Visual Search for Musical Performances and Endoscopic Videos

Degree's Final Project Dissertation
Telecommunications Engineering

Author: Jennifer Roldán Carlos
Advisors: Mathias Lux and Xavier Giró-i-Nieto

Alpen-Adria University of Klagenfurt (AAU Klagenfurt)
Universitat Politècnica de Catalunya (UPC))
2014 - 2015

Abstract

This project explores the potential of LIRE, an en existing Content-Based Image Retrieval (CBIR) system, when used to retrieve medical videos. These videos are recording of the live streams used by surgeons during the endoscopic procedures, captured from inside of the subject. The growth of such video content stored in servers requires search engines capable to assist surgeons in their management and retrieval. In our tool, queries are formulated by visual examples and those allow surgeons to re-find shots taken during the procedure.

This thesis presents an extension and adaptation of Lire for video retrieval based on visual features and late fusion. The results are assessed from two perspectives: a quantitative and qualitative one. While the quantitative one follows the standard practices and metrics for video retrieval, the qualitative assessment has been based on an empirical social study using a semi-interactive web-interface. In particular, a thinking aloud test was applied to analyze if the user expectations and requirements were fulfilled.

Due to the scarcity of surgeons available for the qualitative tests, a second domain was also addressed: videos captured at musical performances. These type of videos has also experienced an exponential growth with the advent of affordable multimedia smart phones, available to a large audience. Analogously to the endoscopic videos, searching in a large data set of such videos is a challenging topic.

Resum

Aquest projecte investiga el potencial de LIRE, un sistema existent de recuperació basat en contingut d'imatge (CBIR) utilitzat en el domini mèdic. Aquests vídeos són enregistraments a temps real de l'interior dels pacients i són utilitzats per cirurgians durant les operacions d'endoscòpia. La creixent demanda d'aquest conjunt de vídeos que són emmagatzemats a diferents servidors, requereix nous motors de cerca capaços de donar suport a la feina dels metges amb la seva gestió i posterior recuperació quan es necessiti. A la nostra eina, les consultes són formulades mitjançant exemples visuals. Això permet als cirurgians tornar a trobar els diferents instants capturats durant la intervenció.

En aquesta tesi es presenta una extensió i adaptació del Lire per a la recuperació de vídeo basat en característiques visuals i late fusion. Els resultats són avaluats des de dues perspectives: una quantitativa i una qualitativa. Mentre que la part quantitativa segueix l'estàndard de les pràctiques i mètriques per vídeo retrieval, l'avaluació qualitativa ha estat basada en un estudi social empíric mitjançant una interfície web semiinteractiva. Particularment, s'ha emprès el mètode "thinking aloud test" per analitzar si la nostra eina compleix amb les expectatives i necessitats dels usuaris a l'hora d'utilitzar l'aplicació.

A causa de l'escassetat de metges disponibles per dur a terme les proves qualitatives, el treball s'ha adreçat també a un segon domini: conjunt de vídeos d'esdeveniments musicals. Aquest tipus de vídeos també ha experimentat un creixement exponencial amb l'arribada dels smart phones i es troben a l'abast d'un públic molt ampli. Anàlogament als vídeos endoscòpics, fer una cerca en una gran base de dades d'aquest tipus també és un tema difícil i motiu d'estudi.

Resumen

Este proyecto investiga el potencial de LIRE, un sistema existente de recuperación basado en contenido de imagen (CBIR) utilizado en el dominio médico. Estos vídeos son grabaciones a tiempo real del interior de los pacientes y son utilizados por cirujanos durante las operaciones de endoscopia. La creciente demanda de este conjunto de vídeos que son almacenados en diferentes servidores, requiere nuevos motores de búsqueda capaces de dar soporte al trabajo de los médicos con su gestión y posterior recuperación cuando se necesite. En nuestra herramienta, las consultas son formuladas mediante ejemplos visuales. Esto permite a los cirujanos volver a encontrar los diferentes instantes capturados durante las intervenciones.

En esta tesis se presenta una extensión y adaptación de Lire para la recuperación de vídeo basado en las características visuales y métodos de late fusion. Los resultados son evaluados desde dos perspectivas: una cuantitativa y una cualitativa. Mientras que la parte cuantitativa sigue el estándar de las prácticas y métricas empleadas en vídeo retrieval, la evaluación cualitativa ha sido basada en un estudio social empírico mediante una interfaz web semi-interactiva. Particularmente, se ha emprendido el método "thinking aloud test" para analizar si nuestra herramienta cumple con las expectativas y necesidades de los usuarios a la hora de utilizar la aplicación.

Debido a la escasez de médicos disponibles para llevar a cabo las pruebas cualitativas, el trabajo se ha dirigido también a un segundo dominio: conjunto de vídeos de acontecimientos musicales. Este tipo de vídeos también ha experimentado un crecimiento exponencial con la llegada de los smart phones y se encuentran al alcance de un público muy amplio. Análogamente a los vídeos endoscópicos, hacer una busca en una gran base de datos de este tipo también es un tema difícil y motivo de estudio.

Key Words

video retrieval, visual descriptors, medical, user study, feature fusion

Acknowledgements

First of all, I would like to acknowledge the ETSETB, for forming me during those hard five years, and to the Alpen-Adria University in Klagenfurt for hosting and mentoring me during those last 6 months.

I would like to express my greatest appreciation to my tutor Professor Mathias Lux for offering to me this great opportunity, for hosting me in this research group, teaching me many things and make me feel comfortable in this city since the first day. Also to my co-advisor Xavier Giró who, from Barcelona, has been willing to solve my doubts, stimulating suggestions and guide me on how to research. Finally, to my colleague and friend Nektarios Anagnostopoulos, whose advice has always been very valuable throughout the days in the laboratory, and who encouraged me not to be afraid of anything and go on.

Last but not least, my thanks goes to my family who has given me the greatest encouragement, and supporting me with every final choice I have ever had, including my departure to Austria and finish my studies. And to my friends, thank you for cheering me up constantly, or showing yourself happy with my achievements.

Agraïments

Primer de tot, m'agradaria donar les gràcies a la ETSETB, per formar-me durant aquests durs cinc anys, i a Alpen-Adria University of Klagenfurt per acollir-me i guiar-me durant aquest últims 6 mesos.

M'agradaria expressar la meva gratitut al meu tutor, el Professor Mathias Lux, per oferir-me aquesta gran oportunitat, per acollir-me en aquest grup de recerca, ensenyar-me moltes coses i fer-me sentir cómoda en aquesta ciutat des del primer dia. També, agraïr al meu company i amic Nektarios Anagnostopoulos, de qui els consells han estat sempre molt valiosos i qui m'ha animat, al llarg dels dies al laboratori, a no tindre por de res i continuar.

Finalment, però no per això menys important, agraïments especials per a la meva familia que m'ha animat i m'ha recolçat amb cada decisió final que he hagut de prendre, incloent el meu viatge a Austria i acabar els meus estudis. I a les meves amigues, que m'han animat constantment i han mostrat una gran felicitat pels meus successos.

Agradecimientos

Primer de todo, me gustaría dar las gracias a la ETSETB, por formarme durante estos duros cinco años, y a la Alpen-Adria University of Klagenfurt por acogerme y guiarme durante estos últimos 6 meses.

Me gustaría expresar mi gratitud a mi tutor, el Professor Mathias Lux, por ofrecerme esta gran oportunidad, por acogerme en este grupo de investigación, por enseñarme muchas cosas y por hacerme sentir cómoda en esta ciudad des del primer día. También, agradecer a mi compañero y amigo Nektarios Anagnostopoulos, de quien los consejos han sido siempre muy valiosos y quien me ha animado, a lo largo de los días en el laboratorio, a no tener miedo a nada y a continuar.

Finalmente, pero no por eso menos importante, agradecer en especial a mi familia que me ha animado y me ha dado soporte con cada decisión final que he hecho, incluyendo mi viaje a Austria y acabar mis estudios. También a mis amigas, que me han animado constantemente y han mostrado una gran felicidad por mis logros.

Contents

List of Figures

Chapter 1

Introduction

New information technologies are increasingly improving our daily life. Thereby, many research efforts try to automatize data processing in several fields where this task has been performed manually. The automatization of these processes provides users with greater efficiency, agility and flexibility.

Furthermore, several professional fields, such as medicine, are applying these new advances in data processing to improve their efficiency and accuracy in their activity. Medical doctors have used the digitization of audiovisual medical content which results in increased information. Through an investigation of image processing can be achieved more effectively and quickness in the daily work of these groups from hundreds video indexing and the criteria of using visual similarity between images that they contain.

1.1 Focus of the Thesis

This thesis focuses on an evaluation investigating whether the application at hand is useful and effective for surgeons, using a quantitative and qualitative study. Nevertheless, this work also adds a second scenario from another video domain to assess the flexibility of the adopted retrieval system.

The project requirements defined at the start were the following:

1. Understand the existing procedure to index videos in the LIRE software, which is oriented to the indexing and retrieval of still images.
2. Explore new visual descriptors for image retrieval.
3. Design and run a quantitative and qualitative study on the endoscopic video domain on both expert (surgeons) and non-expert users to assess

the human-computer interaction for the indexing, search and browsing of endoscopic videos.

4. Extend both the quantitative and qualitative studies to an additional domain to test the flexibility of the solution.

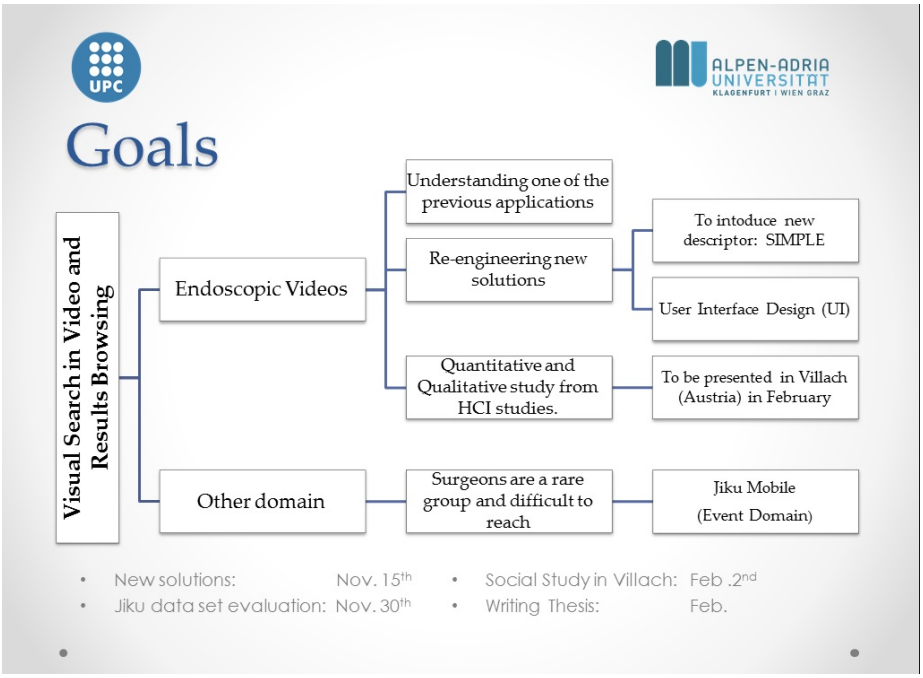

Figure 1.1: Original project planning as presented at the University of Klagenfurt in October 20th

This project was carried out at the University of Klagenfurt, in Austria, in the framework of the European Union Erasmus program for student mobility. This thesis was performed in the framework of project, CODE-MM (Community of Domain Experts in Medical Multimedia) [1]. This scientific project investigates how communities of medical experts can be supported by information systems and innovative multimedia services. The research results are aimed at a real-world scenario, adding automatic annotation and retrieval methods to endoscopic videos. This thesis has been possible thanks to the support and co-operation of the Information Technology department of the Klagenfurt University in Austria. The main initial ideas for this project were provided by my supervisor Professor Mathias Lux, who also provided the software and datasets necessary for the work.

Throughout my research I have been able to collaborate in the writing of two papers submitted and accepted by the 13th International Workshop on Content-Based Indexing (CBMI).

[1] `http://codemm.org/`

1.2 Motivation

Currently, the main tool for documentation of endoscopic procedures are shots from the video, identified and taken during a surgery. With the growth of the large amount of data storage day by day, it is challenging to find the particular scenes taken during the procedure by the surgeons.

In order to re-find easily these shot and summarize better the video content of the surgeries, a graphical user interface is developed, mainly, for covering the surgeons' need to retrieve parts of the videos on demand. This application allows video retrieval based on still image features and a late fusion method. The image-based retrieval is based on comparing a query image with each frame of the indexed videos. The results with the matching frames are presented on the user interface so that the matched frame is presented in the time line of the retrieved videos. One of the central question of this thesis is assessing whether this application is actually useful for surgeons.

In addition, the system is also assessed in the musical performances domain, using an online available data set which was made by means of the personal mobile phone. The larger the event, the larger is the amounts of videos taken there and also, the more videos get shared online. The user study obtained in this domain may also provided insights about the potential of the search engine in the endoscopic domain.

1.3 Outline of the Thesis

The rest of the thesis is structured as follows: Chapter 2 describes the state of the art, focusing on similar systems of the existing demo application. Chapter 3 presents the system requirements that had to be satisfied by this project. Chapter 4 describes the software platform adopted in this project and the different methods and tools of image processing that supports. In the same Chapter a new solution that combines a global scale descriptor with a novel local descriptor: SIMPLE. Chapter 5 describes the semi interactive interface for our tool and the evaluation through a thinking aloud test conducted on five non-surgeon users from the CODE-MM project. In the last section of this chapter, we adapt the new solution and interface to the endoscopic domain and we validate again the tool through a study conducted directly on medical retrieval experts from the same CODE-MM project. Finally, Chapter 6 presents the final conclusions of the study.

1.4 Work Plan of the Thesis

The Gantt chart shown to the Figure 1.2 illustrates the activities of the Thesis conducted over the exchange program in Klagenfurt. The length of each activity

is shown on the horizontal axis throughout the entire semester.

