141
487
THS

THESIS
1
2009

LIBRARY
Michigan State
University

This is to certify that the
thesis entitled

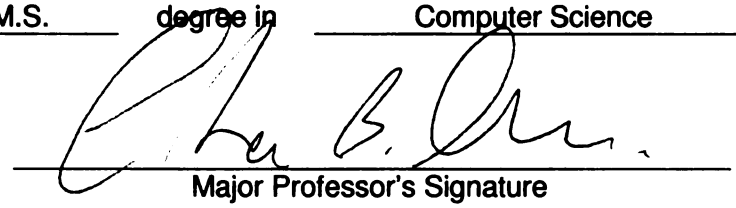
VIRTUAL ENVIRONMENT FOR VIDEO BROWSING: DESIGN AND COMPARATIVE STUDY OF EFFECTIVENESS

presented by

Zubin John Abraham

has been accepted towards fulfillment
of the requirements for the

M.S. degree in Computer Science



Major Professor's Signature

12-05-08

Date

MSU is an Affirmative Action/Equal Opportunity Employer

PLACE IN RETURN BOX to remove this checkout from your record.
TO AVOID FINES return on or before date due.
MAY BE RECALLED with earlier due date if requested.

DATE DUE	DATE DUE	DATE DUE

5/08 K /Proj/Acc&Pres/CIRC/DateDue.indd

VIRTUAL ENVIRONMENT FOR VIDEO BROWSING: DESIGN AND COMPARATIVE STUDY OF EFFECTIVENESS

By

Zubin John Abraham

A THESIS

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

MASTER OF SCIENCE

Computer Science

2008

ABSTRACT

VIRTUAL ENVIRONMENT FOR VIDEO BROWSING: DESIGN AND COMPARATIVE STUDY OF EFFECTIVENESS
By

Zubin John Abraham

The thesis presents the Virtual Video Gallery, a virtual video browsing environment that facilitates browsing and retrieval of video clips from a video repository by using content-based video retrieval techniques. The browsing structure avoids the need to submit a query in the form of a video or an image, while still providing retrieval capabilities that are sensitive to the underlying media content. The browsing environment provides alternative trails of related content that users can browse for video browsing by visually portraying inter-video relationships derived from a composite of content and metadata similarity. The Virtual Gallery is presented as a semi-realistic representation of a real art gallery, as opposed to an artificial computer simulated environment, and has been designed with the intention of exploiting human spatial memory and the ability of humans to recognize visual patterns. The concept is developed in the context of previous research done on content-based image retrieval and browsing environments. This thesis presents a prototype Virtual Video Gallery and the results of user studies that examined the effectiveness of the system. It is shown that there was 12% improvement in successfully finding a video of choice and 26% improvement in finding a video similar to a given video. There was a 60% reduction in the number of queries entered as well as 70% increase in number of video clips viewed per user. This translates to approximately a 300% improvement in the success of retrieving a video of choice, per query entered by the user.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Images in this thesis are presented in color.

Chapter 1

Introduction

The growth of high-speed Internet connections and video-sharing websites such as YouTube and Google Video has spurred a massive increase in the number of multimedia files easily available over the World Wide Web. For this wealth of information to be useful to the general public, it is important to develop better ways to find interesting and relevant content. This is particularly the case for sites that accumulate video content, which remains difficult to search and browse. In most of the current state of the industry, the content producer is prompted to append relevant metadata in the form of textual tags and written commentary to each video to facilitate categorization and organization of the various video clips over the website. Consequent to the growth in the number of video clips and the utilization of these user supplied ad-hoc tags as a primary means for organization and searching, online web-based video libraries increasingly grapple with the problem of organizing and showcasing their large compilations of video clips in a coherent manner. The labeling

with three to five simple tags does not adequately describe video content that can number in the millions.

This thesis presents one approach to the problem of locating content of interest in a large video collection: the Virtual Video Gallery. The Virtual Video Gallery is a *virtual browsing environment* that presents video clips in a setting analogous to that of a physical museum allowing users to rapidly browse the collection by moving among the exhibits. The layout of content within the museum is determined in response to a user-provided query merged with a self-organization based on image content. Visually related content is placed in close proximity to each other, with the physical distance separating content an indicator of similarity. A fundamental element of this approach is the concept of browsing. A search moves a user to the approximate neighborhood of the desired result. From this point forward, the user explores content in a physically related area with a natural and appealing appearance. This physical representation/layout is a metaphor for actual content similarity measures. In the prototype built this research, that similarity is based on color histogram comparisons of representative key frames. The goal of this thesis is to examine the value of a video browsing environment with a museum metaphor as a means to promote exploration within content and the process of locating relevant retrieval result. Consequently, a relatively simple, yet commonly accepted, means is used to compare the video content. Clearly, more advanced similarity measures exist and would be the subject of future work.

Effective management of media data on the web and in computers, particularly temporal content such as audio and video, has been extensively studied. Many approaches exist, but there is no general solution to the problem of locating desired content in large collections. Historically textual document retrieval has attracted the most attention [1-3]. Document retrieval and browsing approaches often employ 2D icons [4, 5] that represent documents so as to take advantage of human spatial recognition [6, 7]. Due to the inherent limitations on the amount of information that can be conveyed using 2D spatial layouts, the use of 3D spatial layouts for exploiting human spatial cognition abilities as also been explored [2, 6-9]. As shown in Figure 1.1, Data Mountain, a document management technique, uses a 3D desktop virtual environments with 2D interaction techniques to improve management of documents in an information space by exploiting spatial memory [4, 10].

This thesis extends concepts incorporated in Data Mountain and other similar digital file and textual content management techniques as a means to improve video selection and retrieval using a 3D Virtual Video Gallery [2, 8] wherein illustrative snapshots of video clips were strategically placed on props and virtual real-estate space within the spatial confines of a gallery appearing as a visual metaphor. Video clips were clustered and arranged within different "rooms" in the gallery based on a combination of similarity measures including metadata similarity and similarity of the underlying visual content. The user navigates in the 3D space using interaction techniques similar to computer games and other traditional 3D applications. The

Virtual Video Gallery attempted to approximate the human trait of arranging documents spatially such that the most recent/relevant video clips were placed closest to the user, while also ensuring that similar video clips remain within close proximity of each other [4].

Figure 1.1 - Layout of the Data Mountain that uses a 3D desktop virtual environment [4].

Content-based image retrieval (CBIR) [11] techniques including color histograms[11-15] and its various variants [13, 16] were used for evaluating similarity of images and video. A variety of content similarity measures exist for video. For this work, color histograms were chosen as a common and accepted basic method. The color histogram of an image remains invariant to translation and rotation about the

viewing axis, as well as other affine transformations. It is well suited to recognize an object of unknown position and rotation within a scene, hence, a good candidate for a basic content similarity approach. Indeed, the focus of this work is less on the best way to compare video clips to each other but more on the use and utility of browsing as a means of rapidly exploring libraries of video content. The most important question is the effectiveness of a virtual browsing environment on locating content in a large video database; specifically the advantages this method has as a new user interface.

The remainder of this thesis examines the design, implementation, and evaluation of the Virtual Gallery. Chapter 2 describes related work that has been published by various authors. Chapter 3 on research methodology describes both Content-based video retrieval and the browsing environment. Chapter 4 describes the concept of the Virtual Gallery followed by a description of a scenario of a hypothetical envisaged system based on the concept of the Virtual Gallery. Chapter 5 details the design of the Virtual Gallery and includes the experimental setup, which describes the prototype of the system that was built along with the results and conclusions from the experimental setup. This is followed by future work, appendix and the bibliography(Chapters 6 through 8)

Chapter 2

Related work

There has been significant progress in the progression from 2D Virtual environments [4] to 3D environments [6, 8]. Ann Smith et al. describe Vista, a 3D virtual environment presented as a web service for exhibition of visual data [17]. The environment dynamically accommodated deletion or addition of information. The environment used "virtual archives" and "space warp trams." The *space warp trams* allowed the user to navigate quickly. Vista presents four domains: a gallery, studio, cinema, and library. In each level of the gallery, at most stored 40 images and 8 sculptures are stored and can be viewed. Though Vista arranged the different types of multimedia in a structured manner such that there is a clear classification between the types of multimedia and a cognitive organization of the data, no emphasis had been given to intra-organization of video content. Rather, the approach assumes self-organization of content.

There has been significant work done in 3D virtual environments for visualization of multimedia data. As seen in Figure 2.1, MediaMetro is a 3D multimedia visualization aid using a city metaphor [2]. MediaMetro utilizes key frames that are generated using a content analysis algorithm for each video that provides a visual summary. These key frames are then applied as facets to the buildings of a cityscape. MediaMetro maintains a directory tree structure for the collection of video clips.

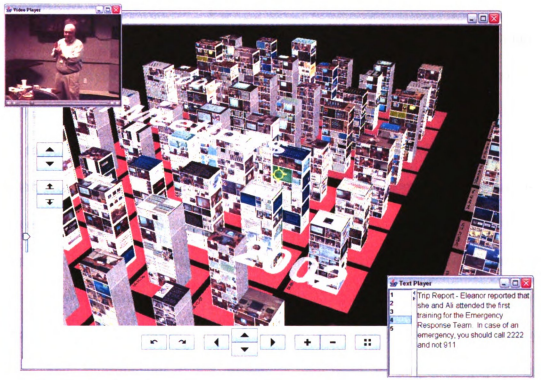


Figure 2.1 – A screen shot of a MediaMetro [2]

Unlike Vista, which uses a Space Warp Tram to navigate, MediaMetro uses a novel concept called 'Swoop' MediaMetro's swooping navigation technique helps users to navigate between high altitude overviews and ground level detail views of

the visual summaries rendered on the buildings in the cityscape. Unlike the swoop navigation technique used by MediaMetro the virtual video gallery system uses a natural navigation paradigm similar to human movement in the real world so as to help sustain a more realistic impression of the environment from the user's perspective.

