrdad Fatourechi, and Rabab Kreidieh Ward. A Robust and Fast Video Copy Detection System Using Content-Based Fingerprinting. *IEEE Transactions on Information Forensics and Security*, 6(1):213–226, March 2011.

[39] Li Chen and F. W. M. Stentiford. Video sequence matching based on temporal ordinal measurement. *Pattern Recogn. Lett.*, 29(13):1824–1831, October 2008.

[40] Changick Kim and B. Vasudev. Spatiotemporal sequence matching for efficient video copy detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 15(1):127–132, January 2005.

[41] Ozgun Cirakman, Bilge Gunsel, Neslihan Serap Sengor, and Sezer Kutluk. Content-based copy detection by a subspace learning based video fingerprinting scheme. *Multimedia Tools and Applications*, 71(3):1381–1409, November 2012.

[42] Roger Weber, H.J. Schek, and Stephen Blott. A quantitative analysis and performance study for similarity-search methods in high-dimensional spaces. In *Proceedings of the International Conference on Very Large Data Bases*, pages 194–205. INSTITUTE OF ELECTRICAL & ELECTRONICS ENGINEERS, 1998.

[43] Piotr Indyk and R. Motwani. Approximate nearest neighbors: towards removing the curse of dimensionality. In *Proceedings of the thirtieth annual ACM symposium on Theory of computing*, pages 604–613. ACM, 1998.

[44] Sébastien Poullot, Olivier Buisson, and Michel Crucianu. Z-grid-based probabilistic retrieval for scaling up content-based copy detection. In *Proceedings of the 6th ACM international conference on Image and video retrieval - CIVR '07*, pages 348–355, New York, New York, USA, 2007. ACM Press.

[45] D. Nister and H. Stewenius. Scalable Recognition with a Vocabulary Tree. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2 (CVPR'06)*, volume 2, pages 2161–2168. IEEE, 2006.

[46] Hervé Jégou, Matthijs Douze, and Cordelia Schmid. Product quantization for nearest neighbor search. *IEEE transactions on pattern analysis and machine intelligence*, 33(1):117–28, January 2011.

[47] Juan Manuel Barrios and Benjamin Bustos. P-VCD: A pivot-based approach for Content-Based Video Copy Detection. In *2011 IEEE International Conference on Multimedia and Expo*, pages 1–6. IEEE, July 2011.

[48] Chih-Yi Chiu, Chu-Song Chen, and Lee-Feng Chien. A Framework for Handling Spatiotemporal Variations in Video Copy Detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 18(3):412–417, 2008.

[49] Yanwei Fu, Yanwen Guo, Yanshu Zhu, Feng Liu, Chuanming Song, and Zhi-Hua Zhou. Multi-View Video Summarization. *IEEE Transactions on Multimedia*, 12(7):717–729, November 2010.

[50] Werner Bailer. Evaluating detection of near duplicate video segments. In *Proceedings of the ACM International Conference on Image and Video Retrieval - CIVR '10*, page 197, New York, New York, USA, 2010. ACM Press.

[51] J Law-To, A Joly, and N Boujemaa. Muscle-VCD-2007: a live benchmark for video copy detection, 2007.

[52] Building video queries for TRECVID2008 copy detection task, 2008.

[53] B. S. Manjunath, Phillipe Salembier, and Thomas Sikora. *Introduction to MPEG-7: Multimedia Content Description Interface*. John Wiley & Sons, Inc., New York, NY, USA, 2002.

[54] Josef Sivic and Andrew Zisserman. Video Google: a text retrieval approach to object matching in videos. In *Proceedings Ninth IEEE International Conference on Computer Vision*, number Iccv, pages 1470–1477 vol.2. Ieee, 2003.

[55] Slava M. Katz. Distribution of content words and phrases in text and language modelling. *Nat. Lang. Eng.*, 2(1):15–59, March 1996.

[56] KW Church and WA Gale. Poisson mixtures. *Natural Language Engineering*, pages 1–24, 1995.

[57] Rasmus E. Madsen, David Kauchak, and Charles Elkan. Modeling word burstiness using the Dirichlet distribution. In *Proceedings of the 22nd international conference on Machine learning - ICML '05*, pages 545–552, New York, New York, USA, 2005. ACM Press.

[58] Qi He, Kuiyu Chang, and Ee-Peng Lim. Using burstiness to improve clustering of topics in news streams. In *Proceedings of the 2007 Seventh IEEE International Conference on Data Mining*, ICDM '07, pages 493–498, Washington, DC, USA, 2007. IEEE Computer Society.