The activities were divided in three larger blocks. First, I got acquainted with the environment of the Thesis and how the previous application worked. In order to get used to the Lire software, I was provided of a new video data set for a better understanding during my training. Thanks to this tool I could introduce a new solution using the novel local descriptor SIMPLE by the end of November. The next steps were preparing the interface for the further musical performance study and check the results with some volunteers. This last activity was done in the middle of January when we could organize the meeting. After getting good results with the previous domain, as a second block the medical domain was again indexed in the demo application and ready for the user test which was done in the middle of March. Finally and for the last block, I collaborated in two papers submitted in the 13th CBMI congress by the end of March.

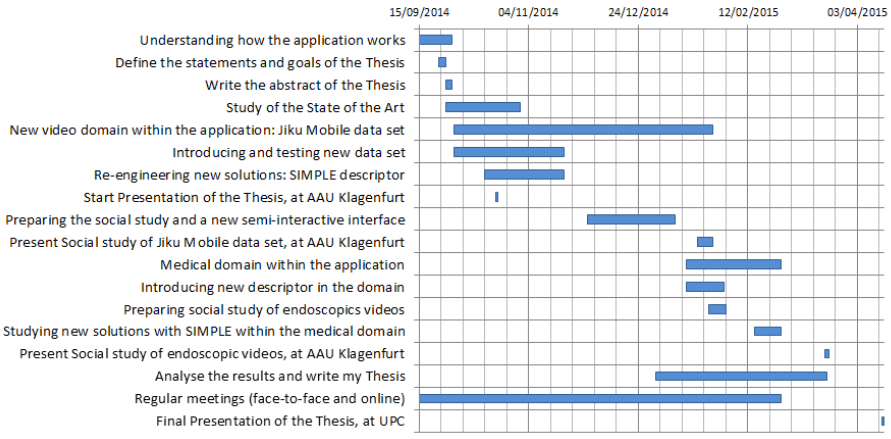

Figure 1.2: Gantt chart. The chart shows the calendar and organization of the Thesis.

Chapter 2

Related Work

In this Chapter we describe the state of the art for visual search in video and results browsing. In order to focus on the two domains covered in this project, we divide this chapter in two different sections. In the introduction chapter, we talked about the endoscopic videos as the main domain of this thesis. This leads to professionals in this field will make use of the final demo application. However, we also discussed to integrate a new domain of this previous work, a dataset of musical performances. We agreed on the final users of this new domain are a wider generic group than the medical one.

In the literature, a wide variety of content-based retrieval methods and systems may be found. Regarding to the user level, we can find a different type of user interface when searching for similar images using a visual query as example. For medical retrieval, CBIR system interfaces are less fancy in design but functionally rich, since it is required to complement the daily work of this professional groups. However, if we focus on a generic group as smart phones users, we found that the user interface design should be more complex and descriptive for easier handling.

The aim of content-based retrieval systems must be to provide maximum support in bridging the semantic gap between the simplicity of available visual features and the richness of the user semantics [10].

A large number of research publications can be found. However, our prototype is different to previous approaches as it incorporates the SIMPLE descriptors, which are local features used for the first time in the field of video retrieval. Most important related initiatives image and video retrieval challenges both in medical imaging and multimedia domain are presented in workshops such as TRECVID [32], MediaEval [19] and ImageCLEF [22].

The first Section is an explanation of existing types of visual search and results browsing on the academic or research interfaces, since the application initially was developed aimed at surgeons in endoscopic field. The goal in this chapter

is to get to know similar systems which have hitherto developed. Some ideas presented in many workshops, as ImageCLEF, will be mining in Section 2.1. It is the main interest for this thesis due to the medical field orientation of this workshop. Then, in the subsequent sections, we will explain other tools which are used.

In the second Section, some interfaces of the second group will be discussed due to the extension of the application domain to social events.

2.1 Medical System Architectures

Content-based image retrieval in the medical domain has been addressed from low-level wavelet-based visual signatures [25] to high level concept detectors [27]. Another way to exploit visual features is to generate automatic text descriptors with computer vision algorithms [18] and use these labels to support text-based queries.

Content-based medical retrieval has greatly benefited from the ImageCLEF [1] benchmark [17], since relevant projects related to medical field were presented there. This workshop has created a strong community of researchers participating in the retrieval of medical images within the *ImageCLEFmed* task. The database for this challenge is a subset of PubMed Central[2] containing in total over 1.5 million images of over 600,000 articles.

Many tasks are presented in ImageCLEFmed workshop and are useful for image retrieval in many reasons. They have purposes on social reasons as education (self, professional, patient), diagnosis, presentations, publications and research in the scientific field. Medical images represent an interesting retrieval domain because annotations are subjective and the context is sensitive.

After one decade of running the ImageCLEF medical tasks, the test collection has grown from 8,000 images in 2004 to over 77,000 images in 2010. The goal of one of these tasks was to classify the images into medical modalities and other images types. The current distribution showing in Figure 2.1 corresponds to that in the PubMed Central data set, much closer to reality than in previous years. As can be seen in the second column, the medical application (endoscopy) of this thesis is considered in this classification.

In user-studies, medical experts have indicated that modality (visual or text) is one of the most important filters that they would like to be able to limit their search by. Using the modality classification, the search results can be improved significantly.

[1] http://www.imageclef.org/
[2] http://www.ncbi.nlm.nih.gov/pmc/

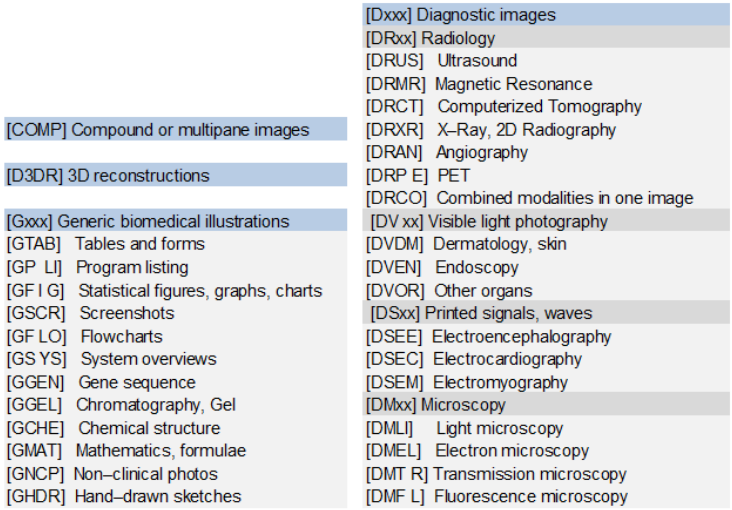

Figure 2.1: Class codes of the modality classification. Source: ImageCLEF @2013

2.1.1 ImageCLEF 2013. AMIA

In the 2013 edition [8], the best textual run achieved the same performance as the best technique using both textual and visual features [9]. Visual-only approaches achieved much worse results than the textual and multimodal techniques. The best visual-based solution [24] was based on the Color and Edge Directivity Descriptor (CEDD), a fuzzy color and texture histogram and a Color Layout Descriptor.

One of the solutions presented in the workshop was NovaMedSearch[3][21], a medical search engine that integrates the two search modalities: text and image. Their goal is to provide an intuitive and simplified way of supporting multimodal queries in medical search. Users can upload their own images to build their query or use existing sample images in their queries. The results are displayed in a ranked list with basic information (e.g. title, keywords, images) and a link to the corresponding article details. The screenshot in Figure 2.2 depicts the search interface which has relevancy on the images and text similarity.

If the search is entirely driven by key words, as in Figure 2.3 the browsing result shows images within many articles related to this search. Once you see the browsing result, you can see the article or search similar images or search similar region. This system allows to consider similarities of a small region within the original image.

[3]http://medical.novasearch.org/

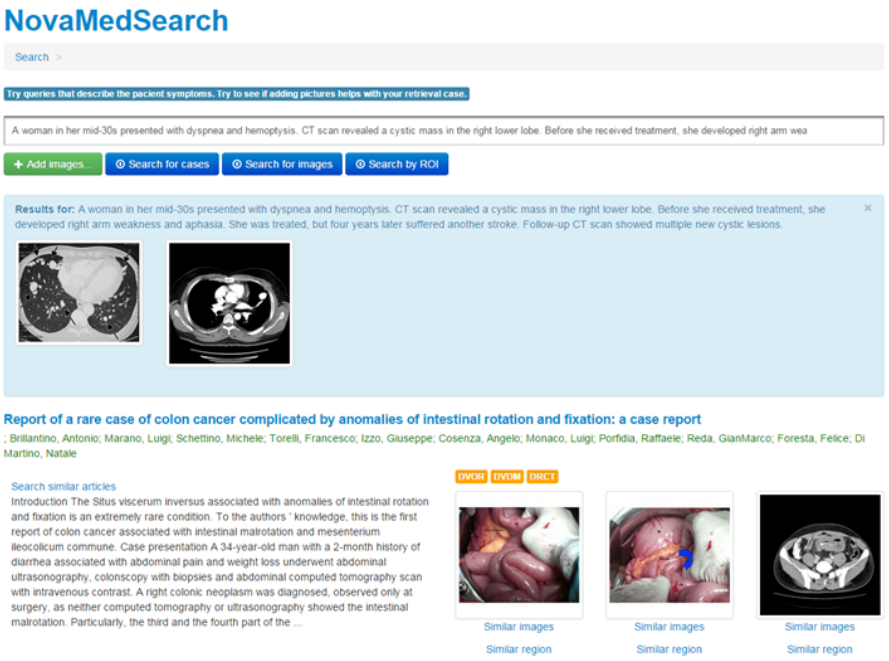

Figure 2.2: Example of a search. Source: NovaSearch System

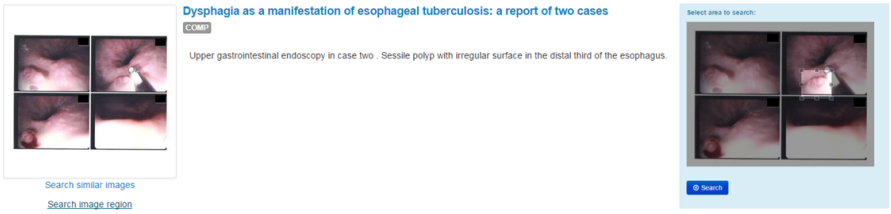

Figure 2.3: Example of a specific search caption. Source: NovaSearch System

2.1.2 ImageCLEF 2014. Liver CT Annotation

The Liver CT Annotation is a task presented in the ImageCLEF Workshop 2014. The main goal of this task is similar this thesis: the medical databases are challenged by the exponential increase in data volumes, but in this case, the project is focused on the radiological domain.

The data of this task was collected as part of the CaReRa project[4] (Case Retrieval in Radiology) using a web-based uploading and manual annotation service depicted in Figure 2.4. This project was a prototype CBCR (Content-Based Case Retrieval) implementation of the CES Platform concept (Clinical Experience Sharing), which focuses on liver cases.

[4]www.vavlab.ee.boun.edu.tr - Research - CaReRa

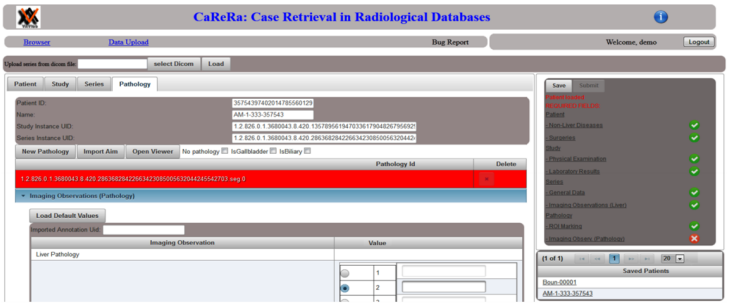

Figure 2.4: CaReRa-Web demo application

There is a desktop application, LIVERworks, to build a CRR query from a given case. The current in-house application targets medical professionals and research groups.

2.1.3 Other Medical System Architectures

Recently, medical retrieval engines have become much more accessible on the web, typically supporting both textual and visual queries. These are the cases of *Goldminer*, *MyPACS* or *Yale Image Finder*.

2.1.3.1 GoldMiner

ARRS GoldMiner [5] [16] helps users find images and articles from peer-reviewed biomedical journals. It is available to all users, but it is intended for health professionals and education. GoldMiner supports medical vocabulary, recognizes abbreviations, synonyms, and kinds of diseases and incorporates standards such as the Medical Subject Heading (MeSH) terms, which are used to index the medical literature in MEDLINE and PubMed. It uses sophisticated techniques from the U.S. National Library of Medicine (part of NIH) to discover medical concepts in free-text figure captions, and uses that information to quickly retrieve relevant images.

Finding images is possible by means of filter the medical classification and some specific words about the disease's characteristics that we expect to see in the result browsing. In this case, a simple search of endoscopic images of adults patients is shown in Figure 2.5. Every result is accompanied with a text annotation of the source of this image and a brief description. If we click on the small "thumbnail" image, a new window appears with the full-size image on the original journal's web site.

[5] `http://goldminer.arrs.org`

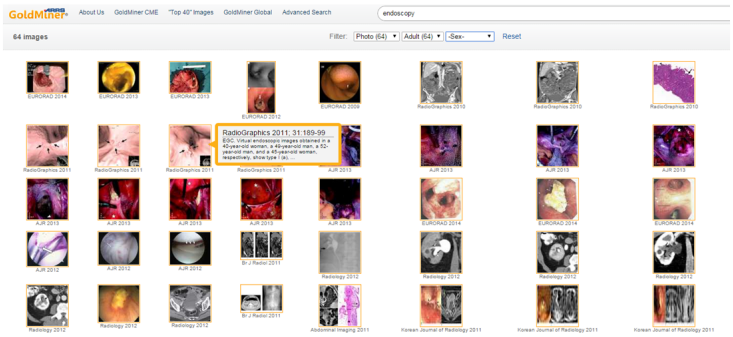

Figure 2.5: Example of a search in application. Source: GoldMiner

2.1.3.2 MyPacs

MyPACS.net [6] is a free service offered to the international radiology community by McKesson Medical Imaging Group as a result of the McKesson acquisition of Vivalog LLC in May 2008. The key to managing this data is having metadata, object schemes, classification systems, and workflow models that accurately reflect the structure of the biomedical expert's domain requirements. Ideally, the domain experts should be able to specify the structure of their domain themselves. We can do the search by means of some filters or writing specific words. The browsing result is shown in Figure 2.6, linking the image with the original source publication

Figure 2.6: Example of a search in application. Source: MyPacs

2.1.3.3 Yale Image Finder

Yale Image Finder [7] can search the actual image content of over 1.5 million Open Access images and figures from PubMed Central by means of search keywords. Yale Image Finder (YIF) is a publicly accessible search engine featuring

[6] `http://www.mypacs.net`
[7] `http://krauthammerlab.med.yale.edu/imagefinder/`

a new way of retrieving biomedical images and associated papers based on the text carried inside the images. Image queries can also be issued against the image caption, as well as words in the associated paper abstract and title. The browsing results are presented in the form of thumbnails with these characteristics. We can click on one of the found images and a new Tab is shown with the original one, the link to the original paper and two kind of images: those that appear in the same paper, and those from other papers with similar image content.