The various stages in the development of a virtual art gallery and the importance of creating an atmospheric environment that will stimulate/engage the user are discussed by Cavazza and Mend [18]. Their work seeks to produce online analogues to physical art galleries as structured environments for browsing modern art. There are two types of organization design principles in a virtual environment that need to be considered. The first influences the structure/layout of the environment and the second influences the use of the environment. The impact that navigational aids and design has on ease of performing tasks in the environment has been found to be considerable. In certain cases, complex 3D environments with limited navigational aids fared worse than their 2D counterparts [19]. "Skilled way-finding" within a virtual environment is greatly dependent upon and enhanced by real world way-finding and environment design principles described in cognitive psychology literature and urban architectural design principles [3]. Similar results were also shown by Robertson et al, who described the impact *spatial memory* has on 3D navigation and on the improvement in Reaction Time [4]. Human spatial memory is the ability to remember where an item is in terms of physical coordinates or orientation

Users of large-scale virtual worlds tend to require structure to effectively navigate within the environment. Way finding augmentations such as direction indicators, maps, and path restriction can all greatly improve both way finding performance and overall user satisfaction. Darken et al,. discuss three critical factors of spatial knowledge which include landmark association, route mapping (procedural knowledge) and a combination of geocentric and egocentric mapping [3].

The virtual environment in this thesis uses corridors and passageways to symbolize 'Paths', walls for 'Edges', floors to symbolize 'Districts', a reception area to act as a 'Node', and a lift to serve as a 'Landmark' in addition to being a Node. These urban metaphors are used to create a spatial hierarchy within the virtual environment for way finding purposes. Experimental studies have demonstrated that a combination of a grid and maps as an absolute orientation frame of reference had the greatest impact in aiding user navigation, typically reducing navigation errors by 15% [3]. Space Warp Trams and swoops, while novel, do not fit within the category of urban metaphors. Hence, this thesis focuses on navigation methods that are grounded within traditional mechanisms of hallways, doors, and rooms. The virtual gallery does diverge from reality in some subtle and non-obvious ways, however. In particular, the virtual rooms are larger physically than is geometrically possible. This divergence allows for the creation of rooms that are more closely spaced than would otherwise be possible in reality, allowing users to be closer to the content. The unreal aspects of the

geometry are hidden from the user since they are never able to see more than one room at a time.

As each room within the virtual environment presents a collection of video clips that are exhibited at locations progressively farther from the user, occlusion of video clips further back in the room is a potent problem. The Cutting Plane interface introduced by Tanaka et al.,, attempts to mitigate occlusion of content in a room through the use of a near clipping plane, that allows users be ability to see beyond occlusions on demand [8]. Tanaka et al., 2004, described a solution to occlusion using the 'Cutting plane' concept and a means to easily identify the hierarchical structure of the nodes by using brightness and thickness as cues on a visualization tool similar to Treecube [8]. However, these virtual concepts again go beyond the physical gallery metaphor and must be selectively enabled. This thesis proposes a tiered gallery that will allow occlusion to be avoided in a natural way.

Data Mountain was envisaged as an alternative to the bookmark option of Internet Explorer [4]. In Data Mountain, links to web pages are shown as snapshots standing upright on a hillside. The user arranges the links to his or her liking anywhere in the environment. The underlying idea is the exploitation of human spatial memory, the inherent human ability to associate items with a spatial coordinate. Losh et al., also exploits spatial memory by the 'Method of Loci' a variation of the various Mnemonic link systems techniques that is commonly used to aid memory. The Method of Loci is a memorizing technique that uses places and

location (spatial memory) to remember order and locations of items on a list. The program structured the snapshots in a way that ensured that occlusion is minimized so as to ensure that as many documents as possible are visible. The hill-side metaphor for managing occlusion is similar to the tiered-gallery structure as presented in this thesis.

Due to the dynamic nature of a virtual environment, efficient indexing, querying and browsing are important. The MUVIS [20] project explored real-time video indexing and supported indexing, irrespective of the format. The MUVIS system is intended to support real-time video (with or without audio) indexing. For efficiency in video recording and indexing, last generation compression techniques such as MPEG-4 and H.263+ are used which also helps in indexing. MUVIS uses intra frames as key-frames for feature extraction

The database used in this study is a large collection of video clips. Directly indexing or comparing video content is impractical due to the massive amount of data involved and the extremely large amount of redundancy in that data. Most systems that seek to index or compare digital video reduce the content to some representation that captures salient information from the original content in a compact form. In this work, video summaries are constructed so as to generate relevant key frames for the video that are the subject to content-based image retrieval techniques.

A video summary is constructed using the process of segmentation, key frame extraction, and summarization. Many algorithms have been proposed for video

segmentation, often using the same basic concepts as image similarity measures (a scene break is a sequential pair of dissimilar images). Key frame extraction then selects a frame from a sequence that is representative of the sequence. This can be as simple as selecting a frame at a fixed point in the sequence (10 seconds in or the middle are common examples) or as complex as creating a composite image from motion analysis of the video sequence [21-24]. Key frames mostly serve two purposes in video retrieval--a visual summary of the video and a visual representative for image-based video retrieval [25]. Apart from the popular automated key frame selection technique of *Middle Image*, most automated key frame selection techniques such as 'Within Event' and 'Cross Event' use event segmentation of the video. 'Within Event' selects an image that has the closest value to the rest of the images in the event, as a key frame while 'Cross Event' additionally requires the image to be dissimilar to the images from other events [26]. The summarization process filters the collected key frames to remove redundancy. The summary comparison requires efficient algorithms specific to the comparison criteria [12, 27, 28]. A video clip consists of a large number of images representing a temporal sequence. This massive quantity of images is highly redundant. Summarization reduces this bulk to a smaller set of images that are representative of the video sequence by removing as much redundancy as possible.

Images selected or synthesized by the video summarization process then need to be compared so as to determine a content-based clip similarity. Many methods to achieve this result exist, of which the simplest and, in general, most robust

mechanism is the color histogram. Color histograms are computed by discretizing the color of each pixel within the image and counting the number of pixels in the image for each color. This information is then represented in the form of a vector. Color histogram results vary only slowly with the angle of view. Translation of an RGB image into the illumination invariant rg-chromaticity space allowed the histogram to operate well in varying light levels A color histogram characterizes an image (key frame from a video sequence) by its color distribution without using difficult to quantify information about object location, shape, and texture. The images are scaled to contain the same number of pixels before the process of converting the image into its equivalent color histogram so as to be invariant to image size. Colors within the image are then mapped into a discrete color space containing a specific subset of colors. This method has its shortcomings, especially if object mapping is required.

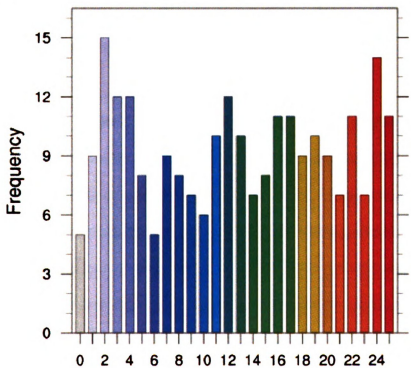


Figure 2.2 – An example of a color histogram having 25 bins

Color histograms in the HSV (hue, saturation, value) color space were preferred for this application compared to RGB color space as the hue and saturation components are intimately related to the way human visual system perceives color. Empirical studies have also shown a greater accuracy by human subjects when using HSV color space when compared to RGB [29].

Some of the other popular features used for the mapping of key frame data are the color or intensity gradient, edge density, textured-ness, number of pixels that differ beyond a particular threshold, and gradient magnitude [12]. Most of these variants of color histograms work well for limited types of images as most other implementations are object association-specific or are too general to classify based on object occurrence. A hybrid approach of using the prominent object, and the relation between objects in the image, to classify was proposed by Tseng et al [30].

Color coherence vectors constitute another intriguing approach which attempts to use the coherence of pixel colors to identify similarity [31]. Color histograms are unable to differentiate between an image having a large red object and an image having many scattered red dots. Color coherence vectors measure not only the colors in an image, but also color proximity. Pass et al. used a modification of the color histogram called the joint histogram and performed studies using over 210,000 images [12]. They demonstrated considerable improvement over traditional color histograms. One of the more prominent features of the new algorithm was the scene

break detection ability, which is useful in the selection of key frames for the video clips utilized in their project.

Chapter 3

Research Methodology

This thesis presents a new approach to video retrieval and browsing, an approach designed to enhance the user experience and the user's perceived comfort while navigating through a video archive. The user's engagement and the perceived feeling of familiarity is a crucial factor in the ability to effectively navigate through the video archive. As shown in Figure 3.1, a twofold approach to accomplish this was made by utilizing content-based video retrieval methods to cluster and group video clips based on dominant hues (Video Retrieval Engine), and then leveraging spatial and visual cognition (Browsing Environment). The Video Retrieval Engine and the Browsing Environment are detailed in the following subsections 3.1 and 3.2, respectively.