[59] F Schaffalitzky and A Zisserman. Automated location matching in movies. *Computer Vision and Image Understanding*, 92(2-3):236–264, November 2003.

[60] H. Jegou, M. Douze, and Cordelia Schmid. On the burstiness of visual elements. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1169–1176. IEEE, June 2009.

[61] Mei-Chen Yeh and Kwang-Ting Cheng. Video copy detection by fast sequence matching. In *Proceeding of the ACM International Conference on Image and Video Retrieval - CIVR '09*, page 1, New York, New York, USA, 2009. ACM Press.

[62] David Arthur and Sergei Vassilvitskii. k-means ++ : The Advantages of Careful Seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, volume 8, pages 1–11, 2007.

[63] Marius Muja and D.G. Lowe. Fast Approximate Nearest Neighbors with Automatic Algorithm Configuration. In *In VISAPP International Conference on Computer Vision Theory and Applications*, volume 340, pages 331–340, 2009.

[64] Savas Ozkan. *Content-Based Video Copy Detection*. M.sc., Middle East Technical University, 2014.

[65] Yvan Saeys, Iñaki Inza, and Pedro Larrañaga. A review of feature selection techniques in bioinformatics. *Bioinformatics (Oxford, England)*, 23(19):2507–17, October 2007.

[66] Huan Liu and Lei Yu. Toward integrating feature selection algorithms for classification and clustering. *IEEE Transactions on Knowledge and Data Engineering*, 17(4):491–502, April 2005.

[67] Grant Schindler, Matthew Brown, and Richard Szeliski. City-Scale Location Recognition. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–7. Ieee, June 2007.

[68] Fayin Li and J Kosecka. Probabilistic location recognition using reduced feature set. In *Proceedings 2006 IEEE International Conference on Robotics and Automation, 2006. ICRA 2006.*, pages 3405–3410. IEEE, 2006.

[69] Panu Turcot and David G. Lowe. Better matching with fewer features: The selection of useful features in large database recognition problems. In *2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops*, pages 2109–2116. IEEE, September 2009.

[70] Nikhil Naikal, Allen Y Yang, and S. Shankar Sastry. Informative feature selection for object recognition via Sparse PCA. In *2011 International Conference on Computer Vision*, pages 818–825. IEEE, November 2011.

[71] Zixuan Wang, Qi Zhao, David Chu, Feng Zhao, and Leonidas J. Guibas. Select informative features for recognition. In *2011 18th IEEE International Conference on Image Processing*, pages 2477–2480. IEEE, September 2011.

[72] Giorgos Tolias, Yannis Kalantidis, and Yannis Avrithis. SymCity: feature selection by symmetry for large scale image retrieval. In *Proceedings of the 20th ACM international conference on Multimedia - MM '12*, page 189, New York, New York, USA, 2012. ACM Press.

[73] Thomas Schürmann. Bias analysis in entropy estimation. *Journal of Physics A: Mathematical and General*, 37(27):L295–L301, July 2004.

[74] B Harris. The statistical estimation of entropy in the non-parametric case. In *Colloquium of the Mathematical Society of János Bolyai*, pages 323–355, 1977.

[75] G. Miller. Note on the bias of information estimates. *In Information Theory in Psychology: Problems and Methods*, pages 95–100, 1955.

[76] A. A. Evans, Ronald J. and Boersma, J. and Blachman, N. M. and Jagers. The Entropy of a Poisson Distribution. *SIAM Review*, 30(2):314–317, 1988.

[77] L. Torresani, M. Szummer, and A. Fitzgibbon. Learning query-dependent prefilters for scalable image retrieval. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2615–2622. IEEE, June 2009.

[78] Junsong Yuan, Ying Wu, and Ming Yang. Discovery of Collocation Patterns: from Visual Words to Visual Phrases. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, June 2007.

[79] C. Lawrence Zitnick, Jie Sun, Richard Szeliski, and Simon Winder. Object instance recognition using triplets of feature symbols. Technical Report MSR-TR-2007-53, Microsoft Research, 2007.

[80] Yimeng Zhang, Zhaoyin Jia, and Tsuhan Chen. Image retrieval with geometry-preserving visual phrases. In *CVPR 2011*, pages 809–816. IEEE, June 2011.