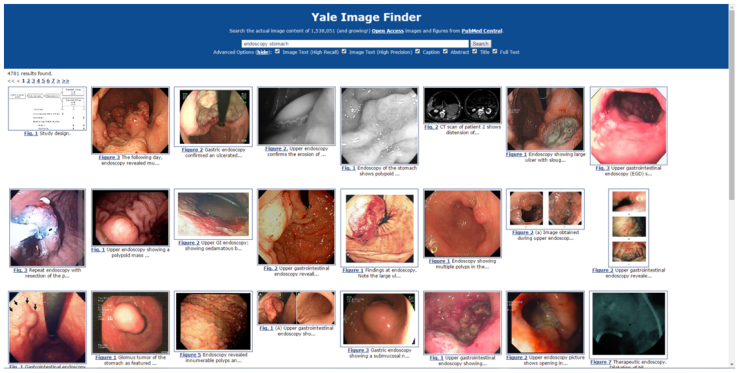

Figure 2.7: Example of a search in application. Source: Yale Image Finder

2.1.4 Content-Based Video Retrieval (CBVR) systems

In contrast to most works on medical CBIR, we address the problem of video retrieval, instead of still images. This venue has been previously explored in the literature. Specifically for real medical videos, [12] proposes a framework that uses principal video shots for video content representation and feature extraction. The classification is mainly implemented by elementary semantic medical concepts, such as "Traumatic surgery" or "Diagnosis". Moreover, [26] presents a framework to retrieve short videos in real time by modeling the motion content with a polynomial model. The system has been successfully applied to automated recognition of retinal surgery steps on a 69-video data set.

2.2 Multimedia in Social Events Architectures

Many people like to share their experiences with friends. A large part of them uses the internet to publish and send pictures and videos from what they have seen, visited and experienced. YouTube alone currently has more than 300 hours worth of videos uploaded every minute[8] and it is hard to keep track on which videos show what.

[8] `https://www.youtube.com/yt/press/`

Our application presents visual search in those videos for near duplicate frames. With such a prototype one can search for visually similar video frames throughout a collection of videos and eventually find those that have been taken from the same scene. For input our system relies on a video frame or an image.

In the literature, a wide variety of content-based retrieval of images and videos from mobile devices may be found. This is an interesting field which is continuously on research. While the tasks of the initiatives are changing, near duplicate frame search either has been a task or it has been used as means to an end for tackling one of the tasks.

In [29] the authors present a system also focusing on videos taken at events. However, they employ a more controlled and holistic approach. Videos recorded with their software are automatically enriched with meta data, ie. sensor readings, which allows for faster and easier retrieval, while we do not restrict the video recording procedure and operate on visual data only. In [23] the authors present a system, which automatically creates an event summary based on different videos from different users and view points. The system, called Jiku Director, operates on the same data set as our prototype does, but relies solely on meta data.

It is a Web-based application that the main contribution is the creation of the summary from event videos uploaded by users. The application uses an algorithm that considers view quality, video quality such blockiness or illumination, and spatial-temporal diversity in order to create this summary. The system, in contrast to our application, is not focused on the retrieval of scenes.

A similar case using a large scale dataset is presented in [1]. The dataset contains of 3,800 hours of newscasts and features 200 queries for retrieval evaluation providing a ground truth. The queries are images and have to be found in the video streams, an approach the authors call image-to-video, I2V. Moreover, the authors present a system operating on the data set in [2].

Chapter 3

Requirements

As we have seen in the previous chapter, video retrieval systems can serve different purposes; they can be developed for specific types of content or different types of users. This chapter we analyse the requirements from the users' perspective.

The goal of our application is to test different visual features as well as various late fusion methods on our use case of re-finding shots within video streams. Using the LIRE software library [20], we can integrate up to 20 different visual features. This application is able to generate automatically a ranked list of videos according to the similarity of their frames to a query image selected by the user. So, we focus solely on visual information, and we report on experiments based on a research data set.

The main difference between our prototype and the related work in Chapter 2 is that additional metadata is not added. Moreover, our system is also different to previous approaches as it incorporates the SIMPLE descriptors, local features, novel in the field of video retrieval.

3.1 Content requirements

One requirement of our application is that the user must be able to quickly find the relevant images referred to the selected query image thanks to the visual features extracted and indexed by the tool.

- Which content must be indexed to correctly assess the search engine ?

We need to extract images from the videos of the data set with a certain sampling rate in order to create a visual index. For the musical performance

dataset, we focus particularly in the different performances within the several videos, different group of people dancing, playing or singing on the stage, outdoor or indoor scenario, point of view... For the medical domain, we focus on the time period of the stream videos inside the patient; the visual information from outside the patient is not useful for this retrieval. The relevant shots are identified and taken during the surgery by the doctors, since for making a difference between a normal or a query image in this domain, knowledge in this field is needed.

- How can we evaluate the visual results taking into consideration the subjective point of view of each user?

The approach should be validated by verifying whether its results fulfill the original user requirements. Our system must be evaluated by participants through an interactive web-interface based on some videos and queries selected from the whole data set. By means of *thinking aloud tests* performed with volunteers users, we record the movements, voice and opinions of the different participants and in this way we are able to evaluate the results of our final visual research.

Chapter 4

Developed solution

This chapter presents the different configuration that have been assessed in this thesis. Global scale descriptors and fusion methods are presented in Section 4.1, while a novel solution combining global features with a novel local scale descriptor is presented in Section 4.2. Both the methods are later tested in Chapter 5.

4.1 Fusion of existing global descriptors

The default LIRE software library [20] used in this work implements up to 20 different visual descriptors for visual indexing. LIRE presents a modular architecture which allows easily adding existing and new descriptors.

The first configuration that was assessed combined three different visual descriptors at a global spatial scale:

Color and edge directivity descriptor (CEDD) [7]: a compact joint histogram of fuzzy color and texture

Auto color correlogram [14]: a color feature that measure how often a color encounters itself in a neighborhood

Pyramid histogram of oriented gradients (PHOG) [6]: a fuzzy gradient histogram organized in a spatial pyramid.

Each descriptor can be considered as an Independent Retrieval Model (IRM) [11] which at some point needs to be fused.

Mainly, two types of fusion schemes can be found in the literature: Early Fusion and Late Fusion [33]. In early fusion the retrieval models are integrated

from the start and afterwards a multimodal representation is learned. Late fusion approaches on the other hand infer similarity directly from unimodal features and integrate results at the end.

In our approach, we employ a late fusion model based on multiple visual global features using a single visual example as a query image. The detailed process of late fusion can be described with the following steps:

1. The query image is compared to the indexed items for each descriptor.
2. An independent ranked list from each IRM in the system is generated, with an associated rank and score to each retrieved item.
3. The objective of late fusion techniques is the combination and re-score or re-rank of the initial result lists into one final list.
4. Apply a normalization strategy, whether based on ranked or on scores. It is required to truncate the initial lists to the top N results and normalize them. In this thesis, two options were considered:
   - Normalization by rank:
   $$\bar{R}_k(n) = \frac{N + 1 - R_k(n)}{N}$$
   - Normalization by score:
   $$\bar{R}_k(n) = \frac{R_k(n) - min(R_k)}{max(R_k) - min(R_k)}$$

   where Rk is the initial result list (rank or score) from the retrieval model k.
5. Aggregate the normalized ranks or scores to generate a single ranked list. In this thesis, two options were considered:
   - *Sum*: $R_t(n) = \sum_k (R_k(n)) = R_1(n) + R_2(n) + ... + R_K(n)$
   - *Sum with combMNZ*: Sum x number of IRM that returned image n where $R_k = 1, ..., K(n)$ are the initial results for image n, K is the number of IRM involved in the system and $R_t(n)$ is the final list for image n.

As a result, the two options for normalization and fusion define four different configurations for late fusion: (i) sum of ranks, (ii) sum of scores, (iii) combMNZ of ranks and (iv) combMNZ of scores.

Fig.4.2 shows the overall architecture of the methodology employed using late fusion methods with global features.

The final ranked list shows the similarities of the query image within the videos of our data set. The objective of the final ranked list is to present a list of videos sorted depend on the items on that list.

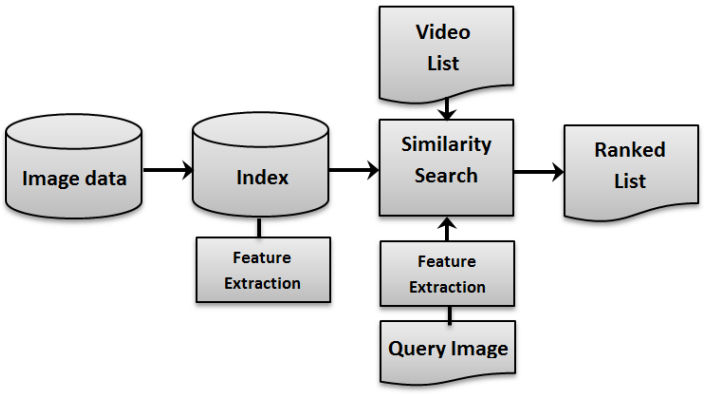

Figure 4.1: CBIR search process which operates on an index previously created

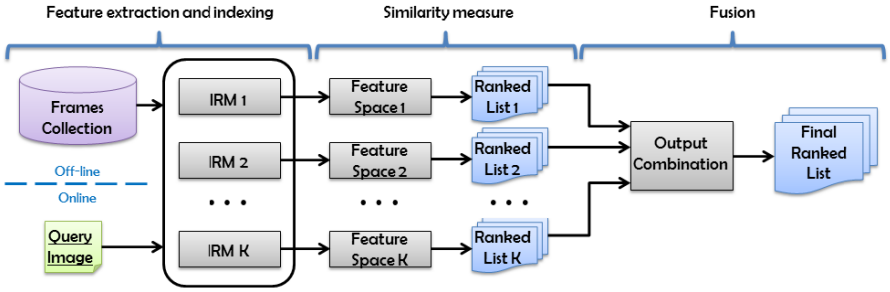

Figure 4.2: Application of late fusion in our approach, illustration is based on the work in [24], [25], [32].

The videos are sorted depend on the position where the images are appearing in the final list (Fig.4.3).

4.2 Introduction of local SIMPLE features

This project also explores a novel solution in LIRE: adding a local scale descriptor to be compared with the global ones.

We adopted a localized version of CEDD using the SIMPLE descriptor (*Searching Images with MPEG-7* (& MPEG-7-like) *Powered Localized dEscriptors*) [15], which has outperformed classical local features in some scenarios. One of the contribution of this thesis is assessing the potential of this descriptor with respect to more classic global scale solutions.

The implentation of SIMPLE we have used in our study is the combination of SURF detector [4] and MPEG-7-like CEDD. Opposite to the first solution of global features using late fusion, in this implementation the indexer is made by

	image List
1	video_A_image_005
2	video_C_image_002
3	video_A_image_001
4	video_A_image_002
5	video_C_image_003
6	video_B_image_006
7	video_E_image_003
8	video_D_image_002
9	video_C_image_001
10	video_E_image_005

	video List	*number of images per video*
1	video_A	*3*
2	video_C	*3*
3	video_B	*1*
4	video_E	*2*
5	video_D	*1*

Figure 4.3: Example of the videos‘ ranking

the extraction of relevant key points using SURF detector and the extraction of the global features using only one descriptor, CEDD. This descriptor uses 24 bins of different colors. 7 different color and 3 sets for each one, plus white, black and grey colors. The detector SURF finds interest points in the image using Hessian matrices of the key points, determining their orientation, and using Haar wavelets in a square region around these points to find intensity gradients.

Following that, for indexing and retrieval, this method uses the *Bag-Of-Visual-Words* (BOVW) model [31] to extract the local features and aggregate them into local features histograms. In our case, and the experiments were reported based on a visual vocabulary of 512 words build with the k-means clustering algorithm.

Chapter 5

Evaluation

In this chapter we evaluate our visual search for musical performance and endoscopic videos. First, we perform the study using the event of Jiku Mobile dataset, as presented in Section 5.1. Afterward, we moved to a dataset of medical videos, presented in Section 5.2. We take a endoscopic video dataset for this study in order to test and verify the goals of the main idea in this project.

The user study presented in this chapter assesses evaluates the fusion methods and global descriptors implemented in LIRE for the video retrieval in both domains. Moreover, we try to verify if the SIMPLE descriptors proposed in Section 4.2 works better in the browsing results depending on the user's needs.

5.1 Application to the social event domain: User-generated videos of musical performances

A first goal in this project was developing a preliminary study in a domain different from the medical one. This domain should not require the assessment of expert surgeons who may be difficult to access and, this way, facilitate the tool development before final tests with medical experts. During the process of testing the previous methods and the new solutions we used a data set made by user-generated videos from musical performances.

Therefore, we used the Jiku Mobile dataset [28] for our study, a collection of 473 video clips taken at five different social events. The videos were recorded by different people from different angles using mobile devices. We must keep in mind that those videos are recorded in low quality and they feature pairwise overlap time- and scene-wise. Each event contains several performances. For our experiments we indexed 356 randomly selected videos based on equidistant

frames, using one frame per second through *ffmpeg*[1] tool. All in all, the demo application uses 88,487 images from those videos.

A set of 412 queries of different performances (cp. Figure 5.1) in the musical performances was created manually. We aimed to cover different aspects, like for instance, outdoor, indoor scenes, colorful, and simple scenes. Manually we obtained the query frames inside each video by means of screenshots using the software VLC player. There is one key frame of each performance.

Figure 5.1: Two query images extracted from one video of the social event.