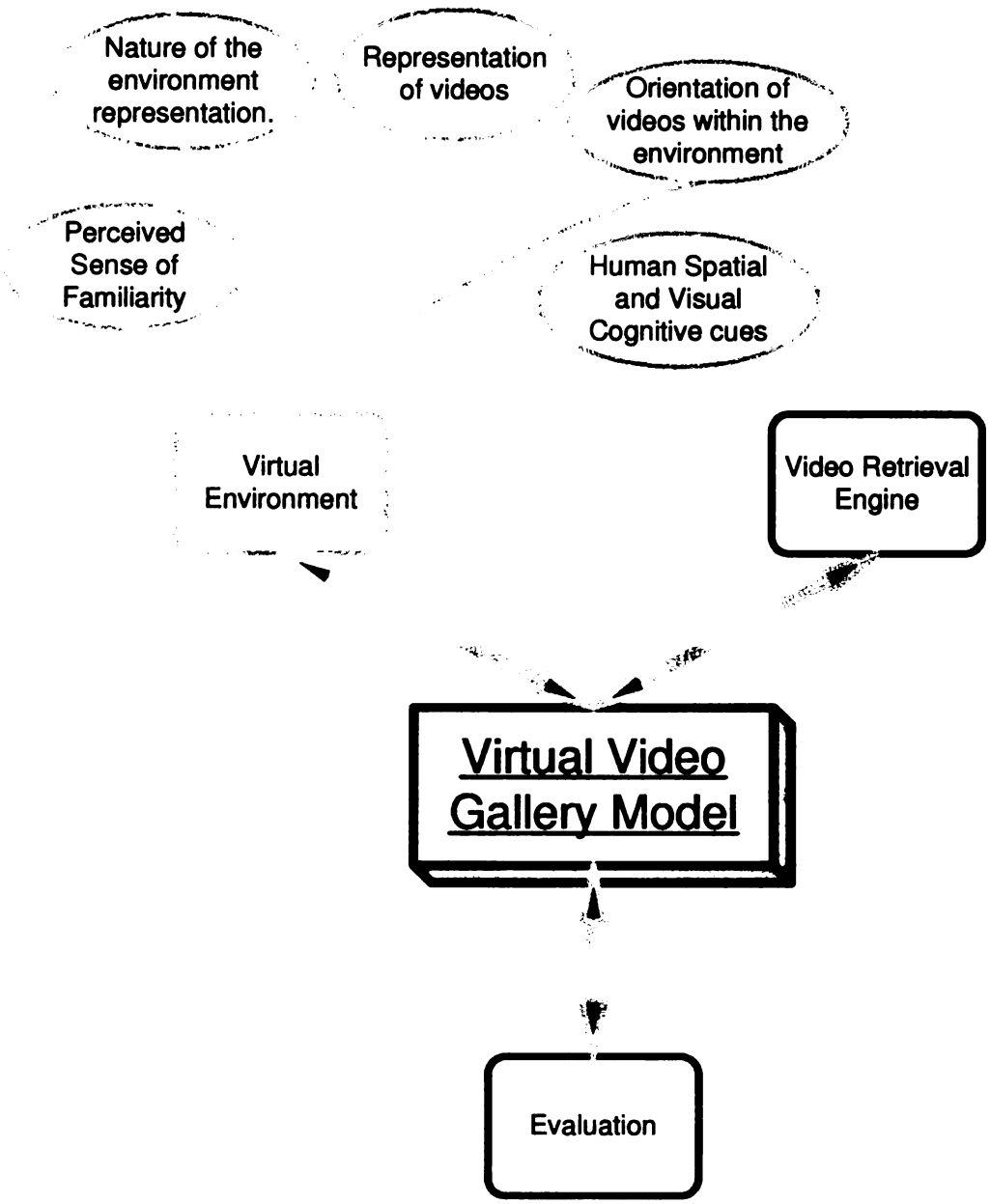


Figure 3.1 – Virtual video gallery architecture

3.1 Video Retrieval Engine

The first step in the video indexing process is to create a framework for video retrieval and grouping. There are numerous flavors of image retrieval [32-35]

available, many of which use histograms of the image as a means for comparison. Unlike some CBVR approaches that focus on the retrieval of the behavior of objects across the collection, this project implemented a simplified content-based retrieval method that is analogous to Content-Based Image retrieval (CBIR) to cluster video clips in the manner shown in Figure 3.2, such that the density and distribution of various colors in the selected key frames prominently contributed as a factor in grouping video clips. Using hue as a predominant influence in clustering allowed a synergic relation between the browsing environment and the grouping of video clips.

The Video Retrieval Engine has two phases. The first phase, groups video clips into clusters so as to pick one cluster (C_{SEED}) that best matches the user provided textual query in phase II. This cluster serves as the starting point of reference within the browsing environment. Metaphorically speaking, this is analogous to identifying which floor within the Museum, the user is sent to. The second phase sub-clusters video clips that belong to cluster C_{SEED} into smaller clusters that will be assigned to separate rooms. Details on Phases I and II follow in section 3.2 and 3.3, respectively.

3.1.1 Phase I

All steps in Phase I are done off-line. Phase I starts by first computing the hash table H_L to help with the retrieval of video clips based on the textual Meta Tags associated with the video clip. The video database contained metadata that is associated with each video. This metadata included a simple description of the video provided by the submitter and user-chosen keywords meant to categorize the content. These textual

meta-tags are used across multiple stages before resulting in the eventual clustering and grouping of video clips, most of which is computed offline. The system creates a hash table (H_L) of the list of video clips. The keys (K_{HL}) to the hash table are the collection of stemmed strings from textual meta-tags.

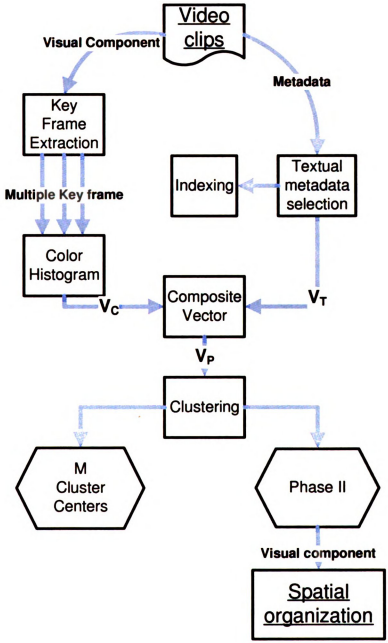


Figure 3.2 – Phase I architecture

For each video clip a key frame is extracted that is representative of the content of the video. As illustrated in Figure 3.2, key frames are extracted for each video, using conventional key frame extraction techniques based on scene break detection [36]. Key frames mostly serve two purposes in video retrieval - a visual summary of the video and a visual representative for image-based video retrieval [26]. Once the collection of key frames for a video is identified, a set of color histograms V_C is generated for the key frames.

A scoring mechanism was derived for identifying weights for each key frame to indicate the relative representative nature of the key frame among the key frames of the same video.. The less the variance in weight of the key frames of a clip the greater the intra-similarity of a video clip. Weight (W_K) for the K^{th} key frame of a video clip is computed using Equation 3.1

$$W_K = \frac{D_K}{\sum_I D_I} \tag{3.1}$$

D_K refers to the dissimilarity or distance between a key frame and all the other key frames that belong to the same video clip. D_K is computed by summing the Cosine similarity of the normalized color histogram vector of K^{th} key frame ($\vec{V}_K$)

with all other 'I' key frames ($\vec{V}_I$) that belong to the same video clip as show in Equation 3.2 and Equation 3.3.

$$D_K = \sum_I \theta_{KI} \tag{3.2}$$

Since all the tuples in the vector $\vec{V}_K$ are positive, the Cosine similarity of any two vectors $\vec{V}_K$ results in a value between zero (90 degrees) and one (0 degrees), where one represents maximum similarity and zero represents minimum similarity (actually complete orthogonality).

$$\theta_{KI} = \cos^{-1}\left(\frac{\vec{V}_K \bullet \vec{V}_{KI}}{\| \vec{V}_K \| \| \vec{V}_{KI} \|}\right) \tag{3.3}$$

Normalization is done by dividing $\vec{V}_K$ by $\| \vec{V}_K \|$ where $\| \vec{V}_K \|$ is the length of the vector ($\vec{V}_K$) that has 'J' tuples as shown in Equation 3.4.

$$\hat{V}_K = \frac{\vec{V}_K}{\| \vec{V}_K \|} = \frac{\vec{V}_K}{\sqrt{\sum_J \vec{V}_{KJ}^2}} \tag{3.4}$$

Now that weights have been assigned to each key frame, we cluster the Key Frames. The Key Frames are then clustered using K-means clustering [37-39] to return M clusters, where each video clip may belong to more than one cluster as a

video clip may have more than one Key Frame which have been assigned different clusters.

From the M clusters, the M cluster centers are identified. M is computed using techniques similar to Hartigan index that aims to make the sum of squared distance of data points in a cluster to its cluster center not monotone.

$$H(K) = \gamma(K)\frac{E_K - E_{K+1}}{E_{K+1}} \tag{3.5}$$

Where, $\gamma(K) = n - k - 1$ and E_K is the sum of squared distance of data points in a cluster to its cluster center, for all the clusters.

For the computation of the M clusters, each data point-($\vec{V}_P$) is represented as a combination of the content based component of the key frame($\vec{V}_C$) and a second vector representing the textual description of the video the key frame belongs to ($\vec{V}_T$).

$$\text{i.e., } \vec{V}_P = (\vec{V}_C, \vec{V}_T) \tag{3.6}$$

The color histogram ($\vec{V}_C$), which is a vector is then normalized ($\hat{V}_C$) to a length equal to 1 unit resulting in the terms of the vector having a range between 0

and 1. Normalization is done by dividing $\vec{V}_C$ by $\| \vec{V}_C \|$ where $\| \vec{V}_C \|$ is the length of the vector ($\vec{V}_C$).

$$\hat{V}_C = \frac{\vec{V}_C}{\|\vec{V}_C\|} = \frac{\vec{V}_C}{\sqrt{\sum_I \vec{V}_{CI}^2}} \tag{3.7}$$

Since, all the terms in ($\vec{V}_C$) are positive on account of the properties of the histogram vector, normalization of ($\vec{V}_C$) results in all the terms being positive and less than one. Normalization is performed so as to be invariant to video length or resolution, as a larger video with higher resolution would result in a larger number of frames in the clip which consequently would increase the range of values of the bins. For the same purpose, the video clips are converted to a common video file format (.WMV) and common screen resolution before the color histogram is computed.

V_T is a binary vector with each component of the vector indicating the presence or absence of a key (K_{HL}) to the lexicographical hash table (H_L) from the video clips textual meta-tags. K-Means algorithm works by minimizing the sum of distances/costs of each cluster (D_{Total}), where cost refers to the sum of the distances of each Key Frame to its respective cluster-center (C_{Cp}). The distance between C_{Cp} and a video $\vec{V}_P$ is computed in two parts – the distance function D_C and D_T.

D_C refers to the similarity of the visual component of the two data points, and was computed based on the color histogram values. D_C was found using the Cosine similarity of the $\vec{V}_C$ component of the two datapoints $\vec{V}_{P1}$ and $\vec{V}_{P2}$. Since all the tuples in the vector $\vec{V}_C$ are positive, the Cosine similarity of $\vec{V}_{P1}$ and $\vec{V}_{P2}$ would result in a value between zero (0 degrees) and one (90 degrees), where one represents maximum similarity and zero represents minimum similarity (actually complete orthogonality).

$$\theta_C = \cos^{-1}\left(\frac{\vec{V}_{P1} \bullet \vec{V}_{P2}}{\| \vec{V}_{P1} \| \| \vec{V}_{P2} \|}\right) \tag{3.8}$$

D_C is normalized so that it is comparable with D_T. Normalization of D_C is done by dividing the Cosine similarity value by its range, i.e., 90.

$$D_C = \frac{\theta_C}{90} \tag{3.9}$$

Similarly, for computing the distance function D_T, Jaccard similarity coefficient was used. The value of D_T is in the range 0 to 1, where 0 represents least similarity and 1 represents maximum similarity.

$$D_T = \frac{|\vec{V}_{T1} \cap \vec{V}_{T2}|}{|\vec{V}_{T1} \cup \vec{V}_{T2}|} \tag{3.10}$$

The objective function of K-Means algorithm is to minimize D_{Total} while learning the weights. W_C and W_T. The algorithm minimizes D_{Total} by minimizing D_C and maximizing D_T . i.e.,

$$Min \sum_{M} D_{TOTAL} = \sum_{M} ((W_C * {}_{MIN} \; D_C) + (W_T * {}_{MAX} \; D_T))_p \qquad (3.11)$$

Where,

D_{Total} = Sum of distances of all data points to their respective Cluster Centers

D_C = Cosine Similarity $(VC)_p$ to cluster center.