CURRICULUM VITAE

AHMET SARACOĞLU

PERSONAL INFORMATION

Nationality	Turkish (TC)
Date of Birth	1982
Place of Birth	Eskişehir
Marital Status	Married
E-Mail	asaracoglu@gmail.com
Mobile	+90 533 3215979

EDUCATION

Degree	**Institution**	**Year of Graduation**
M.S.	METU - Department of Electrical and Electronics Engineering	2007
B.S.	METU - Department of Electrical and Electronics Engineering	2004
High School	TED Ankara College	2000

PROFESSIONAL EXPERIENCE

Year	**Place**	**Enrollment**
2012-Present	Kuartis Technology and Consulting	Managing Partner, Co-Founder
2012-Present	KuartisMED Medical Research	Managing Partner, Co-Founder
2010-2012	TUBITAK Space Technologies Research Institute	Technical Leader
2006-2010	TUBITAK Space Technologies Research Institute	Senior Researcher
2004-2006	METU Department of Electrical and Electronics Engineering	Research Assistant

PUBLICATIONS

D. Anuk İnce, A. Ecevit, B. Oskay Acar, A. Saracoğlu, A. Kurt, M.A. Tekindal, A. Tarcan, "Noninvasive evaluation of swallowing sound is an effective way of diagnosing feeding maturation in newborn infants," Acta Paediatrica 2014.

M. Soysal, K. B. Loğoğlu, M. Tekin, E. Esen, A. Saracoğlu, et. al., “Multimodal concept detection in broadcast media: Kav Tan,” Springer Multimedia Tools and Ap-

plications, 2013

M. Tekin, A. Saracoğlu, E. Esen, et. al., "Multimodal Concept Detection on Multimedia Data – RTÜK SKAAS KavTan System," IEEE SİU 2012, Muğla Türkiye.

M. Ali Arabaci, M. Soysal, A. Saracoğlu, E. Esen, "Flag Detection Using Spatial-Color Joint Probability Functions," IEEE SİU 2012, Muğla Türkiye.

A. Saracoğlu, E. Esen, et al., "TUBITAK UZAY at TRECVID 2010: Content-Based Copy Detection and Semantic Indexing," in the Proceedings of TRECVID 2010, Maryland USA.

K. B. Loğoğlu., A. Saracoğlu, Ersin Esen, A. Aydın Alatan, "Gender Classification via Gradientfaces," ISCIS 2010, London UK.

H. Sevimli, E. Esen, A. Saracoğlu, et al., "Adult Image Content Classification Using Global Features and Skin Region Detection," ISCIS 2010, London UK.

A. Saracoğlu, M. Tekin, E. Esen, et al., "Generalized Visual Concept Detection," SİU 2010, Diyarbakır Türkiye.

T. K. Ateş, E. Esen, A. Saracoğlu, et al., "Content Based Video Copy Detection with Local Descriptors," IEEE SİU 2010, Diyarbakır Türkiye.

A. Saracoğlu, E. Esen, M. Soysal, et al., "TUBITAK UZAY at TRECVID 2009: High-Level Feature Extraction and Content-Based Copy Detection," in the Proceedings of TRECVID 2009, Maryland USA.

A. Saracoğlu, E. Esen, T. K. Ateş, B. Oskay Acar, Ü. Zubari, E. C. Ozan, E. Özalp, A. A. Alatan and Tolga Çiloğlu, "Content Based Copy Detection with Coarse Audio-Visual Fingerprints" in CBMI 2009, Crete Greece.

E. Esen, A. Saracoğlu, T. K. Ateş, B. Oskay, Ü. Zubari and A. A. Alatan, "Content Based Video Copy Detection with Coarse Features," IEEE SİU 2009, Antalya Türkiye.

B. Oskay Acar, Ü. Zubari, E. C. Ozan, A. Saracoğlu, E. Esen and T. Çiloğlu, "Voting System Based Robust and Efficient Audio Copy Detection," IEEE SİU 2009, Antalya Türkiye.

A. Saracoglu, E. Esen, A. A. Alatan, et al., "COST292 experimental framework for TRECVID2008" TRECVID 2008 Workshop, 2008.

E. Esen, M. Soysal, T. K. Ateş, A. Saracoglu and A. A. Alatan, "A Fast Method For Animated TV Logo Detection," CBMI 2008, London, England.

T. K. Ateş, E. Esen, A. Saracoglu and A. A. Alatan, "Boundary Matching Based Translucent TV Logo Detection," IEEE SIU 2008, Antalya, Türkiye.

A. Saracoglu and A. A. Alatan "MRF-based VideoText Detection" International Conference on Systems, Signals and Image Processing, 21-23 September 2006, Budapest-Hungary.

A. Saracoglu and A. A. Alatan “Automatic Video Text Localization and Recognition” Signal Processing and Communications Applications, 2006 IEEE 14th, 17-19 April 2006.

A. Saracoglu, A. A. Alatan, et.al., “COST292 experiments for TRECVID 2006” TRECVID 2006 Workshop, 2006.