For testing, we split the events in different folders called "EVENTyymmdd" where are indicated the day of the performance. Depending on the event, we can find 50, 80 or 66 videos inside. The goal is to find similar images that show us the same instant of the performance of the different videos capture by all mobile devices. The procedure is the same as explained in the previous sections.

The results are presented in a visual form in HTML5 [13] and can be viewed in a recent version of common browsers. For each query an HTML file is generated displaying the query image, the list of similar images that the demo application finds, and the videos where both the query image and the rest of the frames belong to. All of the items appear along a timeline where the images were taken. The screenshot presented in Fig.5.2 shows the results of a shot query. Instead of showing the image results, only their positions in the video are indicated in the time line. Due to the nature of similarity search results look very much like the query, so showing them would not help the user in re-finding them in the video streams.

Based on the top 40 hits for truncate the list of each query, we automatically determined the four best matching videos and present them to the user, highlighting the time location where the matching frames have been actually found, as shown in Fig.5.2. As the search process is based on frames within the videos and the result list is also composed of video frames, our system aggregates the frames as a last step. For this reason, the final ranked list of videos is based on their best matching frame, ie. the most similar frame defines the best matching video, the next most similar frame of a different video defines the second best matching video, etc.

[1] `https://www.ffmpeg.org/`

Figure 5.2: Jiku Mobile data set. Screen shot of the application showing a query and the first four results

To sum up, in order to get the online results, we had to convert all the video data set from MPEG-4 [30] to WebM [3] video formats and compress them to 640x360 by means of *ffmpeg*[2].

5.1.1 Quantitative study

Based on the event database explained above, we present the results from the four late fusion methods using global features as descriptors and the results from the new solution by local features using SIMPLE. Based on the 412 queries, we created a benchmarking data set. We tested if all the queries are to be found within the video data set. Our tests have shown that the video from which the query frame was extracted was ranked at the first position for 96% of the cases

[2] `https://www.ffmpeg.org/`

(cp. Fig. 5.3). Therefore, this quantitative study confirms that the subsampling of one frame per second is enough for the data set to yield meaningful and accurate results with our approach.

	Late Fusion using Global Features				Local Features
	Sum of Ranks	Sum of Scores	combMNZ of Ranks	combMNZ of Scores	SIMPLE-CEDD
Video 1	397	398	397	399	377
Video 2	5	4	5	4	9
Video 3	1	1	1	0	2
Video 4	1	2	2	2	3

Figure 5.3: Results of the tests of the social event domain on where that actual video can be found in the results. The first four columns give the four different tested feature fusion approaches, the fifth one gives the results on the use of the SIMPLE-CEDD descriptors.

As we can see in the table, using local features we obtain mostly the same number of query image found in the first video on the top of the list. Without implementing late fusion methods, we obtain with SIMPLE-CEDD features a good approach similar to the previous methods.

5.1.2 Qualitative user study

5.1.2.1 Evaluation Method

The qualitative study considered each of the five different music events separately. In each event, the user can find several performances. In order to do the study, we prepared three configurations of the search engine. Due to the similar results between the four late fusion methods we employ, we want to show to the users only two of them contain two different fusion methods and the search by means of the global features extraction. The third one is developed by means of the new solution with the Simple descriptor. In this way, we will be able to know their opinions between the two solutions based on global and local features, respectively.

We tested which of the three search engines satisfies the user's needs, and which of them gives better results. All the opinions regarding to the results which are similar to the query image, narrowed down to the scenario or performances and wrong approaches, are valuable for our evaluation test.

This application was not available for online search, but we precomputed 5 results and asked users to choose depending on the user's preferences. From those 5 events, the user could select keyframes as queries and try among the

three search engines for each query chosen. The task of the user was to use the search engine, by picking one query for each of the events and browse the results. Finally, user reported their opinion on the obtained results.

The methodology that we employed was a *thinking aloud* test, as described in [5]. Thinking-aloud protocols involve participants thinking aloud as they are performing a set of specified tasks. Users are asked to say whatever they are looking at, thinking, doing, and feeling as they go about their task. This enables the observers to see first-hand the process of task completion. The observers objectively take notes of everything that user's say, without attempting to interpret their actions and words. Test sessions are audio- and video-recorded so that we can go back and refer to what participants did and how they reacted. The purpose of this method is to make explicit what is implicitly present in subjects who are able to perform a specific task.

The evaluation session consisted of two parts; part one being a hands-on experience by different participants who used our tool. Part two is an open interview reflecting his experience with the tool; we ask the volunteers during the interview what they think about the tool and which conclusions they extract from this test. Is it a useful tool? Does this tool cover their expectations?

5.1.2.2 Evaluation Procedure

We created a semi-interactive web page using some images as queries for each event (cp. Fig. 5.4 and 5.5). From these pages users could access to each event and find the different key frames, choose one and explore the results. We created three search approaches which were blindly labeled as A,B or C. For the first one we implemented two search engines, A and B, which belong to the results of the indexing with global features. The difference between them is the Late Fusion method implemented. We assess (i) sum of ranks and (ii) sum of scores. On the other hand, we develop a third search engine, which is implemented by the new solution using the index of SIMPLE based local features. The final results are shown in new tabs to let users compare among these three search engines. This way, we avoided any bias of the subjects towards any of the three approaches.

The thinking aloud tests let us know the participants' opinions and which method works better according to their needs and expectations. At the end of the recorded interviews, the volunteers express their opinion about the browsing result and whether it covers the main visual search they expect.

After the user finished the test, we could compare their reasoning and the results with our hypotheses.

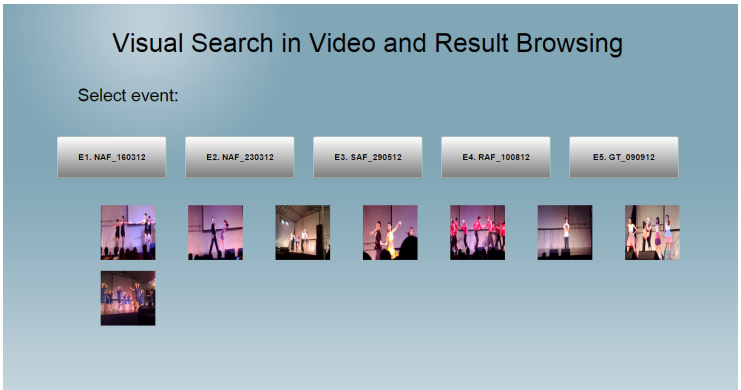

Figure 5.4: Main window of the demo application.

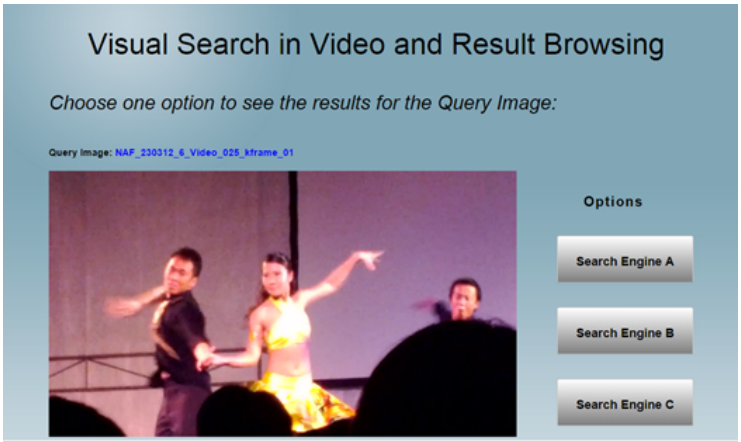

Figure 5.5: Second window of the demo application.

5.1.2.3 Hypotheses

To evaluate the effectiveness and quality of our proposed visual search in video and browsing result developed, the algorithm needs to be tested. We defined some hypotheses that we wanted to verify in this test.

- Hypothesis 1

  The results from the previous application made by the extraction of different global features show a browsing result not as specific as the results with the new solution developed by the SIMPLE features. As commented in Chapter 4, for one selected query which belongs to one performance, we find similar images within the same performances in different videos by means of SIMPLE features. However, by means of global descriptor we find similarities in the scenario of the query, even taking different performances within the same event.

  We notice that for one specific search as for example one specific per-

formance, we can find different videos from different perspectives of this one by means of the new solution using SIMPLE. On the other hand, if the user needs are to see many performances from only one perspective, the extraction with global features covers these expectations. It could be useful for record and re-view the whole event, for example.

- Hypothesis 2

  Search engines A and B do not show many differences between them and always find similar images and show the results. On the other hand, with search engine C (made by SIMPLE features) mostly shows only similarities in the results if it finds the specific search within the videos. That means that if in the dataset only exist one video which has recorded this specific performance, the result only show this video, even if we ask to the demo application to see four of them.

- Hypothesis 3

  Users are able to recognize the performances within the videos.

5.1.2.4 Participants

The tests for the musical performances domain were run with 6 researchers within the Information Technology department in the University of Klagenfurt. Four of them belonged to the CODE-MM project. We expected more specific answers from them about the results due to their daily work on image processing and a following opinion regarded to the medical domain. They were also able to suggest whether the results of the previous application or the new solutions would fit with the endoscopic video dataset.

For the hands-on experience, we indicated the methodology of the participants the *thinking aloud test* while they were provided of a Personal Computer running our demo application tool. The sessions were recorded with one video camera over their shoulders capturing his mechanical interaction with the tool. The camera captures the whole scene of the screen including the voice (cp. Figure 5.6.

5.1.2.5 Results

Most users chose as queries the four images presented in Figure 6.1. As a general overview, we noticed the users were expecting to see similar scenes or the same performances in the results of the search. All of them appreciated the similarities of the background, the scenario or the number of people which are shown in the results. However, the main expectation was to find the similar musical performance from different points of view.

The participants chose the query image based on their intuition of what would result interesting, ie. they were driven by their own curiosity. There are many

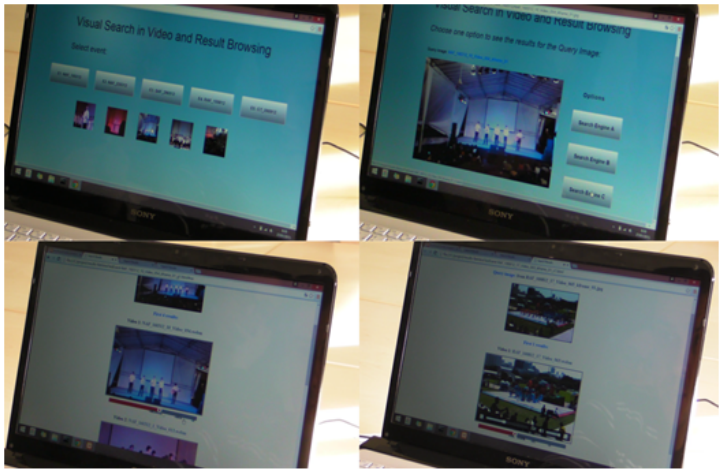

Figure 5.6: Screenshots of the different movements from the user's test 1.

Figure 5.7: Most used query images in the user test (left to right).

reasons as, for example, the simplicity of the scenario with specific people on it, as in the top left image in Figure 6.1. Colourful scenario outdoor crowded of people as in the top right image could be another factor. And, finally, other reasons were the shocking background colour or a specific show with shocking movements as in the bottom right picture.

After choosing a query image, some users were expecting to see directly videos showing the performance. They realized later that the results were shown in the timeline. They considered the timeline as a great tool to use in the demo application, because it allows the user to go directly to the final results without the need to watch the whole video. They did not appreciate too many differences between some search engines, so to have the results in different tabs was useful in order to check better the differences and similarities to compare the results.

Participants were also searching for other query images as the six images presented in Figure 6.2, in order to see more results and being able to express an overall of the search and browsing result. Users chose many pictures which are similar to the first query images. As an example, one of the key frames shows again a unique person on the stage but, this time, without this uniform red colour than the last selected query. And, finally, users were interested in seeing the results from a simple query image as for example, a piano on the scene or a couple dancing. Users investigated other examples about outdoor images, two of them plenty of colour, recorded one in the night and another in the day, and another isolated one to make a comparison between them.

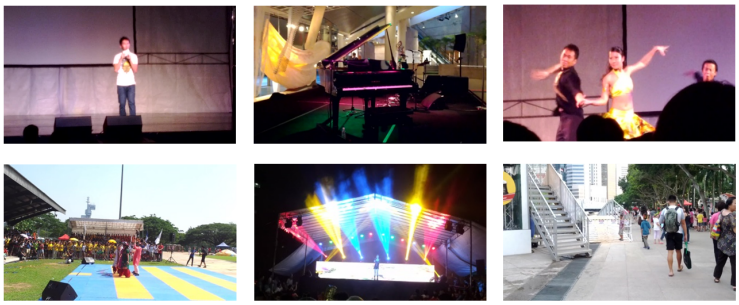

Figure 5.8: Other used query images in the user test (left to right).

Overall impression:

Search Engine A and B:

These two search engines show an abstract view from the whole event. Users can see different performances within the same event. Even so, the participants can notice similarities among all the hits of the results. They show similarities in the colour from the background, among the number of people who appears in the scene, among the view point from where the performances were recorded... All of the users have some knowledge in the image processing and regarding to their opinion, it could be possible due to the features used. If there is similar surrounding and similar building in the background, you get other results of the same event. It depends of what are you looking for. Search engine A and B are good approaches when users expect the same background or overview of the event, in an exploratory search mode with a query as an example.

Search Engine C:

All the users agree this is the search engine which fits better if the user is searching for similar content. Users realized especially for the first key frame they chose with 5 people in the stage. This result always showed the same performance with different viewpoints and it is a great result for the user. As we commented above, users expect as a main result to see the same performance as the query image shows. So, in this case, this engine covers better the user's

needs. The participants realize that, in this case, they can see only the hits they were looking for. If there are fewer hits, the results are less confusing for the user. The third search engine works better because it reduces the number of steps the user has to do until reach the right moment, the moment where the same performance of the query image shows. On the other hand, there are videos which show better results in earlier positions. As users, they are expecting to see the results quickly and not waiting until the last video to see more good results.