D_T = Jaccard Similarity Coefficient $(VT)_p$ to cluster center.

W_C = Weight of Content vector (VC)

W_T = Weight assigned to Meta Tags Vector (VT)

Since a video may have multiple key frames associated with it, it is possible that a video may belong to multiple clusters simultaneously. The cluster centers serve the purpose of reducing scalability problems by making it efficient to assign a new video to an existing cluster based on the similarity of the new video to the cluster centers. A batch process can occasionally re-cluster the collection off-line using KNN (K-nearest Neighbor).[40]

For each cluster, an intra-cluster similarity matrix is computed. These matrices are of varying sizes and on the order of N x N where N is the number of

video clips belonging to that particular cluster. Issues of quadratic growth were avoided by limiting cluster sizes. This intra-cluster similarity matrix is fed into phase II along with the M clusters created in phase I.

3.1.2 Phase II

This chapter details the design of Phase II of the Video Retrieval engine. Part of this phase is done at run-time.

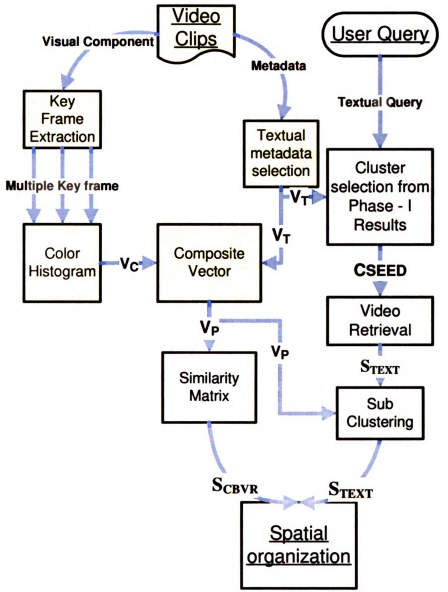


Figure 3.3 – Phase II architecture

Once the M clusters and M cluster-centers are identified at the end of Phase I, Phase II could begin with the selection of Cluster (C_{SEED}). The video clips belonging to (C_{SEED}) are further clustered as detailed below into sub clusters, such that each sub cluster is assigned a separate room in the museum metaphor.

As shown in Figure 3.3, the user entered textual query is needed to start Phase II. Based on the textual query provided and using the index created from V_T, the key frames of the videos that belong to the results of the textual query are identified and their associated weights (W_K) summed together for each of the M clusters from Phase I. A cluster (C_{SEED}) with the highest sum of weights associated with the key frames of the videos retrieval results is chosen as a starting point of the virtual gallery environment. The results of the textual query search is based on textual search of the video keywords (Meta Tags), a well understood and scalable process. However, this search is not meant to locate the desired content, but rather intended to put the user in the right neighborhood or floor of the museum (i.e., video clips that belong to C_{SEED}), such that a content-based method can then allow browsing of similar content.

Since it is not reasonable for a human user to specify a query as a video or image, this two-phase approach uses a texture query method as a starting point and a means to select one or more sample images that serve as the seed for visual browsing, which is followed by sub clustering of only the video clips present in the cluster chosen (C_{SEED}), so as to assign sub-clusters to the various rooms in the browsing environment for spatial ordering.

Using the same textual query provided by the user all the videos from within the cluster (C_{SEED}) that match the keywords in the textual query are retrieved. The retrieval of video was done solely based on the V_T Component of the Key Frame.

This subset of video's retrieved (S_{TEXT}) are then again clustered using K-means clustering to form K clusters, where K symbolically represented the number of rooms in the browsing environment. The equation used is

$$Min \sum_{M} D_{TOTAL} = \sum_{M} ((W_C *_{MIN} D_C) + (W_T *_{MAX} D_T))_{PTEXT} \quad (3.12)$$

Where,

D_{Total} = Sum of distances of all datapoints to their respective Cluster Centers.

D_C = CosineSimilarity(VC)p to clustercenter.

D_T = Jaccard SimilarityCoefficient (VT)p to clustercenter.

W_C = Weight of Content vector (VC) computed in Phase I

W_T = Weight assigned to Meta Tags Vector (VT) computed in Phase I

$P_{TEXT,} \in S_{TEXT}$

This subset of video clips retrieved (S_{TEXT}) formed the basis for a starting point from which additional video clips (S_{CBVR}) are retrieved before being eventually displayed in the environment.

The selection of the room to which each cluster belongs is done such that the cluster that has the cluster center video with the highest score based on the textual search query is placed in the room closest to the user. The cluster center of the retrieval result (S_{TEXT}) is placed precisely in the middle of the room at an elevation

higher than any other video clip (e.g. ceiling) as seen in the Figure 3.4. Also, way finding augmentations like additional props that complemented the scene of the museum like a sofa, staircases, etc, are added to the scene to provide points of reference and markers to help the user navigate the virtual environment as seen in Figure 3.5. These way finding augmentations were added to improve both way finding performance and overall user satisfaction as discussed by Darken et al, [3]

The video clips that have been assigned to a room are oriented from the middle of the room keeping in mind the orientation of the entrance to the room, such that the centroid video is placed precisely in the middle of the room at an elevated position. The video closest to the centroid video belonging to the sub cluster assigned to the room is placed at the far end of the room so that it directly faces the entrance and is in the line of sight of the user as the user enters the room. This is done so that the video is the first video that the user sees on entering the room. Apart from the centroid, the first video directly in the line of sight of the user as the user enters the room is the video closest to the centroid in the cluster center. Subsequent video clips belonging to the cluster are oriented along the circumference of the circle that is formed with the centroid at the center, such that farther the distance of the video from the cluster center, the farther from the first video on the circumference they were placed.

Figure 3.4 – A screen short of the CBVR environment showing a cluster center

The system next computed a similarity matrix (S_C) for the video clips in the database using both component (textual and contextual) of the composite vector. This step may be done offline. For each video oriented in the room, the two video clips closest according to the similarity matrix S_C were identified and oriented right above the respective video in the room, to form a second tier of video clips in the room. Similarly, for each of the video clips that were oriented on the second tier, their respective closest similar two video clips according to S_C were oriented above them to form three tiers, which ended up looking similar to the design of an amphitheatre. The video clips that were retrieved to fill the second and third tier of the room (S_{CBVR}) are the video clips that were retrieved based on the visual content, i.e., CBVR. The

videos were oriented such that each higher tier had a larger diameter so as to accommodate the larger number of videos that belonged to the tier.

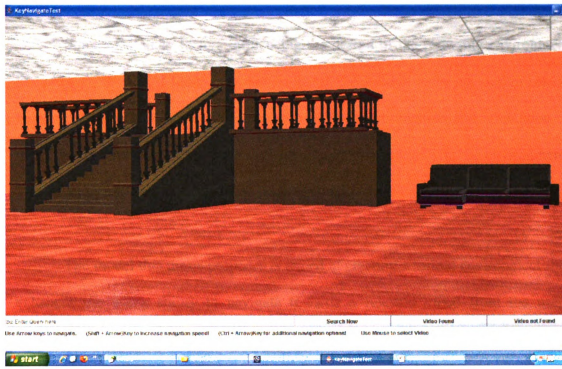


Figure 3.5 – A screen shot of the CBVR environment showing props in environment

3.2 Browsing Environment

Preliminarily, this work envisioned the browsing environment [41-44] as a modern day real world museum where each room within the museum is shaped like a small amphitheatre having retrieval results adorning the tiered, gradually ascending, wall Cognitive science has extensively studied the impact of environment layout on

human perception [45-47] . This work as used as a basis on which to design a browsing environment layout cognizant of human spatial abilities and likely to be an effective and useful visual representation of information.

Some of the key factors influencing the design of the browsing environment include a perception of familiarity; the impact the real world metaphor design has on human cognition and its resulting ease of use. The representation and orientation of the retrieval results should be such that they seamlessly merge with the environment. The effect should be that of museum exhibits and paintings on the wall [48, 49].

The model this thesis proposes is aimed at creating a virtual environment that enhances the user's ability to navigate towards a relevant video from the users perspective by using techniques incorporated in CyberMap [1] and MediaMetro [2]. Results from psychological and cognitive research were used to help identify a layout of the virtual environment that will increase the users 'ease of use' and familiarity. For example, identifying the impact of using real world metaphors on the usability [2] would be feasible .

Any proposed method is only an idea and the human-computer interface improvements due to that method are only theory until improved performance is demonstrated with a user evaluation. A study was conducted to determine any impact the combination of content based and metadata based retrieval of video clips has on user efficiency and estimate the extent of the enhancement on efficiency. A

comparison was made between conventional keyword search methods and the Virtual Video Gallery approach. The rubric for examination will be the ability to locate specific video results and the time it takes to achieve that result.

Chapter 4

Browsing Environment Concept

The browsing environment begins when the user enters the environment. The entry point is determined through the use of a textual query that guides the user to the relevant floor of the virtual gallery building where all the video clips returned as results from the textual query were present, i.e., from only video clips that belong to the clusters that was chosen as the seed (C_{SEED}). Alternative means of entry would include random placement in the world for exploration or a return to a previous room the user found interesting or relevant. The video clips that are returned by the textural query are placed within close proximity of the central hallway of the floor, though they may be distributed over the various rooms on the floor. Color histograms [14, 15, 31] of key frames, along with textual metadata, form a composite feature vector that is used to compare and cluster retrieval results. Multiple similarity matrices representing each cluster were maintained for browsing organization based on relevance.

The video clips are assigned to rooms based on the composite distance measure returned as a function of the color histograms vector component of the composite vector. The video clips selected by the textual query were arranged such that they are closest to the entrance of their respective room. In addition to these video clips, the room has other clips that are determined based on the content similarity to the retrieval results returned from the textual query. These were placed at locations farther into the room and behind the video clips that share similar content. Since each room is a collection of video clips that have similar color histograms, the rooms are presented using a color theme that is a contrasting color to the predominant hues of the retrieved content. This concept of grouping video clips based on the most prominent hue is based on Gestalt's Law of Similarity where the human mind groups similar elements to an entity based on the similarity relation of color, form and size [50].

Since users tend to limit their search to the first few returned results, the number of video clips returned based on contextual similarity is limited to a relatively small number of at max 18 per room. Based on the number of times a video is retrieved and the resulting/corresponding contextual similarity, the video is assigned a larger showcase and provided more space around it. The color of the frame encasing the video is an indicator of the luminance of the video as calculated from the color histograms. The richness in color of the frame or the showcase that encases a video is

representative of the size of the video. Similarly other facets of the video are embedded as visual cues around the video.

Darken and Silbert found that the addition of real world landmarks, such as paths, boundaries and directional cues, can greatly benefit navigation performance in a virtual environment; hence, the emphasis on creating an environment that closely approximates the real-world [3]. The goal of the Virtual Video Gallery design is to achieve this real world appearance, while providing a large number of cues to the contents in the room. These cues include the location of contents relative to the center of the room and to other contents in the room, the room theme itself, and appearance of the frame.

4.1 Scenario

This section presents an example scenario, describing the process a user would expect to see in the Virtual Gallery. A typical retrieval scenario begins with the user entering the gate/entrance of the Virtual Gallery; the starting point of the museum tour. Once the user enters the Gallery, the user is presented with a snapshot of the foyer with a dialog box that allows a query to be entered, as well as buttons for alternative general browsing with no initial query. The buttons either take a user to the central hallway of the gallery or to a random location in the gallery ("I'm feeling

lucky"). The dialog box serves as a means for the users to enter a query, which will refine the search results and provide a starting point for browsing. The snapshot serves as a landmark [3] and helps the user maintain navigational cues for efficient navigation within the environment.

The general browsing button gives the user access to all the floors within the Gallery, allowing free exploration of the gallery in its entirety. Each room on a floor is a collection of video clips that comprise a sub cluster. Though the grouping of clusters may be predetermined, the layout of the rooms and the floors are determined at runtime based upon a predefined threshold that limits the number of rooms per floor. Similarly the orientation and position of a video clip within a room is determined by its similarity relationship to the cluster center. The size of each room is governed by the size of the associated cluster and hence may vary from room to room.

The rooms have a hemispherical structure with a flat roof so as to provide a panoramic view to the user without occluding the video clips adorning the gradually ascending wall. The ascending curved walls provide a tiered surface upon which the video clips are mounted. The illusion that this provides is similar to the seating arrangement in a small amphitheatre where the rows of seats are replaced by video clips. To provide a more realistic and practical design, each tier had a glass railing and a means of approach using one of the aisles. The video representing the cluster center is hung precisely in the middle of the room. More distant a video is from the center of the room the less similar the video is to the cluster center. The texture of the

frame of a video provides information about the video such as the duration of the video.

Alternatively, and probably typically, the user could enquire at the reception desk by typing in a query. Then the user is sent to the hallway of a floor wherein the results of the search query could be found. All rooms within a floor share a collection of retrieval results with a common theme derived from the search query. Each room is distinguished from the other on the floor by the color of the entrance and representative images on the door, such that no two rooms within the floor are identified by the same color. This color code of the room serves as an identifier and indicator of the most prominent hue among the video clips within the room. From the user's perspective, each room is a collection of video clips that match the user's query such that video clips within each room share their most prominent colors belonging to a common spectrum. This helps the user select a room based on the color approximation of the visual content of the video that the user is interested in locating.

As an example, if the user supplied a textual query such as 'Mountaineering', it would be most likely that a majority of rooms of the floor the user is sent to would be color coded with hues representing snow, rock and mountains, representing the video clips from snow climbing, rock climbing and general mountain climbing. Thus, based on the color code of the rooms, the user may choose which room to enter. This process is assisted by representative images on the doors.

The room closest to the user in the hallway is the room containing the video returned with the highest similarity score in response to the query. The rooms are similar in design to the rooms observed in the general browsing environment with the exception that there were a few video clips placed upon pedestals in the center of the room. These are video clips that belong to the cluster and those that are among the top search results returned by the users query. The clips on the various tiers just behind the clips on the pedestals were video clips with high similarity scores within the cluster to the video clips on the pedestals.

Chapter 5

Experiment Setup

The experiment setup provided to the volunteers of the study involved two environments- a test environment and a control environment as shown in Figure 5.1 and Figure 5.2. The two environments were run on the same computer so as not introduce any bias on account of the setup. The content returned by the control environment for a particular query were solely based on the textual meta-data unlike the video returned in the test environment (CBVR), which used the proposed Virtual Gallery approach for the purpose of returning video clips and orienting them in the environment.

The prototype is meant to be a proof of concept and hence a relaxed design and a simplified version of the design is used for the CBVR environment. The prototype uses weights that helped collapse one or more stages of the design specifications into single stage for the purpose of simplicity. For example, the prototype assigned M=1 that effectively created a cluster C_{SEED} that contained all the

video of the repository. This decision was motivated by the limited no. of video clips (570) the system used. Similarly, in computing the color histogram of a video clip, instead of using key frames generated using video segmentation, all the frames of the video clip were used with equal weights assigned. Also, since the prototype wasn't going to be a 'live system' that would constantly have its database of videos updated, clustering related to cluster centers for the purpose of scalability were omitted in the design of the prototype.

The prototype also assumes the use of only one key frame for representing each video clip. This was done, as the primary reason to have more than one Key Frame based on Scene Break Detection was to help cluster video clips in Phase I while still allowing video clips to belong to more than one cluster at a time, which is not needed as we already have the Cluster C_{SEED}. The second motivation to have multiple Key Frames per video was to for the purpose of representing the video clip in the Browsing environments. For the purposes of this study, the video is not segmented during the video summarization process, since the collection represented all short-form content [51-55]. Instead, a single frame from a time point half way through the video clip (middle image) is selected as representative. This is a naïve method for key frame selection and does not guarantee that the key frame will be a good representation, but is actually common practice in many systems due to the complexity of automatically deciding a representative key frame from among the frames of a clip as noted by Smeaton et al,[25]. Selecting the middle frame provides a base-line performance for the prototype evaluation; it can only be assumed that better

chosen key frames will improve performance, especially since the system uses the key frame for the dual purpose of providing a visual summary as well as a means to assist image-based video retrieval.

A color histogram is computed for each video clip. The histogram is computed using all the frames of the video. The color histogram ($\vec{V}_C$) is a summation of color histograms of all the frames such that it is represented by the distribution of the colors of each pixel, across each frame. During the computation of the color histogram, the RGB color space is subdivided into 125 regions using a uniform subdivision of the intensity of each component into five discrete values. Each region is mapped to one of 125 histogram bins. The histogram is constructed by counting each pixel from every frame. A total of 125 histograms bins is used as a result of choosing 5 bins each for Red, Green and Blue. The selection of a 125 bin histogram is a common choice for a 3 dimensional histogram [56-59]. This was done for the purpose of expediency as a vast majority of the video content, on account of the nature of the source of the collection (www.youtube.com) were short durational video clips, consequently with fewer event segments, which otherwise would have dramatically skewed the color histograms if computed on the whole clip.

For this thesis, the prototype assigned equal weights to W_C and W_T for the two distance functions D_C and D_T, for the computation of D_{TOTAL} . Also, for the experiment, the prototype used only a limited amount of stemming (e.g., eliminating plurals, 'ing', etc) of the meta-tags.

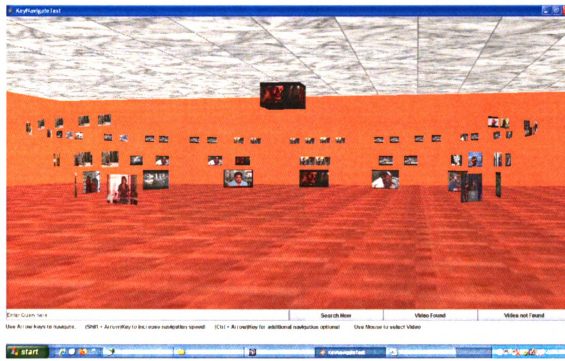


Figure 5.1- A screen shot of the CBVR environment

The two graphical environments were rendered using Java3D API. The video clips used during the study were retrieved from Youtube.com. As the experiment needed to physically have the database entirely available for analysis of and clustering, it was not practical to have the model retrieve video clips on the fly from the net, nor could the reliable availability of the content be assured during the experiment. Hence, the video clips used in the study were retrieved and stored into a local repository. A collection of 570 video clips was constructed for use during the study. The clips were retrieved from YouTube.com based on categories (at the time of retrieval) including 'Top Rated', 'Top Favorites', 'Most Discussed', 'Most Viewed', etc. To include a more realistic sample of the content actually available on popular video browsing websites, the experiment included video clips from categories

such as 'Recently Featured' that were likely to have less textual meta-tags associated with the video clips on account of the video being recently added.

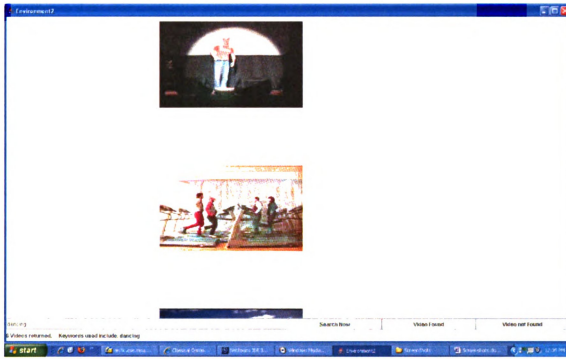


Figure 5.2: A screen shot of the control environment

Among the 570 video clips available, there were a number of redundant video clips on account of belonging to multiple categories or on account of being uploaded multiple times. The experiment excluded a considerable number of redundant video clips as the approach of grouping and orienting video clips in the CBVR environment inevitably clustered video clips that had similar visual content even if they didn't share the textual keywords entered by the user or didn't share the same meta-tags. This was the intention of the CBVR environment, but having too many duplicates video clips would have partially defeated one of the purposes of the study in that the

study tried to gauge if the participants were able to identify a pattern in the orientation of the video clips in the environment, as it would have made it extremely obvious to the participants in the study that video clips of similar visual content were grouped together in the same room.

Eventually, 354 video clips were used out of the 570 videos after removing redundant videos. The redundant video clips were not necessarily 100% identical, as most of them were essentially the same video uploaded by different people having different meta tags and possibly not necessarily having the same video length. The similarity matrices we computed D_C for the video clips were used to identify redundant videos. This in itself showed that similarity of visual content of the video clips based on color histograms D_C was working as expected.

Along with each video that was used, its associated metadata, especially the meta-tags were saved and processed, so as to be used for textual search capability.