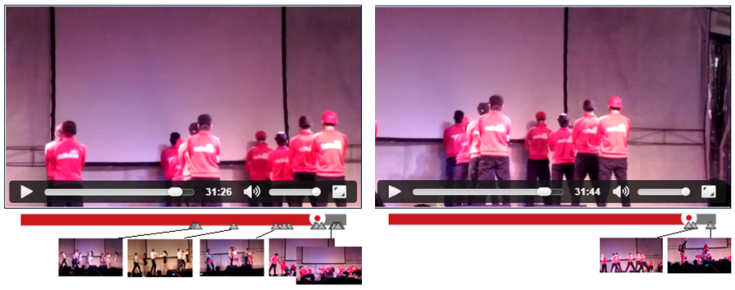

Figure 5.9: Results of a specific query image for both global and local features solutions.

5.2 Application to the medical domain: Endoscopic videos

Our application was also tested in the medical domain for covering the main goal of this thesis. The development was done on a data set made by surgeons in Austria of 1,276 video clips, and it covers roughly 33 hours of anonymized video data of 54 laparoscopy procedures. Due to the size of the video archive which has been used in HQ, a temporal subsampling was needed as in the previous Section using the same tools. In this case, for the indexing we extracted five frames per second, all in all 593,446 frames. These images are sorted in this case in a list of 10 results and the demo application shows the similar images as a result from the first three videos in that list. Average search time for combining the three retrieval models mentioned above was 30 seconds.

In order to define the experiments, we created a test dataset of query images. For this purpose, we used the shots generated by the surgeons during real procedures whenever they wanted to document a specific event that they consider important in the course of the surgery. This way we exploited the interaction from experts in endoscopic videos to determine the highly informative frames in the video, assuming that given the original intention queries in a retrieval system would be from a similar nature. Notice that, as a result, our set of queries is a new group of images different from the uniformly sampled frames from the video dataset. Even more so, as the shots are taken from the live and not the recorded video, we assume that some of them are not even in the recorded clips. Using experts, we cleaned out the query set aiming to remove

stills that do not reflect a recorded video frame, ie. out-of-patient shots, survey shots, etc., resulting in 600 queries.

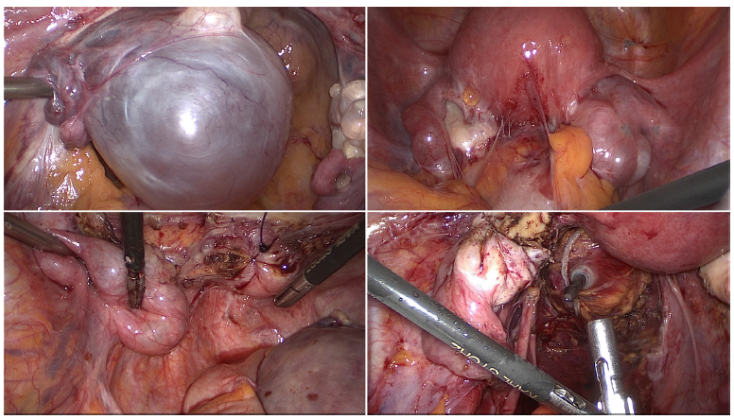

Figure 5.10: Shots (photos) manually created from the surgeon in the course of the procedure.

The relation of photos taken by a surgeon to the actual video streams is depicted in Fig.5.10. These photos are merely frames (still images) that have been saved at the time of operation on request of the surgeon, so they are also part of the video stream itself. Most important, what distinguishes them from the other frames of the video is that the surgeon intentionally directed the camera to a view to capture an optimal picture for later reference.

The screenshot presented in Fig.5.11 shows the results presented in in a visual form in HTML5 of a shot query with their position indicated in the time line.

On the other hand, in a non-visual way, the application shows in which video we find within the query image. For the different methods chosen, we check, in different text files, how many queries of the whole set appear in the top positions of the video list. We find a good approach when the video that contains the query image is shown on first position in the top of the list.

Fig.5.13 shows, using the same procedure of the previous domain, the three best matching videos, in this case, and present them to the user. The list is based on the top 10 hits for each query.

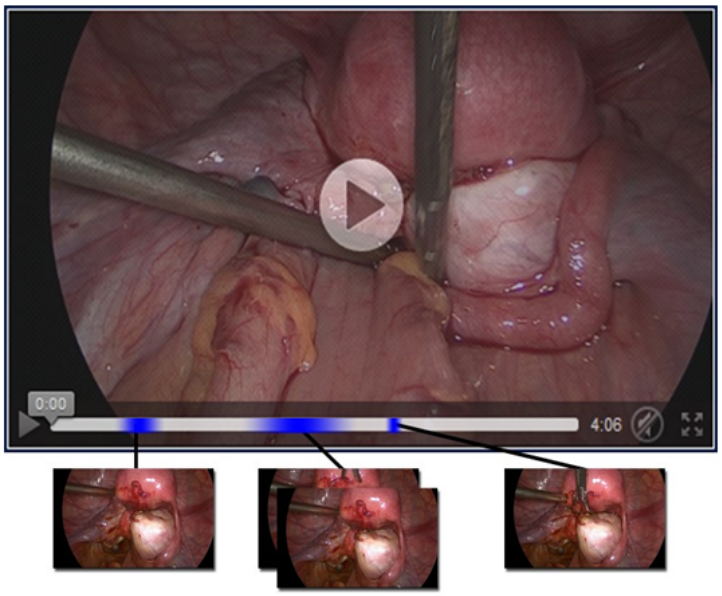

Figure 5.11: All results are presented in HTML5 and can be viewed in a recent version of common browsers.

5.2.1 Quantitative study

This thesis extends a previous work where the first results of the late fusion methods and the global descriptor were presented. We implement the same new approach used in the previous domain in order verify whether it works better to the medical field.

Due to the inside of the subject in the surgeries look pretty similar, with similar background color and textures, all the frames in the whole data set are near in image content. This fact brings us to find new solutions which work better for our visual search of our demo application. In this chapter, we present the new implemented solution with SIMPLE features, due to the success in the musical performance domain studied above.

Based on the medical database of endoscopic videos introduced in above and keeping the format in the methodology of the previous domain, we present the results from the four late fusion method using global features as descriptors and the results from the new solution by local features using SIMPLE. We created the same benchmarking data set. Based on the whole set of queries, our tests have shown that for 470 out of 600 (78.3%) of the queries, the source video was at the first position of the result list. In 84.2% of the queries the source video was among the top three positions for the sum of ranks approach and the combMNZ of ranks, a very similar figure was obtained also for the sum of scores and the combMNZ of scores. Local SIMPLE descriptor led to slightly better results, as in 79.8% of the queries the source video was in the first place, while

in 84.6% of the queries the matching video was the first three videos (cp. Figure 5.12).

	Late Fusion using Global Features				Local Features
	Sum of Ranks	Sum of Scores	combMNZ of Ranks	combMNZ of Scores	SIMPLE-CEDD
Video 1	470	471	471	471	479
Video 2	21	20	20	20	21
Video 3	14	15	14	15	8

Figure 5.12: Results of the tests of the medical domain on where that actual video can be found in the results. The first four columns give the four different tested feature fusion approaches, the fifth one gives the results on the use of the SIMPLE-CEDD descriptors.

This indicates that the subsampling of five frames per second is enough for the used dataset to yield meaningful results. Note at that point that the shots are not necessarily in the video frames as they were taken from the live videos, so the ground truth at hand is more on a semantic level than mimicking a near duplicate task. Taking in account the similarities between all the frames in the medical data set and the good approach of the late fusion methods using global descriptors, the new solution using SIMPLE features improves even more the solutions in our application.

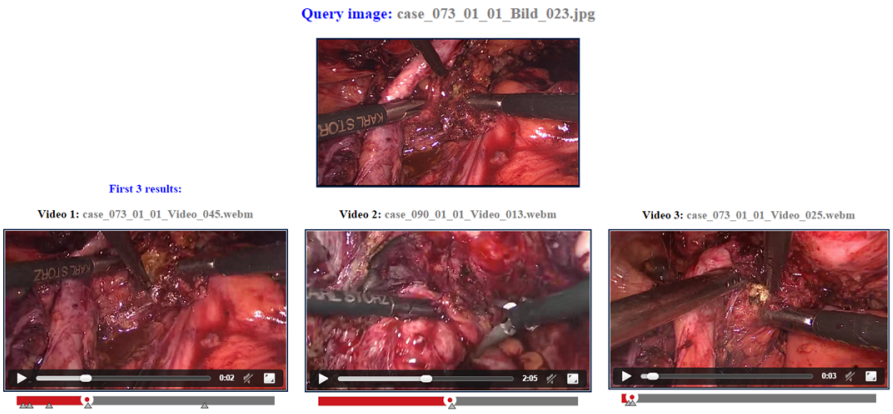

Figure 5.13: Screenshots of the result presentation showing the three top videos and the query image. All results can be viewed in recent browsers supporting HTML5 videos and JavaScript. Best matching frames are indicated by triangles in the red and gray time line below the video player.

5.2.2 Qualitative user study

5.2.2.1 Evaluation Method

The main idea is the following. It is a dataset of different surgeries' cases. For each case, we have many videos, due to the long duration of the surgery. However, we treat all the videos of each case as an only one set. We expect to find similarities in the images within the different cases of our query image. However, we treat as a good approach to find, in the first position, the video where the specific query image belongs.

The methodology in this case is the same we followed in the Jiku Mobile data set. The user test will be done by means of a thinking aloud test [37]. The procedure had the same characteristics than the previous one in Section 5.1. The task of the user is now to try and use this search engine pick some endoscopic queries and investigate the retrieved results.

5.2.2.2 Evaluation Procedure

For testing, we use the same interface used in the previous dataset with three different search engines (cp. Figure 5.14). However, it is modified in order to introduce the different surgeries' cases and the new implementation of the SIMPLE descriptor. In this case, we consider ten different surgeries cases. For each of them, the user has the possibility to choose between 4 different query shots available for search. The three search approaches are blindly labeled as search engine A (for sum of ranks fusion of global features), search engine B (for sum of scores fusion of global features) and search engine C (for the use of SIMPLE based local features). This was we avoided any bias of the subjects towards any of the three approaches.

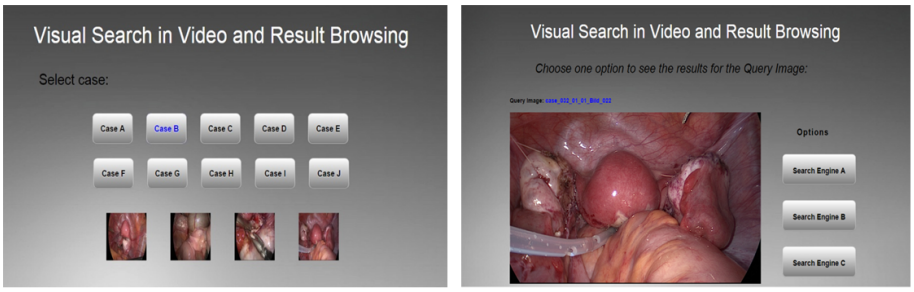

Figure 5.14: Main and second window for the medical semi-interactive interface.

5.2.2.3 Participants

At the beginning of the thesis, one of the goals was to test the application directly with surgeons in Villach, Austria. Due to the professionals in this field are a difficult user group to reach, we had only the possibility to do the user test with some of the same researchers who also did the user test above using the musical performances dataset. These people work within the Information Technology department in the University of Klagenfurt, especially for the CODE-MM project. They have got a good background on the medical domain by working many years in video processing and object recognition and tracking in this domain. They are aware what surgeons are searching for with this application, for this reason, their participation in the test have a high value for our evaluation.

5.2.2.4 Results

In the second experiment – the thinking aloud test – users in general expected to see the same background in several shots within the videos, which are similar to the query image. The participants choose the query image based on their intuition of what would result interesting, ie. they were driven by their own curiosity. They were driven by many reasons, as for example the simplicity of the background with specific organs on it, or specific movements of the surgeons as for instance cut tissue. Other reasons are a specific background, ie. bloody or damaged tissue, or a specific event using different instruments, which lets the user relate to a specific part of the procedure. Based on the overall state of tissue seen in the scene, ie. if it has been cut or cauterized, users know a rough time point within the surgery from the video. It gives them an orientation about the specific moment of the intervention, ie. they know whether the video is from in the beginning, during or the end of the procedure. After choosing a query image, the participants were expecting to see directly videos showing similar interventions.

Due to the length of the videos, users consider an useful tool in the application when the results are marked in the time line; it allows them to find the right moment without the need to watch the whole video.

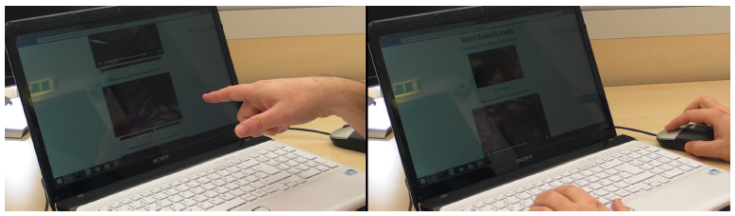

Figure 5.15: Still frames from the thinking aloud test recordings. Test participants pointed out and explained the utility of particular results.

As an overall impression, for the search engines A and B, user commented

they are good approaches showing in the top results the most relevant shots within the videos. However, in many cases the videos with higher ranks in the results show content which is semantically dissimilar by for instance featuring a different organ, instrument or background. For search engine C, which is based on the SIMPLE local features, users agreed it is the search engine that fits better when searching for semantically similar content. This technique also tends to retrieve fewer hits, which is (i) less confusing for the user and (ii) users need fewer steps to reach the right time point.

As we indicated above, the dataset employed in this research is 33 hours approximately. As we were indexing only 54 procedures, it is difficult to provide semantically similar in higher ranks of the result list. Users consider search engine C a good approach because it only shows videos which contain real similarities with the query image, without showing false shots in the last positions. The participants indicated that this application is a good approach in order to re-find the video where the query image belongs within the whole, eventually huge, data set. Mostly, this result appears in the first top video on the list. They considered this a useful tool to the doctors, who day by day record a huge amount of data which is difficult to access and retrieve ad hoc when needed.

Chapter 6

Conclusions and Further Work

This thesis report has presented a novel application for retrieving shots within endoscopic video streams, which is based on a real world use case from laporoscopic surgery. We adapted and assessed and the existing LIRE tool for content-based video retrieval.