5.1 Experiment Participants and Tasks

20 volunteers participated in the study. Each of the participants was asked to perform six tasks, with the two environments having three tasks each. In the interest of fairness, the subjects were asked to alternate environments between tasks and the environment that was chosen for the first task which was chosen at random for each

subject. No explanation was provided to the participants regarding the basis of the ordering and orientation of video clips in the environment so as to help gauge whether the users of the environment were explicitly or implicitly able to identify by themselves, if there were identifiable patterns that governed the grouping and orientation of the video clips in the two environments.

Each of the six tasks required the participant to find a video that matched the description provided. Once the video was found, the participant was asked to find another video that the subject believed was similar in visual content. Therefore, there was a requirement for the identification of two video clips per task. Since 'similarity of a video based upon visual content' is subjective, it was mentioned to the participant that the basis of what the participant considered similar visual content is subjective to his or her perception. A screen shot of the participants having selected a video is shown in Figure 5.3 and Figure 5.4

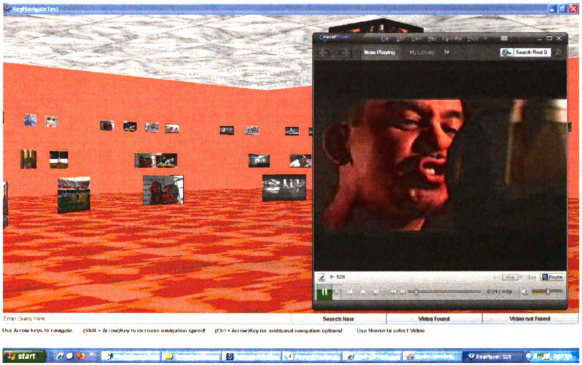


Figure 5.3 - A screen shot of a video being played in the CBVR environment

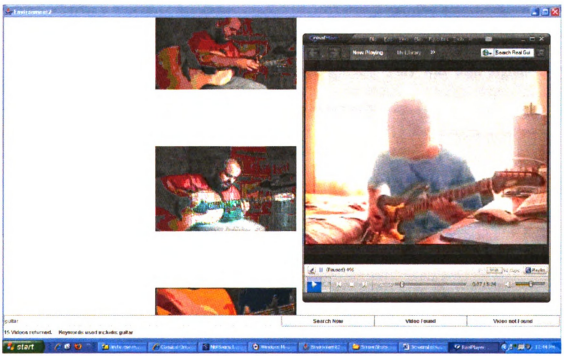


Figure 5.4 - A screen shot of a video being played in the control environment

Among the tasks that were provided for each environment, a clear description of the video to search for in the task was provided for two out of the three tasks. The 3rd task was open-ended where the participant was asked to search for a video that was of interest to the user. This was done so as to introduce a more realistic real-world scenario where the user decides to search for a video based on his/her own volition, while also reducing any bias that may creep in as a result of an exhaustive list of search descriptions used during the study.

The participant confirmed success of completion of a task by pressing a button that specified success was achieved. In the event that the participant was unable to accomplish the task, or did not believe it was possible to complete the task (because the subject believed the video that the participant was looking for didn't exist) or if the participant was bored or fed-up with the task at hand and wished to move on, the subject would signal that the task wasn't completed. This was also signaled by pressing a button that specified that success was not achieved.

5.2. Results

Based on the results taken from the 20 subjects who participated in the user study, it was found that 95% of the subjects had better or at least equal success in retrieving video clips. The average percentage of improvement using the CBVR environment in

success rate over all tasks was found to be 23.98%. As shown in Figure 5.5, this improvement in the overall task completion success rate worked out to increased Mean of Difference (M_D) = 9.68 points (on a scale of 100) with a level of significance of 3.009 and df =18 and a confidence of 99.62% for the directional test. The estimated standard deviation of sampling distribution of the mean of the difference between the two environments $est.\sigma_{MD}$ = 3.218.

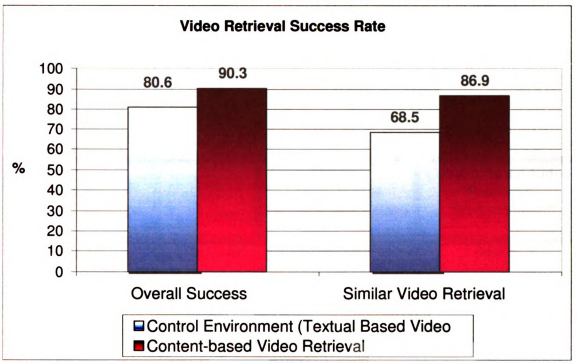


Figure 5.5 - Video retrieval success rate

The M_D in successfully finding a video that had similar visual content according to the criteria set by the subjects was found to be an increase of 18.36 points (on a scale of 100) with a confidence of 99.24% and $est.\sigma_{MD}$ = 6.845 for the

directional test. Here also, all but one of the study participants had a better or equal success rate. The only participant that had a lower success rate in the CBVR environment was the same participant who had a lower overall success rate. The percentage of improvement in finding a video that had similar visual content was found to 30.17%. This statistic doesn't include one of the participants who had 100% success rate in the test environment and a 0% success rate in the control environment as the percentage improvement would result in a division by 0 and it is not valid to compare successful results to complete failure. Further task completion success rate statistics are provided in the Table 5.2.1.

	Control Environment (Textual based video Retrieval)	**Content-based Video Retrieval Environment**
Overall video retrieval success rate	80.6 %	90.3 %
Similar video retrieval success rate	68.5 %	86.9 %
Overall video retrieval success rate per Query	4.5 %	17.9 %
Similar video retrieval success rate per Query	3.6 %	17 %

Table 5.2.1: Comparison of success rate in terms of percentage for the two environments

There was no statistical difference between the two environments as far as general categorical search of a video was concerned, which supports the argument that the search capability of the two environments were comparable and similar.

One of the more prominent trends was that the average number of queries used to accomplish the task was considerably lower in the test environment (CBVR environment) as compared to the control environment. The mean of the difference (M_D) in the number of queries entered by the participants of the study per successful video found was found to be 4.25 queries/video_found with a confidence of over 98.5% and $_{est}.\sigma_{MD}$ =2.774. This increase in the number of queries per video found was more than twice the average number of queries used to successfully find a video in the CBVR environment, which was 1.69 queries/video_found as shown in Figure 5.6. Further statistics regarding a quantitative comparison of queries entered is shown in Table 5.2.2

	Control Environment (Textual based video Retrieval)	**Content-based Video Retrieval Environment**
Average number of queries entered.	23.5	9.4
Average number of queries entered per successful retrieval of a video	4.9	1.7
Average of users average number of queries entered per successful retrieval of a video	6.0	1.8
Average number of video clips viewed per user	12.6	21.6

Table 5.2.2: Comparison of query related statistics for the two environments

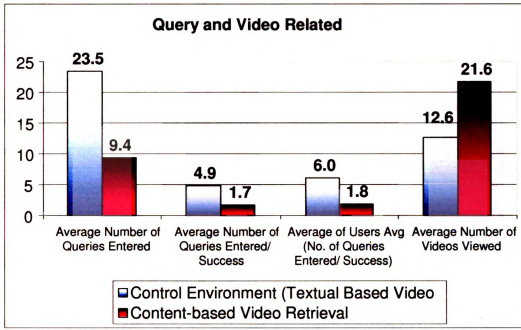


Figure 5.6 - Query and video related statistics

In terms of success rate per query entered, the M_D in success rate per query for the two environments showed an increase of 13.46 points (on a scale of 100) for the test environment with a confidence of 98.87% and $est.\sigma_{MD}$=4.769 for a non-directional test. A similar trend was also noticed in the success rate of finding a similar video per query for the two environments, with the CBVR environment showing an increased $M_D$of 13.45 points with $est.\sigma_{MD}$= 4.225 and a level of significance of 0.005 for a non-directional test as shown in Figure 5.7.

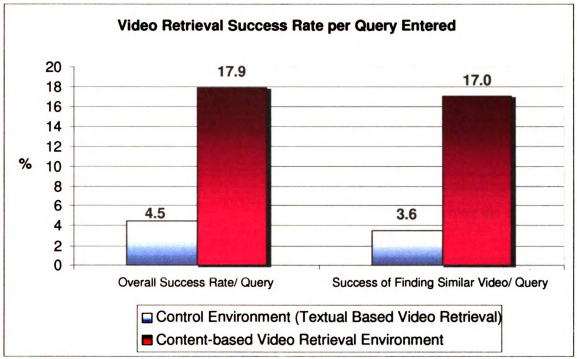


Figure 5.7 - Video retrieval success rate per query entered

During the study, 89% of the participants entered fewer queries per task in the CBVR environment as shown in Table 5.2.3. The average number of queries entered by the participants in the study was found to be 14.1 queries less with a $est.\sigma_{MD}$=2.475 in the CBVR environment than in the control environment. The level of significance of this directional test was found to be 5.69 with a p-value of 99.99%. This worked out to an average reduction of 50.79% in the number of queries entered by the user for the accomplishment of a task in the test environment as compared with the control environment.

Statistic	Improvement in the CBVR environment over the control environment
95%	Users had higher success in video retrieval in the CBVR environment.
89%	Users entered fewer queries per task in the CBVR environment.
89% improvement	In reduction of queries entered per task in the CBVR environment over the control environment.
23.98% improvement	In task completion success rate in the CBVR environment over the control environment.
Statistically insignificant difference	In categorical search between the two environments.
4.25 fewer	Queries entered per successful retrieval of video(in mean of difference- M_D terms)
13.5% improvement	in M_D of overall success rate per query entered.
13.5% improvement	in M_D of success rate of retrieving a similar video, per query entered.

Table 5.2.3 - Summary of statistical improvement of the CBVR environment in comparison with the control environment

The post-experiment survey showed that the participants in the study felt that they viewed more video clips in the test environment than the control environment. This increase in the video clips viewed according to the participants of the survey was a mean average of 2.4 points (on a scale of 10). This was found to be statistically

significant with a confidence of 99.96% in a non-directional test. When this result was compared with the actual data collected from the experiment it was found that the user's estimation of the environment was accurate and that there was in fact an 88.10% increase in the number of video clips viewed per task at an average in the CBVR environment. This can partially be attributed to the observation that the participants enjoyed exploring and browsing the CBVR environment more than the test environment. This was reflected in the post-experiment survey where the participants reported a 65% increase in enjoyment level with 99.99% confidence for the non-directional test.

While studying the participant's propensity for choosing a video from among the list of returned video clips for checking whether the video matches the user's criteria for selection, it was found that 33.82% of video clips viewed by the user were video clips that were returned to the user as a result of content-based video retrieval (S_{CBVR}). Even more interesting was the fact that when it came to the participant's task of searching for a video, 44.84% of successful attempts at selecting the video clips as the video of choice were S_{CBVR} video clips. Similar was the case when the participant eventually identified a video that matched the video content the user was looking for, with 45.47% of the video clips chosen belonging to S_{CBVR} (Table 5.2.4).

The video in each room that represented the centroid of the cluster was chosen by only 10% of the participants' completed action as the video of final choice, with a low level of significance for the degree of freedom.

Attributes of viewing by percentage of S_{CBVR} video clips set	**Percentage**
Percentage of video clips viewed by the user belonged to the set S_{CBVR}.	33.8%
Percentage of video clips that marked a successful attempt, that belonged to the set S_{CBVR}	44.8%
Percentage of video clips that marked a successful attempt of finding a similar video, that belonged to the set S_{CBVR}	45.5%

Table 5.2.4: Summary of percentage of S_{CBVR} video clips that were viewed by the user in the CBVR environment

In the post-experiment survey, participants were asked how fast they felt they could perform a particular task in the two environments; the participants felt that they were 19.84% slower in the test environment. This was mostly attributed to the fact that the CBVR environment was considerably slower in rendering the video clips in the environment, a characteristic of the prototype implementation and the general observation that rendering multiple video images in a 3D environment is more complex computationally than the presentation of a simple list. This observable increase in time for the CBVR environment when it came to returning to the video clips after a query was entered by the participant was a result of the implementation of graphical rendering on the computer and not a result of the conceptual implementation.

For the questions from the post-experiment survey, which asked how 'efficient' the environments were at retrieving a particular video, the results showed that the participants favored the CBVR environment over the control environment by 30.68%. Similarly, for the question of how 'confident' the participants were that they had identified or found the video that best matched the search, the survey found a 13.51% increased favor for the CBVR environment as shown in Table 5.2.5.

Survey result	Control environment (Text-based video retrieval)	Content-based video retrieval environment
Total number of video clips viewed	Less	More
Time spent in the environment	Less	More
'**Efficiency**' of the environment in retrieving a video	Less	More
'**Confidence**' of the user of having found the best video that matched the task	Less	More
'**Prominence of a pattern**' that grouped video clips	Statically insignificant	Statically insignificant

Table 5.2.5- Post-survey feedback

As far as the question of how prominent the pattern that grouped the video in the test environment when compared to the control environment, there was no statistically significant difference. Correlating the results from the post-experiment survey about the grouping of video clips and the results from the actual experiment when it came to finding a similar video, it stands to reason that patterns that grouped

video clips based on CBVR were more implicit that explicit. An interesting statistic related to the perception of a pattern that grouped the video clips in the two environments was that 100% of those participants that had lower success rate when it came to searching a video in the test environment as compared with the control environment claimed that the control environment had a more prominent pattern in the grouping of video clips, when in fact the orientation of the video clips in the control environment was purely random.

5.3 Discussion and Conclusion

The results from the study showed a strong relationship supporting the use of CBVR in combination with other video metadata retrieval method in improving the ability of the user to retrieve video clips, especially if the search is based on visual content. The enjoyment factor of browsing video clips was also found to be substantially higher in a more realistic browsing environment (a museum metaphor). This also contributed to the implicit identification of patterns in the grouping of the video clips.

Chapter 6

Conclusion and Future Work

The enhanced video browsing environment presented in this thesis primarily aims at providing a means for the user to search video databases using content-based video retrieval in a manner which does not require the user to provide a video clip as an example for query-by-content. Most Content-based Image retrieval (CBIR) techniques require the user to provide an image to return results. Since it is not always feasible for users to have access to video clips similar in visual content to the video intended to be retrieved, alternative means for enabling retrieval, such as accepting textual queries or displaying similarity relationships between video clips, is proposed in this thesis as an alternative.

The other advantage of the enhanced video browsing environment presented is the ease of use attributed to the user's perceived sense of familiarity when navigating within the environment. This sense of familiarity is attributed to the realistic modeling of the environment whose design has been extended from real world

metaphors. The browser can run in a Java applet in a conventional browser and appear to be a portal into a video browsing universe.

The next stage in this work would be to examine means of further improvement and exploration of the balance between the weights of content-based scores and conventional textual-based video retrieval scores, in returning to video clips and the orientation of the video clips in the environment to further exploit human cognition for pattern discovery. One of the potential areas of improvement involves the selectivity of the extraction of key frames from each video, such that the key frames are selected are a better indication of the video. This can be realized with the help of event segmentation. Also variants of image segmentation like the combinational approach of 'Within Event and Image Quality Fusion' and 'Cross Event and Image Quality Fusion' may be employed [26].

Another area of improvement with regard to similarity of video clips based on visual content is the use of individual key frames during the comparison of video clips as against the present method used where key frames is combined together before similarity of video clips is computed. The present model uses the basic content-based image retrieval approach of color histograms for CBVR, but, as discussed in the related work section, there are various alternative approaches to CBIR that can be experimented with to see how they perform in a virtual environment.

The textual query search engine used in the current model can be vastly improved using some of the various contemporary string-matching algorithms. Also, one of the other parameters that has significant potential for improvement is the weight set that is assigned in balancing V_C and V_T during the computation of similarity of video clips.

There is a tremendous potential for improvement by creating a more realistic virtual environment to further exploit the user's ability to find patterns in the environment to aid with the browsing of video clips. This first experiment was rather crude visually, focusing on layout and organization of the space, rather than the details of a finely crafted virtual gallery. Many of the proposed concepts of a virtual gallery, including contrasting color rooms, luminance keyed frames, and the impact of a more realistic virtual world is yet to be examined.

Additional variations of the use of physical metaphors to enhance the impact of spatial memory on 3D navigation for the video gallery remain to be explored. Similarly, innovative methods to use video metadata to describe content in the browsing environment can be tested.

Chapter 7

Appendix

Pre-Experiment Survey

1) What is your Age?	_____				
2) Gender (Male, Female, Do not wish to disclose)	Male ☐	Female ☐	Do not Disclos ☐		
	YES	NO			
3) Do you visit websites that store videos like 'youtube.com' or Google videos?	☐	☐			
4) Do you play 3D computer games?	☐	☐			
	1	**2**	**3**	**4**	**5**
5) How often do you use a computer? (1- Never, 5- Very Often)	☐	☐	☐	☐	☐
6) How often do you play 3D computer games? (1- Never, 5- Very Often)	☐	☐	☐	☐	☐
10) How often do you visit websites like 'youtube.com' and Google Videos? (1- Never, 5- Very often)	☐	☐	☐	☐	☐
11) Do you like trying new technologies? (1- Don't like, 5- Very much)	☐	☐	☐	☐	☐
12) When you hear a website has become popular, how likely are you to visit that site? (1- Not likely at all, 5- Absolutely certain)	☐	☐	☐	☐	☐

Post-Test Survey

The following questions are related to "Conventional Text-based Video Search Environment". 1 2 3 4 5 6 7 8 9

1) How favorable an impression did the environment make?
(1-Not favorable, 10- Very favorable)

2) How fast can you perform a task of searching and retrieval of a video in this environment?
(1-Very Slow, 10-Very Fast)

3) How easy was it to search for a video using this environment?
(1-Very Hard, 10-Very Easy)

4) How enjoyable was it to search for a video using this environment?
(1-No fun, 10-Lot of Fun)

5) How good was this environment in grouping similar videos?
(1-Very bad, 10-Very good)

6) How prominent was the pattern that grouped videos together in this environment?
(1-Not obvious 10-Very obvious)

7) For a large collection of similar videos, how efficient do you think this environment would be in retrieving a particular video?
(1-Not efficient 10-Very efficient)

8) Using this environment, did you end up watching a lot of videos before selecting the video of choice?
(1- Not many 10-A lot)

9) How confident are you that you found the video that best matched the search criteria you were looking for?
(1-Did not find, 10-Found)

The following questions are related to the "3D-video Browsing Environment". 1 2 3 4 5 6 7 8 9

10) How favorable an impression did the environment make?
(1-Not favorable, 10- Very favorable)

11) How fast can you perform a task of searching and retrieval of a video in this environment?
(1-Very Slow, 10-Very Fast)

12) How easy was it to search for a video using this environment?
(1-Very Hard, 10-Very Easy)

13) How enjoyable was it to search for a video using this environment?
(1-No fun, 10-Lot of Fun)

14) How good was this environment in grouping similar videos?
(1-Very bad, 10-Very good)

15) How prominent was the pattern that grouped videos together in this environment?
(1-Not obvious 10-Very obvious)

16) For a large collection of similar videos, how efficient do you think this environment would be in retrieving a particular video?
(1-Not efficient 10-Very efficient)

17) Using this environment, did you end up watching a lot of videos before selecting the video of choice?
(1- Not many 10-A lot)

18) How confident are you that you found the video that best matched the search criteria you were looking for?

Chapter 8

Bibliography

1. Gloor, P.A. CYBERMAP: yet another way of navigating in hyperspace in Proceedings of the third annual ACM conference on Hypertext 1991. San Antonio, Texas, United States ACM Press.

2. Chiu, P., et al. MediaMetro: browsing multimedia document collections with a 3D city metaphor in Proceedings of the 13th annual ACM international conference on Multimedia 2005 Hilton, Singapore: ACM Press.

3. Darken, R.P. and J.L. Sibert, Navigating large virtual spaces Int. J. Hum.-Comput. Interact., 1996. 8 (1): p. 49-71.

4. Robertson, G., et al. Data mountain: using spatial memory for document management in Proceedings of the 11th annual ACM symposium on User interface software and technology 1998. San Francisco, California, United States ACM Press.

5. Erol, B., K. Berkner, and S. Joshi. Multimedia thumbnails for documents in Proceedings of the 14th annual ACM international conference on Multimedia MULTIMEDIA '06. 2006. Santa Barbara, CA: ACM Press

6. Dengel, A., et al. Human-centered interaction with documents in Proceedings of the 14th annual ACM international conference on Multimedia MULTIMEDIA '06. 2006. Santa Barbara, CA: ACM Press

7. Huotari, J., K. Lyytinen, and M. Niemel, Improving graphical information system model use with elision and connecting lines ACM Trans. Comput.-Hum. Interact. , 2004. 11 (1): p. 26-58.