We can divide the experiments of this project in three main steps: the video indexing, the implementation and the user test. Once we have the uniform sampling of the input video, the implementation block is able to find the shots in the respective videos within the first results on the top of the list. The user test assessed whether the application satisfies the user' needs for each search in video and browsing result. Although this application was initially developed for surgery videos, we implemented a previous prototype in order to study the best solutions for retrieval trying new methods and extending the search to other domains.This implementation of the thinking aloud test in another domain gave me the opportunity to be ready to interact with different professionals in retrieval of the medical field and check how they make use of the indexing, visual search and result browsing of the first implementation of endoscopic videos.

For our first experiments, we reported on tests using a public available research dataset of musical performances. Those experiments have indicated that both methods employed, (i) late fusion of global features and (ii) use of SIMPLE based local features, have their merits for different types of queries in the investigated use case. Depending on the user's need, each method is useful in different scenarios. However, SIMPLE search is more accurate and can retrieve the same scene from different angles, while global features present a broader picture, match scenes with similar background and allow for a more exploratory type of search.

For our experiments within the medical dataset we could run a small study

with two expert users employing the same web interface modified for this case. They also indicate that such a tool is of value for the everyday work routine of a surgeon. One obvious approach is video hyperlinking, ie. to find visually similar scenes in different video streams, and therefore, allowing for non-linear video browsing.

The new solution works well in the Jiku Mobile data set due to the characteristics of the images in the whole database. We can find different scenarios, different kind of people, dresses, backgrounds, objects, etc. SURF detector can find good key points on these frames and CEDD can extract many colors from the events. However, in the medical domain, we find mostly red color, and the frames of the whole data set are different from each other only based on textures, the position of the organs, the stuff used by the surgeons rarely. The background only makes differences between outside the patient, inside or in the way until the right position of the surgery. There are no blocks, edge or shapes to let SURF detector finds good key points, and CEDD descriptor mostly requires 2 or 3 colors from the whole 24 bins it uses.

In order to find a better approach in this complicated domain for retrieval, other implementations of SIMPLE could give better results. Due to we have seen mostly red color in medical images, implementing the CEDD-modified descriptor will result in a wide range of red tonalities and reduce the sets of other colors in the spectrum of the descriptor. An example can be to not distinguish among the different sets of blue and green color, and amplify the sets of red and orange ones. Moreover, due to most of the information in medical images are concentrated in the center, to use a Gaussian Random keypoint detector instead of the evenly distributed random one could improve the results.

For further work in this field, we find that another interesting experiment would be to employ this approach to ad-hoc search within surgery procedures. Surgeons may take a shot and search the database for similar situations. Next steps in this project are a user study involving multiple surgeons, a large scale evaluation on our test data set including 600 shots. However, for practical use of our method we have to take into account the amount of indexing time. Local features of course take additional time as a code book has to be created, so for practical use the code book should be pre-computed. We have to investigate indexing strategies which allow for faster search time. We further aim at reducing the number of frames to be indexed by an automated method of frame selection for indexing. We aim also to fuse local and global features, which may allow us to get best of both worlds, and, for practical use, we want to speed up indexing time.

The results have shown that our approach is able to properly extract visual instances that shows similar content of the videos of the two data set we have employed in this project. This study finally has led to the publication of two Papers which were accepted by the IEEE/ACM 13th International Workshop on Content-Based Indexing (CBMI) [1]: "Event Video Retrieval using Global and Local Descriptors in Visual Domain" and "Visual Information Retrieval in

[1] `http://siret.ms.mff.cuni.cz/cbmi2015/`

Endoscopic Video Archives". Both of them are presented in the Appendix II of this Thesis.

Appendix I

This Appendix describes individually each of the queries the participants picked up during the user tests in the event domain.

Due to the Jiku Mobile data set was used in order to understand the Lire software and the previous demo application, all the exploratory search of new solutions was made using this group of images. For this reason and taking into account that more volunteers did this test using this domain, we have an exhaustive evaluation of each query image.

Figure 6.1: Most used query images in the user test (left to right).

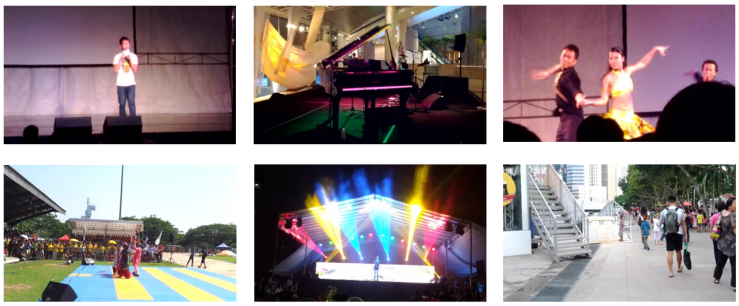

Figure 6.2: Other used query images in the user test (left to right).

The most used queries are shown in Figure 6.1.

- Query frame 1:

  Query image in the top left corner of the figure 6.1. This key frame belongs to the first set of images in the first event. It shows five men separated and singing in the middle of the stage. Many users chose this picture as a result of its simplicity searching for similar images with a few actors in the scenario or just due to it was the image just in the middle of the whole set. By means of this query image, the user expects to see music performances with a small group of people singing and dancing in the stage, probably with the same colours.

  Search engine A and B: All videos look like the same event and users were surprised to see the video in the first place where the query image was taken from. The user characterizes this fact as a great result. The user realizes that there are many different performances on the same stage of the same event in all the result within the videos. All the participants agree that there are not too many differences between the two first search engines. Moreover, they are always taking in account the videos, not the hits within them. Mostly, users wonder themselves why, in some query images, the third and the fourth video do not show so good results. Regarding to their opinion, a bad result is that one which doesn't show the same people dancing in the stage. For example, in the first key frame, they cannot see the guitar on the stage that the query image shows. Even though, they prefer search engine A because of the fourth video shows better results than in the second engine; at least, it shows people with the same clothes and it makes sense. As a summary, all of the participants can see similarities between the resulting hits; however, they realize that there are frames between the results which are also similar but they are not marked.

  Search engine C: There are fewer results than the other two search engines. Users are interested in this search engine due to he can always get the same performance, the same scenery and group of people on the stage, from different viewpoints in the event. It is a really good result for all of them. It is the best result for the user because only one hit shows one image from other performances. This is, therefore, what they were expecting to see.

- Query frame 2:

  Query image is in the top right corner of the figure6.1. This keyframe belongs to the fourth event. It shows an open space; outdoor image. The user wants to see the contrast between the results with the last query image. It is a colourful picture in a park with people playing music and dancing together. By means of this query image, the user expects to see music performances with a big group of people singing and dancing in the stage with the same colourful background and by this far point of view.

  Search Engine A and B: Users were interested in the first video. It shows the video in the first position where the query image was taken. It seems to show, in all of the videos, the same performance from the

chosen image, same surrounding in the background with people dancing with umbrellas. The best result for the user is to see, in the first position, the video where the query belongs. The other videos show different sub-events. In these cases, they can see more results than in the third search engine. Users did not expect that but it is also interesting because, at least, shows the same event. It is a good engine for getting an overview from the event. There are other performances different to the query image. But the results are always people dancing in the stage from the same perspective.

Search engine C: They are surprised that there is only one video result. The user was surprised and wondered himself about why the third result only shows one video. They wonder themselves that maybe there is only one video in the whole database which contain this specific sub-event with people dancing with the umbrella. It covers their expectations, so it is a good search engine for the user. All the volunteers express that there are not false hits and it is also great that the application does not show other videos if they are not containing the same performance.

- Query frame 3:

  Query image in the bottom left corner of the figure6.1. This key frame belongs to the first set of images in the first event. It shows an indoor photography. The image was taken from the right view point and the background has a uniform red colour. We can see only one person talking or singing on the stage. By means of this query image, the user expects to see whatever performance but only one person on the stage, probably with the same colour in the background.

  Search Engine A and B: Even the users could not see too many differences between search engine A and B, they find better results in the second engine due to the order of the videos. The first video shows many hits with the same performance. It might be also because of the significant colour of the background. The next videos also show the same stage and in some hits the same performances but without the same red colour as background. There are also different performances as results. Two of the participants suggest that it is a wrong result when the application shows other performances different to the chosen query image. However they can see the same background and comment that it is a great result. If the results show another performances and different coloration in the background, the user refers to the hits as a completely wrong result.

  Search Engine C: The best results are shown in this search engine because the user finds exactly the results which she was interested in. On the other hand, the videos in the same position show better results than in the two last search engines. As a user, the volunteer is expecting to see the results quickly and not waiting until the last video to see more good results. However, there are frames with similar red appearance but definitely different performances. It is a good approach but the user expected to see the performances in fact.

- Query frame 4:

  Query image in the bottom right corner of the figure6.1. This key frame belongs to the third event. It shows an indoor photography. We can see

different people dancing with specific movements on the stage with red shirts. By means of this query image, the user expects to see the same group of people dancing with the same clothes as the query image. For the users, if the results show different movements in the dancing or different perspective of the image is not too much important, on the contrary, it is a great result.

Search Engine A and B: The user does not know the whole database and wonder himself if all the videos are singing and dancing performances. A good result would be found the same group of people with red jackets dancing. The results are good but the user can see different sub-events. One of the participants suggests that when she wants to see the performance which she is interested in, she has to go through all of the results until reach this one. She suggests it is not useful because she could go to the performances by herself, without taking in account the marked hits. She also suggests that the user must not guess where must be the right moment, the right performances, and with these search engines is difficult to reach the right moment the user is looking for. It is also a good approach due to it always shows people dancing in the stage with the same background, but they would prefer to see the right performance directly. Finally, users comment that for them is difficult to make differences between the two search engines. Even the same wrong moments are shown in the two results.

Search Engine C: Every result is correct, it is a totally superior search engine, and it is the best of all of them. Users can see only the hits they were looking for. If there are fewer hits, the results are less confusing for the user. Other users said: "without any doubt, this result is much better than the two first search engines. It always shows the best results, the same performance as the query image contains. There are not false hits".

Other interesting queries for the users are shown in Figure 6.2.

Other indoor images

- Key frame 5:

  Search Engine A and B: The results are quite similar. There are results which are not the same person or performances, but at least it shows only one person on the stage, It could be a woman or a man, but one person alone. The user doesn't find many differences between the two search engines.

  Search Engine C: The top results show good hits, quite similar to the query image.

- Key frame 6:

  Search Engine A and B: The first search engine fits better than the second one in this case, due to three of the videos show at least the piano, which is the query the user chose.

Search Engine C: For the user is totally great that the first video where the query belongs is shown in the first position. However, in this case with this key frame the other three results are not good enough; even they don't show the same performance or event.

- Key frame 7:

  Search Engine A and B: When the user chooses one key frame he expects to see people dancing in the stage, with the same colours if it could be possible. The user likes the video which contains the query image inside is shown in the first position. It is a good result for him taking in account that the user doesn't know the whole database. It is what he was expecting. The user cannot see too many differences between the two first search engines, he realize that the two engines show the same sequence of videos. He pays more attention to which videos are shown than the hits within those videos.

  Search Engine C: The third result doesn't fit as well as the two last ones due to his expectations.

Other outdoor images

- Key frame 8:

  Search Engine A and B: They are a good results, every moment are relevant. They are not the moment the user was looking for, but at least they are hits of the same event with the same background. Again, the correct video is on the top of the list, good result for the user. However, the user suggests the second search engine is better due to the first videos show near results to the query image.

  Search Engine C: The top result is correct; however, the user prefers the second search engine.

- Key frame 9:

  Search Engine A and B: The user was expecting these results. The two search engines show good results because we can see the stage with the different colour lines. The fourth video is not as good as the one in search engine 1 for the user because it shows more than one person on the stage, different than the query image. User comments that when the images are crowded the similarities are better. Definitely the user prefers approach A, because of the colours and the people.

  Search Engine C: The first video result is quite good for the user, but from the third video the results are not which the user expected to see.

- Key frame 10:

  Search Engine A and B: It shows a perfect results, four different videos recording the same scenery. Honestly, the user suggested that it would be better to see more videos or more images in the same screen. However, the user can see what he was expecting.

  Search Engine C: The third result doesn't fit as well as the two last ones due to his expectations.

Appendix II

This Appendix contains the submitted version of the two scientific papers related to this thesis that have been accepted for presentation the IEEE/ACM 13th International Workshop on Content-Based Multimedia Indexing (CBMI), which will be hold in Prague (Czech Republic) between the 10th and 12th June 2015. The articles have been a collaboration of my supervisor in Austria, Mathias Lux, my supervisor in Barcelona Xavier Giró-i-Nieto, Pia Muñoz who created the previous implementation, and finally Nektarios Anagnostopoulos who collaborated with the new solution using local features.

For the musical performance domain we wrote with 4 pages as a Demo Paper under the name "Event Video Retrieval using Global and Local Descriptors in Visual Domain". For the endoscopic videos we wrote a 6 pages as a Special Session under the name "Visual Information Retrieval in Endoscopic Video Archives".

Event Video Retrieval using Global and Local Descriptors in Visual Domain

Jennifer Roldan Carlos*, Mathias Lux*, Xavier Giro-i-Nieto†, Pia Munoz* and Nektarios Anagnostopoulos*
*Klagenfurt University
Klagenfurt, Austria
Emails: jroldancarl@gmail.com, mlux@itec.aau.at, piamunozt@gmail.com, nek.anag@gmail.com
†Universitat Politecnica de Catalunya
Barcelona, Catalonia/Spain
Email: xavier.giro@upc.edu

***Abstract*—With the advent of affordable multimedia smart phones, it has become common that people take videos when they are at events. The larger the event, the larger is the amount of videos taken there and also, the more videos get shared online. To search in this mass of videos is a challenging topic. In this paper we present and discuss a prototype software for searching in such videos. We focus only on visual information, and we report on experiments based on a research data set. With a small study we show that our prototype demonstrates promising results by identifying the same scene in different videos taken from different angles solely based on content based image retrieval.**

I. Introduction

Many people like to share their experiences with friends. A large part of them uses the internet to publish and send pictures and videos from what they have seen, visited and experienced. YouTube alone currently has more than 300 hours worth of videos uploaded every minute[1]. Especially for large events where lots of people attend, it is common to find multiple videos from the same time and same location on YouTube, Facebook and alike, and it is hard to keep track on which videos show what.