8. Tanaka, Y., Y. Okada, and K. Niijima. Interactive interfaces of Treecube for browsing 3D multimedia data in Proceedings of the working conference on Advanced visual interfaces 2004. Gallipoli, Italy ACM Press.

9. Cockburn, A. and B. McKenzie. 3D or not 3D?: evaluating the effect of the third dimension in a document management system in Proceedings of the SIGCHI conference on Human factors in computing systems 2001 Seattle, Washington, United States ACM Press.

10. Rautiainen, M., T. Seppänen, and T. Ojala. On the significance of cluster-temporal browsing for generic video retrieval: a statistical analysis in Proceedings of the 14th annual ACM international conference on Multimedia MULTIMEDIA '06. 2006. Santa Barbara, CA: ACM Press

11. Datta, R., J. Li, and J.Z. Wang. Content-based image retrieval: approaches and trends of the new age. in Proceedings of the 7th ACM SIGMM international workshop on Multimedia information retrieval 2005. Hilton, Singapore ACM Press New York, NY, USA

12. Pass, G. and R. Zabih, Comparing images using joint histograms Multimedia Syst., 1999. 7 (3): p. 234-240.

13. Stehling, R.O., M.A. Nascimento, and A.X. Falcão. A compact and efficient image retrieval approach based on border/interior pixel classification. in Proceedings of the eleventh international conference on Information and knowledge management CIKM '02 2002. McLean, Virginia, USA: ACM Press.

14. Brown, M.G., et al. Automatic content-based retrieval of broadcast news in Proceedings of the third ACM international conference on Multimedia. 1995. San Francisco, California, United States ACM Press.

15. Flickner, M., et al., Query by image and video content: the QBIC system. IEEE Computer, 1995. 28: (9): p. 23-32

16. Wada, N., S.i. Kaneko, and T. Takeguchi, Using color reach histogram for object search in colour and/or depth scene Pattern Recogn, 2006 39(5): p. 881-888.

17. Smith, A. and M. Webster, Developing a Global Repository and Showplace for Imagery Data Proceedings of the Theory and Practice of Computer Graphics 2003, 2003 p. 178.

18. Cavazza, M. and S.J. Mend, VIRTUAL ART GALLERIES: A NEW KIND OF CULTURAL OBJECTS? IEEE Computer, 2001: p. 590-593.

19. Hendricks, Z., J. Tangkuampien, and K. Malan. Virtual galleries: is 3D better? in Proceedings of the 2nd international conference on Computer graphics, virtual Reality, visualisation and interaction in Africa 2003. Cape Town, South Africa: ACM Press.

20. Gabbouj, M., et al., MUVIS: A Multimedia Browsing, Indexing And Retrieval System.

21. Aoki, H., et al. (1997) A shot classification method of selecting effective key-frames for video browsing. Proceedings of the fourth ACM international conference on Multimedia Volume, 1 - 10

22. Bezerra, F.N. and E. Lima. Low cost soccer video summaries based on visual rhythm in Proceedings of the 14th annual ACM international conference on Multimedia MULTIMEDIA '06. 2006. Santa Barbara, CA: ACM Press

23. Smeaton, A.F., et al. Automatically selecting shots for action movie trailers. in Proceedings of the 14th annual ACM international conference on Multimedia MULTIMEDIA '06. 2006. Santa Barbara, CA: ACM Press

24. Tse, T., et al., Dynamic key frame presentation techniques for augmenting video browsing. AVI -Proceedings of the working conference on Advanced visual interfaces 1998 p. 185 - 194

25. Smeaton, A.F. and P. Browne, A usage study of retrieval modalities for video shot retrieval. Information Processing and Management, 2005..

26. Doherty, A.R., et al. Investigating keyframe selection methods in the novel domain of passively captured visual lifelogs. in Proceedings of the 2008 international conference on Content-based image and video retrieval. 2008. Niagara Falls, Canada

27. Fleischman, M., P. Decamp, and D. Roy. Mining temporal patterns of movement for video content classification in Proceedings of the 14th annual ACM international conference on Multimedia MULTIMEDIA '06. 2006. Santa Barbara, CA: ACM Press

28. Setia, L., et al. Image classification using cluster cooccurrence matrices of local relational features in Proceedings of the 14th annual ACM international conference on Multimedia MULTIMEDIA '06. 2006. Santa Barbara, CA: ACM Press

29. Schwarz, M.W., W.B. Cowan, and J.C. Beatty, An experimental comparison of RGB, YIQ, LAB, HSV, and opponent color models. ACM Trans. Graph. , 1987. 6(2): p. 123-158

30. Tseng, V.S., C.-J. Lee, and J.-H. Su, Classify By Representative Or Associations (CBROA): a hybrid approach for image classification Proceedings of the 6th international workshop on Multimedia data mining: mining integrated media and complex data 2005 Chicago, Illinois: ACM Press. 61-69.

31. Pass, G., R. Zabih, and J. Miller. Comparing images using color coherence vectors in Proceedings of the fourth ACM international conference on Multimedia 1996 Boston, Massachusetts, United States ACM Press.

32. Jing, F., et al. IGroup: a web image search engine with semantic clustering of search results in Proceedings of the 14th annual ACM international conference on Multimedia MULTIMEDIA '06. 2006. Santa Barbara, CA: ACM Press

33. Kherfi, M.L., D. Ziou, and A. Bernardi, Image Retrieval from the World Wide Web: Issues, Techniques, and Systems ACM Comput. Surv., 2004: p. 35-67.

34. Luo, H., et al. Large-scale news video retrieval via visualization in Proceedings of the 14th annual ACM international conference on Multimedia MULTIMEDIA '06 2006. Santa Barbara, CA: ACM Press.

35. Gao, Y.i., H. Luo, and J. Fan. Searching and browsing large scale image database using keywords and ontology in Proceedings of the 14th annual ACM international conference on Multimedia MULTIMEDIA '06. 2006. Santa Barbara, CA: ACM Press

36. Owen, C.B. and J.K. Dixon, Adaptive video segmentation and summarization, in Handbook of Video Databases. 2003, CRC Press LLC. p. 299-318.

37. Redmond, S.J. and C. Heneghan, A method for initialising the K-means clustering algorithm using kd-trees Pattern Recogn. Lett., 2007. 28(8): p. 965-973.

38. Ding, C. and X. He. K-means clustering via principal component analysis in Proceedings of the twenty-first international conference on Machine learning 2004. Banff, Alberta, Canada ACM Press.

39. Ahmad, A. and L. Dey, A k-mean clustering algorithm for mixed numeric and categorical data. Data & Knowledge Engineering 2007. 63(2): p. 503-527

40. Soucy, P. and G.W. Mineau. A Simple KNN Algorithm for Text Categorization. in Proceedings of the 2001 IEEE International Conference on Data Mining 2001.

41. Adams, B., S. Greenhill, and S. Venkatesh. Browsing personal media archives with spatial context using panoramas in Proceedings of the 14th annual ACM international

conference on Multimedia MULTIMEDIA '06. 2006. Santa Barbara, CA: ACM Press

42. Hürst, W., T. Lauer, and R. Kaschuba. Interfaces for interactive audio-visual media browsing. in Proceedings of the 14th annual ACM international conference on Multimedia MULTIMEDIA '06 2006. Santa Barbara, CA: ACM Press.

43. Schmandt, C. Audio hallway: a virtual acoustic environment for browsing in Proceedings of the 11th annual ACM symposium on User interface software and technology UIST '98 1998. Santa Barbara, CA: ACM Press

44. Welch, G., et al., Projected Imagery in Your "Office of the Future". 2000 IEEE Computer Graphics and Applications, 2000: p. 62-67.

45. COLAVITA, F.B., Human sensory dominance (auditory versus visual stimulus) Perception and Psychophysics. Vol. 16. Oct. 1974. 409-412.

46. Vicki Bruce, P. and M.G. R Green, Visual Perception: Physiology, Psychology and Ecology 1996: Psychology Press (UK) 416.

47. Gibson, J.J., The Ecological Approach to Visual Perception 1986: Lawrence Erlbaum Associates.

48. V. Vinayagamoorthy, A.B., M. Gillies, M. Slater, A. Steed. An Investigation of Presence Response across Variations in Visual Realism. in Presence: The 7th Annual International Workshop on Presence. . 2004.

49. Peter, L., L. Patrick, and C. Alan, Psychophysically based artistic techniques for increased perceived realism of virtual environments, in Proceedings of the 2nd international conference on Computer graphics, virtual Reality, visualisation and interaction in Africa. 2003, ACM Press: Cape Town, South Africa.

50. Dempsey, C., D. Laurence, and E.T. Juhani, Gestalt theory in visual screen design: a new look at an old subject, in Proceedings of the Seventh world conference on computers in education conference on Computers in education: Australian topics - Volume 8. 2002, Australian Computer Society, Inc.: Copenhagen, Denmark.

51. He, L., et al. Auto-Summarization of Audio-Video Presentation. in ACM Multimedia. 1999. Orlando, Fl, USA: ACM.

52. Hanjalic, A. and R.L. Lagendijk, Automated High-Level Movie Segmentationi for Advanced Video-Retrieval Systems. IEEE Transaction on circuits and systems for video technology, 1999. 9(4).

53. Vasconcelos, N. and A. Lippman, A spatiotemporal Motion Model for video Summarization, in IEEE. 1998.

54. Toklu, C., S.-P. Liou, and M. Das. VideoAbstract: A Hybrid approach to generate semantically meaningful video summaries. in IEEE. 2000.

55. Zhang, H.J., A. Kankanhalli, and S.W. Smoliar, Automatic Partitioning of full-motion Video. Multimedia Syst., 1993. 1: p. 10-28.

56. Kerminen, P. and M. Gabbouj. IMAGE RETRIEVAL BASED ON COLOR MATCHING. in Proceedings of FINSIG. 1999.

57. Benitez, A.B. and S.-F. Chang. Image classification using multimedia knowledge networks. in International Conference on Image Processing, 2003. ICIP 2003. 2003. Dept. of Electr. Eng., Columbia Univ., New York, NY, USA.

58. Gong, Y. and X. Liu. Generating optimal video summaries. in IEEE International Conference on Multimedia and Expo (ICME). 2000. New York, NY, USA.

59. Lee, H.Y., H.K. Lee, and Y.H. Ha, Spatial color descriptor for image retrieval and video segmentation. IEEE Transactions on Multimedia, 2003. 5(3): p. 358- 367.

MICHIGAN STATE UNIVERSITY LIBRARIES
3 1293 03062 5242