In this paper we present a prototype for near duplicate visual search in videos. With such a prototype one can search for visually similar video frames throughout a collection of videos and eventually find those that have been taken from the same scene. For input our system relies on a video frame or an image. With the given query the system finds videos, where similar frames occur, ranks them by the relevance of the frames and the amount of frames found, and returns a list of videos with the relevant frames highlighted (cp. Figures 1, 2). For indexing we sample equidistant frames and use both, global and local features, for search. Result aggregation is done by late fusion.

The overall goal of the prototype is to give a proof of concept that visual search can be used to identify videos from events, where multiple videos have been recorded from the same scene. Based on the visually similar frames we assume videos can be hyperlinked or even roughly synchronized. We show the applicability of our approach by using the Jiku Mobile data set [1], which features videos taken from different users from a set of events; including dancing, singing and sports. For each event, multiple temporally and spatially overlapping videos are available.

[1]https://www.youtube.com/yt/press/, visited 2015-03-03

Fig. 1. Screen shot of a single result with the frames visually similar to the query highlighted.

A. Related Work

Related work in this field is, while it being an obvious approach, rather sparse. Most important related initiatives video and multimedia retrieval challenges such as TRECVID [2], MediaEval [3] and ImageCLEF [4]. While the tasks of the initiatives are changing, near duplicate frame search either has been a task or it has been used as means to an end for tackling one of the tasks. Our prototype is different to previous approaches as it incorporates the SIMPLE descriptors, which are local features used for the first time in the field of video retrieval.

In [5] the authors present a system also focusing on videos taken at events. However, they employ a more controlled and holistic approach. Videos recorded with their software are automatically enriched with meta data, ie. sensor readings, which allows for faster and easier retrieval, while we do

not restrict the video recording procedure and operate on visual data only. In [6] the authors present a system, which automatically creates an event summary based on different videos from different users and view points. The system, called Jiku Director, operates on the same data set as our prototype does, but relies solely on meta data. The main contribution is the creation of the summary, not the retrieval of scenes.

A similar case using a large scale dataset is presented in [7]. The dataset contains of 3,800 hours of newscasts and features 200 queries for retrieval evaluation providing a ground truth. The queries are images and have to be found in the video streams, an approach the authors call image-to-video, *I2V*. Moreover, the authors present a system operating on the data set in [8].

II. Our prototype

In our demo application we focus on content based video indexing and retrieval to match example query content to target video content by extracting and indexing visual feature descriptors. Each descriptor can be considered as an *independent retrieval model* [9] which at some point needs to be fused. Mainly, two types of fusion schemes are considered. In *early fusion* the retrieval models are integrated from the start and afterwards a multimodal representation is learned. *Late fusion* approaches on the other hand infer similarity directly from unimodal features and integrate results at the end [10].

In the demo, we employ a late fusion model based on multiple global features using a single visual example. The goal of late fusion techniques is the combination and re-score or re-rank of the initial result lists into a single final list. Before fusing the top hits from different lists it is required to truncate to the top N results and normalize them either by rank

$$\bar{R}_k(n) = \frac{N + 1 - R_k(n)}{N}$$

or by score

$$\bar{R}_k(n) = \frac{R_k(n) - min(R_k)}{max(R_k) - min(R_k)}$$

where R_k is the initial result (rank or score) from the retrieval model k. For our demo we apply the sum approach, where either normalized ranks or scores are summed up (cp. fusion strategies in [11]):

$$R_t(n) = \sum_k (R_k(n)) = R_1(n) + R_2(n) + ... + R_K(n)$$

For late fusion we used three different global features, (i) *CEDD* [12], a compact joint histogram of fuzzy color and texture, (ii) the *auto color correlogram* [13], a color feature that measures how often a color encounters itself in a neighborhood, and (iii) the *pyramid histogram of oriented gradients* (PHOG) [14], a fuzzy gradient histogram organized in a spatial pyramid.

In addition to the global descriptors, we also introduce localized version of CEDD employing the SIMPLE model [15], which has outperformed classical local features in many scenarios. SIMPLE uses a key point detector to find salient points on different scales. Based on the scale the point has been found, a local image patch is indexed with a compact and composite descriptor. Following that, the *bag of visual words* model is used to aggregate local features into histograms. We used SIMPLE with the CEDD feature, the SURF key point detector [16], and k-means to create a visual vocabulary of 512 visual words. All of the features were extracted with the open source library LIRE [17].

Fig. 2. Screen shot of the application showing a query and the first four results.

As search process is based on frames within the videos and the result list is also composed of video frames, our system aggregates the frames as a last step. Based on the top 40 hits for each query, we determine the four best matching videos and present them to the user while visualizing the time location where the matching frames have been actually found, as shown in Fig. 2. For this reason, the final ranked list of videos is based on their best matching frame, ie. the first frame defines the best matching video, the first frame of a different video in the result list of frames defines the second best matching video, etc.

III. Experiments

We used the Jiku Mobile data set [1] for our study, which is a set of 473 video clips taken at five different social events. The videos were recorded by different people from different angles. They feature pairwise overlap time- and scene-wise. For our experiments we indexed 356 randomly selected videos based on equidistant frames, using one frame per second. A set of 412 queries of different performances (cp. Figure 3) in

Fig. 3. Sample queries showing scenes from indoor and outdoor events as well as different points of view.

the social events was created manually. We aimed to cover different aspects, like for instance, outdoor, indoor scenes, colorful, and simple scenes.

Based on the 412 queries we created a benchmarking data set. We tested if all the queries are to be found within the video data set. Our tests have shown that the video from which the query frame was extracted was ranked at the first position for 96% of the cases (cp. Table I). This confirms that the subsampling of one frame per second is enough for the data set to yield meaningful and accurate results with our approach.

TABLE I
RESULTS OF THE TESTS ON WHERE THAT ACTUAL VIDEO CAN BE FOUND IN THE RESULTS. THE FIRST TWO COLUMNS GIVE THE TWO DIFFERENT TESTED FEATURE FUSION APPROACHES, THE THIRD ONE GIVES THE RESULTS ON THE USE OF SIMPLE-CEDD.

	Sum of Ranks	**Sum of Scores**	**SIMPLE**
Precision @ 1	0.964	0.966	0.908
Precision @ 2	0.976	0.976	0.927
Precision @ 3	0.978	0.978	0.927
Precision @ 4	0.981	0.983	0.927

In order to test our prototype, we implemented a semi-interactive web based interface which allows users to dynamically select a query image and see the search results from three search configurations. In particular, the interface presents to the users a manually selected set of query frames from five social events of the Jiku Mobile data set. Users can explore the results from three configurations, named *search engines* for the sake of the test. These three approaches have been labeled as *search engine A* (for sum of ranks fusion of global features), *search engine B* (for sum of scores fusion of global features) and *search engine C* (for the use of SIMPLE based local features).

We asked the users to test which of the three search engines satisfied the users' needs, and which of them gives subjectively better results by mining ie. more accurate or broader. We did not want to give the users a goal beside explaining them what the data set and the queries meant. It was up to them to decide if the search engines returned what seemed *natural* to the

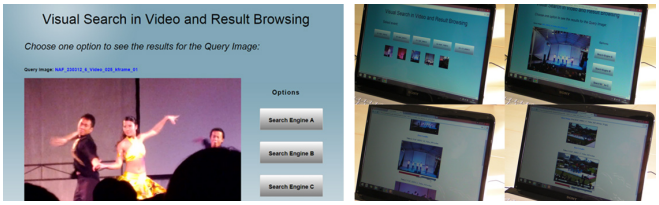

Fig. 4. Screen shot of the application showing the query interface (left) and a view on the test subjects environment (right).

users. Each user freely chose different queries and investigated the results provided by the three search engines. In that sense it was a heuristic evaluation asking experts on the overall performance. All six test subjects had been working in the field of computer science, and especially multimedia research for several years.

For evaluating the results we employed a *thinking aloud test* setup as described in [18]. It consisted of two parts. (i) Part one is a hands-on experience by different participants who used our tool. In this part we asked the participants to voice their thoughts and we did not interfere or encourage them. The sessions were recorded with one video camera over their shoulders capturing the mechanical interaction with the tool. (ii) Part two is an open interview reflecting his experience with the tool. Users are asked during the interview what they think about the tool and which conclusions they extract from this test. Is it a useful tool? Does this tool cover their expectations? After the tests we reviewed and transcribed all the interviews and test sessions. Based on the transcripts and the notes taken we discussed the results and concluded on the test.

As a general overview, we noticed the users were expecting to see *visually* similar scenes or *the same performances* in the results of the search. They particularly looked out for hints that this is a video showing the very same event, and eventually the same part of the event. All of them appreciated the similarities in the background, the stage or the number of people which are shown in the results. However, the main expectation they had was to find the same performance from *different point of view*.

The participants choose the query image based on their intuition of what would result interesting, ie. they were driven by their own curiosity. They were driven by many reasons, as for example the simplicity of the scenario with specific people on it, or colorful scenario outdoor crowded of people. Other reasons are the out-of-the-ordinary background color or a specific performance with out-of-the-ordinary movements.

After choosing a query image, some users were expecting to see directly videos showing the performance. They realize later that the results are shown in the time line. They expressed their view about the time line as a great tool to use in the demo application, as it allows the user to go directly to the final results without the need to watch the whole video. To investigate subtle and non-obvious differences between the different search engine, participants opened multiple tabs in the web browser and compared the results by switching

between them.

As an overall impression, for the search engines A and B, which are the sum of ranks and sum of scores fusion of global features, user comment they are good approaches for abstract exploratory search with a query as an example, and when searching for scenes with the same viewpoint of the stage, even with different sub-events. For search engine C, which is the SIMPLE based local features approach, all the users agree on this is the search engine that fits better when the user is searching for semantically similar content. Mostly, it shows the same performance with different viewpoints. Moreover, this search engine tends to retrieve fewer hits, which is (i) it is less confusing for the user and (ii) users need fewer steps to reach the right time point.

IV. Conclusion

In this paper we have presented a prototype implementation for video search based on frames and visual information retrieval. We further reported on tests using a freely available research data set. Our experiments have indicated that both methods employed, (i) late fusion of global features and (ii) use of SIMPLE based local features, have their merits for different types of queries in the investigated use case. Using SIMPLE search is more accurate and can retrieve the same scene from different angles, while global features present a broader picture, match scenes with similar background and allow for a more exploratory type of search.

However, for practical use of our method we have to take into account the amount of indexing time. Local features of course take additional time as a code book has to be created, so for practical use the code book should be pre-computed.

In the future we want to try our method, especially visual search in videos based on SIMPLE on larger data sets and we want to compare it to more traditional local feature approaches like SIFT/SURF BoVW [19]. We further aim to fuse local and global features, which may allow us to get best of both worlds, and, for practical use, we want to speed up indexing time.

Acknowledgements

This work was supported by Lakeside Labs GmbH, Klagenfurt, Austria and funding from the European Regional Development Fund and the Carinthian Economic Promotion Fund (KWF) under grant KWF-20214/25557/37319. It has also been developed in the framework of the project TEC2013-43935-R, financed by the Spanish Ministerio de Economa y Competitividad and the European Regional Development Fund (ERDF).

References

[1] M. Saini, S. P. Venkatagiri, W. T. Ooi, and M. C. Chan, "The jiku mobile video dataset," in *Proceedings of the 4th ACM Multimedia Systems Conference*, ser. MMSys '13. New York, NY, USA: ACM, 2013, pp. 108–113. [Online]. Available: http://doi.acm.org/10.1145/2483977.2483990

[2] A. F. Smeaton, P. Over, and W. Kraaij, "Evaluation campaigns and trecvid," in *MIR '06: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*. New York, NY, USA: ACM Press, 2006, pp. 321–330.

[3] M. Larson, B. Ionescu, X. Anguera, M. Eskevich, P. Korshunov, M. Schedl, M. Soleymani, G. Petkos, R. Sutcliffe, J. Choi, and G. J. Jones, Eds., *MediaEval 2014 Multimedia Benchmark Workshop*, ser. CEUR Workshop Proceedings, vol. 1263, Oct. 2014.

[4] H. Müller, P. Clough, T. Deselaers, and B. Caputo, *ImageCLEF: Experimental Evaluation in Visual Information Retrieval*, 1st ed. Springer Publishing Company, Incorporated, 2010.

[5] P. Seshadri, M. Chan, W. Ooi, and J. Chiam, "On demand retrieval of crowdsourced mobile video," *Sensors Journal, IEEE*, vol. PP, no. 99, pp. 1–1, 2014.

[6] D.-T.-D. Nguyen, M. Saini, V.-T. Nguyen, and W. T. Ooi, "Jiku director: A mobile video mashup system," in *Proceedings of the 21st ACM International Conference on Multimedia*, ser. MM '13. New York, NY, USA: ACM, 2013, pp. 477–478.

[7] A. Araujo, J. Chaves, D. Chen, R. Angst, and B. Girod, "Stanford I2V: A News Video Dataset for Query-by-Image Experiments," in *Proc. ACM Multimedia Systems*, 2015.

[8] A. Araujo, D. Chen, P. Vajda, and B. Girod, "Real-time query-by-image video search system," in *Proceedings of the ACM International Conference on Multimedia*, ser. MM '14. New York, NY, USA: ACM, 2014, pp. 723–724. [Online]. Available: http://doi.acm.org/10.1145/2647868.2654867

[9] H. J. Escalante, C. A. Hérnadez, L. E. Sucar, and M. Montes, "Late fusion of heterogeneous methods for multimedia image retrieval," in *Proceedings of the 1st ACM international conference on Multimedia information retrieval*. ACM, 2008, pp. 172–179.

[10] C. G. Snoek, M. Worring, and A. W. Smeulders, "Early versus late fusion in semantic video analysis," in *Proceedings of the 13th annual ACM international conference on Multimedia*. ACM, 2005, pp. 399–402.

[11] K. Mc Donald and A. F. Smeaton, "A comparison of score, rank and probability-based fusion methods for video shot retrieval," in *Image and video retrieval*. Springer, 2005, pp. 61–70.

[12] S. A. Chatzichristofis and Y. S. Boutalis, "CEDD: color and edge directivity descriptor: A compact descriptor for image indexing and retrieval," in *Computer Vision Systems, 6th International Conference, ICVS 2008, Santorini, Greece, May 12-15, 2008, Proceedings*, 2008, pp. 312–322.

[13] J. Huang, R. Kumar, M. Mitra, W. Zhu, and R. Zabih, "Image indexing using color correlograms," in *1997 Conference on Computer Vision and Pattern Recognition (CVPR '97), June 17-19, 1997, San Juan, Puerto Rico*, 1997, pp. 762–768.

[14] A. Bosch, A. Zisserman, and X. Munoz, "Representing shape with a spatial pyramid kernel," in *Proceedings of the 6th ACM international conference on Image and video retrieval*. ACM, 2007, pp. 401–408.

[15] C. Iakovidou, N. Anagnostopoulos, A. C. Kapoutsis, Y. Boutalis, and S. A. Chatzichristofis, "Searching images with MPEG-7 (& mpeg-7-like) powered localized descriptors: the SIMPLE answer to effective content based image retrieval," in *12th International Workshop on Content-Based Multimedia Indexing (CBMI)*. IEEE, 2014, pp. 1–6.

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

[17] M. Lux and S. A. Chatzichristofis, "Lire: lucene image retrieval: an extensible java CBIR library," in *Proceedings of the 16th International Conference on Multimedia 2008,*, Vancouver, Canada, Oct 2008, pp. 1085–1088.

[18] T. Boren and J. Ramey, "Thinking aloud: Reconciling theory and practice," *Professional Communication, IEEE Transactions on*, vol. 43, no. 3, pp. 261–278, 2000.

[19] J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos," in *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*. IEEE, 2003, pp. 1470–1477.

Visual Information Retrieval in Endoscopic Video Archives

Jennifer Roldan Carlos*, Mathias Lux*, Xavier Giro-i-Nieto†, Pia Munoz* and Nektarios Anagnostopoulos*
*Klagenfurt University
Klagenfurt, Austria
Emails: jroldancarl@gmail.com, mlux@itec.aau.at, piamunozt@gmail.com, nek.anag@gmail.com
†Universitat Politecnica de Catalunya
Barcelona, Catalonia/Spain
Email: xavier.giro@upc.edu

***Abstract*—In endoscopic procedures, surgeons work with live video streams from the inside of their subjects. A main source for documentation of procedures are still frames from the video, identified and taken during the surgery. However, with growing demands and technical means, the streams are saved to storage servers and the surgeons need to retrieve parts of the videos on demand. In this submission we present a demo application allowing for video retrieval based on visual features and late fusion, which allows surgeons to re-find shots taken during the procedure.**

I. Introduction

While maintaining large video archives is an expensive venture for clinics and hospitals, more and more countries require the storage of those videos for legal reasons. Therefore, a growth of video archives over the next years is expected, especially related to endoscopic videos. As a consequence clever methods for indexing and retrieval are needed. Users of such an archive should be able to retrieve information on specific procedures, types of procedures or similarities between different procedures with ad hoc searches.

There are mostly two main approaches for the creation of stored endoscopic videos depending on the doctors in charge of the procedure. (i) Those surgeons who are aware of the space requirements of videos and the tedious work of identifying relevant section in hour long recordings, typically turn on and off recording to just document the most important steps or results of the procedure. (ii) Surgeons, who just want to document their procedures for legal reasons and are not bound to re-visit them later, record the whole procedures including even large parts of the preparations and clean-up afterwards, which are typically out-of-patient recordings of less importance. However, in both cases surgeons rely on the same *photo function*, which allows them to grab a frame from the video stream and store it, ie. to put it in a report later on.

In this paper we focus on the relation of *photos* taken by a surgeon to the actual video streams as depicted in Fig. 1. These photos, which we call *shots* throughout the paper, are merely frames (still images) that have been saved at the time of operation on request of the surgeon, so they are also part of the video stream itself. Most important, what distinguishes them from the other frames of the video is that the surgeon intentionally directed the camera to a view to capture an optimal picture for later reference.

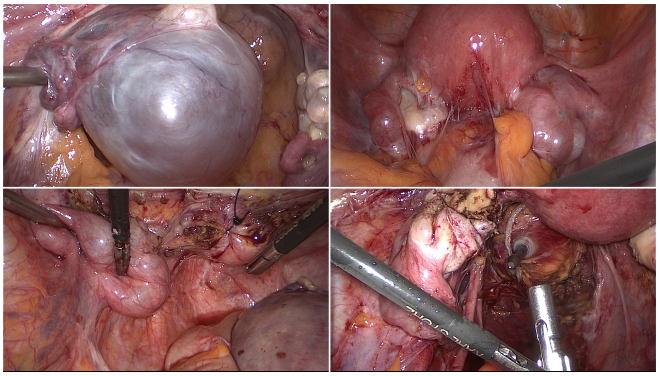

Fig. 1. Shots (photos) manually created from the surgeon in the course of the procedure.

In the framework presented in this paper, we focus on *re-finding* those shots within video streams, i.e. we assume that the shots are known, but we want to (a) find the part of the original video where the shot was taken, and (b) find videos with visually similar frames to identify semantically similar scenes in different procedures. Ultimately, we believe such a system can be used for supporting medical research, education and training. We tested our application on a set of 1.276 videos ($\approx$ 33 hours) from 54 procedures.

The remainder of the paper is organized as follows. After surveying the most important related work, we first present the methods used for our approach, and then outline our application. After describing the test setup and presenting the results of a retrieval evaluation experiment and a qualitative result study, we conclude our paper and outline the next steps.

II. Related work

In literature, a large number of research publications in medical imaging can be found, chiefly for gray scale images such as X-rays or magnetic resonance imaging (MRI). [1] describes potential applications of medical image retrieval and reviews some existing medical CBIR systems. [2] also introduces different types of medical images used in CBIR systems as

well as a large variety of techniques, potential applications and future lines. [3] provides a more recent review, emphasizing the multi-dimensional (2D and 3D) and multi-modality nature of the medical retrieval scenario. Nevertheless, medical image and video retrieval remains an area of active research.

For example, the ImageCLEF benchmark [4] has created a strong community of researchers participating in the retrieval of medical images. A task for image-based retrieval was organized between 2004 and 2013. This case differs from the one addressed in this work because they were defined with 1-7 sample images accompanied by text. In the 2013 edition [5], the best textual run achieved the same performance as the best technique using both textual and visual features [6]. As in previous years, visual-only approaches achieved much lower results than the textual and multimodal techniques. The best visual-based solution [7] was based on the Color and Edge Directivity Descriptor (CEDD), a fuzzy color and texture histogram and a Color Layout Descriptor.

Content-based image retrieval in the medical domain has been addressed from low-level wavelet-based visual signatures [8] to high level concept detectors [9]. Another way to exploit visual features is to generate automatic text descriptors with computer vision algorithms [10] and use these labels to support text-based queries.

Nowadays, medical retrieval systems have already become much more accessible on the web, typically supporting both textual and visual queries. These are the cases of NovaMed-Search [11] or GoldMiner [12].

In contrast to most works on medical CBIR tasks, we address the problem of video retrieval, instead of still images. This venue has been previously explored in the literature. Specifically for real medical videos, [13] proposes a framework that uses principal video shots for video content representation and feature extraction. The classification is mainly implemented by elementary semantic medical concepts, such as "Traumatic surgery" or "Diagonosis". Moreover, [14] presents a framework to retrieve short videos in real time by modeling the motion content with a polynomial model.

III. Methods

In our approach we focus on content based video indexing and retrieval to match example query content (still images) to target video content by extracting and indexing visual feature descriptors. For tests on the utility and usefulness of different approaches, we implemented three methods for visual retrieval: two of which use global features and feature fusion, and the third one which employs local features based on a recent model.

A. Global and Local Features

In our study we have tested three different types of global features: (i) *Color and Edge Directivity Descriptor (CEDD)* [15], a compact joint histogram of fuzzy color and texture, (ii) the *auto color correlogram* [16], a color feature that measure how often a color encounters itself in a neighborhood, and (iii) the *pyramid histogram of oriented gradients* (PHOG) [17], a fuzzy gradient histogram organized in a spatial pyramid.

A local feature solution has also been adopted to be compared with the global ones. We employ a localized version of CEDD using the SIMPLE model [18] which has outperformed classical local features in many scenarios. SIMPLE uses a key point detector to find salient points on different scales. Based on the scale the point has been found, a local image patch is indexed with a compact and composite descriptor. Following that, the bag of visual words model is used to aggregate local features into histograms. We used SIMPLE with the CEDD feature, the SURF key point detector [19], and k-means to create a visual vocabulary of 512 visual words.

Following the extraction of local features, the *bag of visual words* model [20] is applied to generate local feature histograms. The experiments reported in this paper were based on a visual vocabulary of 512 words build with the k-means clustering algorithm.

B. Late Fusion by Rank and by Score

For fusion, each descriptor can be considered as an *independent retrieval model* [21]. To incorporate more characteristics than just one feature vector, independent retrieval models can be fused. Mainly, two types of fusion schemes are typically adopted. In *early fusion* the different retrieval models and feature spaces are integrated from the start, and afterwards a multimodal representation is learned. *Late fusion* approaches on the other hand infer similarity directly from unimodal features by creating a relevance score or ranked list for each of them, and integrate results at the end [22] by fusing different scores or ranks.

Fig. 2 shows the overall architecture. First, in an offline process, frames are collected and indexed. Based on the index and ad hoc search, similarity in different retrieval models is computed. For each of the feature spaces we get a ranked list, which are then fused to get a final ranked result list.

In our approach, we employ a late fusion model based on multiple visual global features using a single query image. The objective of late fusion techniques is the combination and re-scoring or re-ranking of the initial result lists into one final list. Typically one truncates the initial lists to the top N results and normalizes them either by rank

$$\bar{R}_k(n) = \frac{N + 1 - R_k(n)}{N}$$

or by score

$$\bar{R}_k(n) = \frac{R_k(n) - min(R_k)}{max(R_k) - min(R_k)}$$

where R_k is the initial result (rank or score) from the retrieval model k. For our approach we apply the sum approach, where either normalized ranks or normalized scores are summed up (cp. fusion strategies in [24]), testing two approaches, sum of ranks and sum of scores:

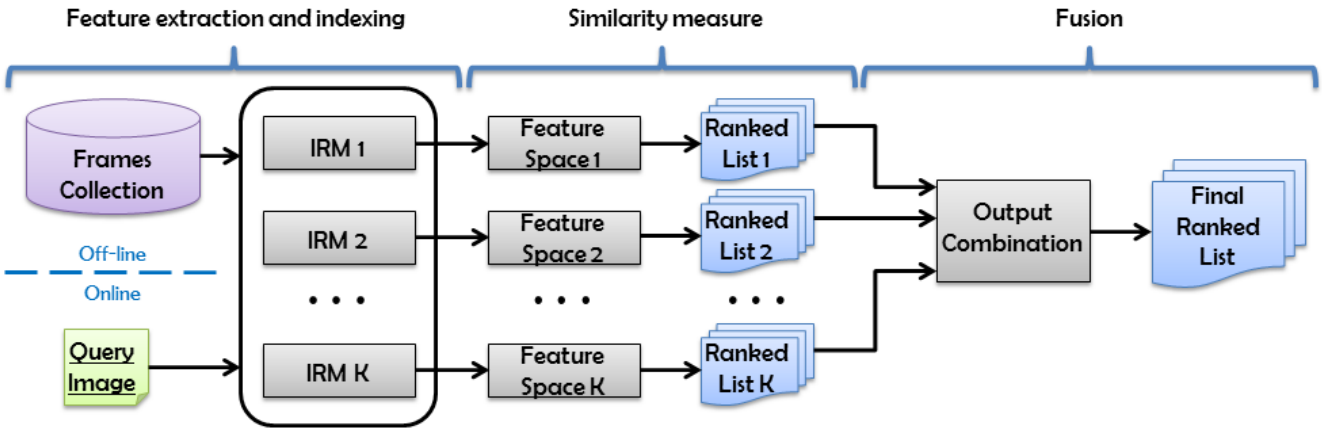

Fig. 2. Application of late fusion in our approach, illustration is based on the work in [22], [23], [1].

$$R_t(n) = \sum_k (R_k(n)) = R_1(n) + R_2(n) + ... + R_K(n)$$

IV. Our Application

The goal of our application is to test and compare the different visual features and fusion methods presented in Section III for the retrieval of endoscopic videos. In particular, we addressed the use case of re-finding shots within video streams with a query still image.

This application was developed on a dataset of 1,276 video clips that were temporally sampled at 5 frames per second. In order to define the experiments, we created a test dataset of query images. For this purpose, we used the shots generated by the surgeons during real procedures whenever they wanted to document a specific event that they consider important in the course of the surgery. This way we exploited the interaction from experts in endoscopic videos to determine the highly informative frames in the video, assuming that given the original intention queries in a retrieval system would be from a similar nature. Notice that, as a result, our set of queries is a new group of images different from the uniformly sampled frames from the video dataset. Even more so, as the shots are taken from the live and not the recorded video, we assume that some of them are not even in the recorded clips. Using experts, we cleaned out the query set aiming to remove stills that do not reflect a recorded video frame, ie. out-of-patient shots, survey shots, etc., resulting in 600 queries.

The test frames were indexed using the LIRE software library [25], a highly versatile image retrieval engine that can extract and integrate up to 20 different visual features. All features and fusion strategies described in Section III were implemented and assessed on this platform. Given that the reported experiments are a proof of concept, we did not explore at this stage additional indexing strategies such as index splitting, hashing or metric indexing.

Our application presents the results in a visual form in HTML5 for a recent version of common browsers. For each query an HTML file is generated displaying the query image, the list of similar images that the demo application finds, and the videos where both the query image and the rest of the frames belong to. All of the items appear along the time line where the images were taken. The screenshot presented in Fig. 3 shows the results of a shot query. Instead of showing the image results, only their positions in the video are indicated in the time line. Due to the nature of visual similarity search, retrieved frames look very much like the query, so showing them would not help the user in re-finding them in the video streams.

Based on the top 10 hits for each query, we determine the three best matching videos and present them to the user, highlighting the time location where the matching frames have been actually found, as shown in Fig. 3. As the search process is based on frames within the videos and the result list is also composed of video frames, our system aggregates the frames as a last step. For this reason, the final ranked list of videos is based on their best matching frame, ie. the most similar frame defines the best matching video, the next most similar frame of a different video defines the second best matching video, etc.

V. Evaluation

Our data set covers roughly 33 hours of anonymized video data of laparoscopy procedures. For each of the procedures we had several shots manually taken by the surgeons. The videos were taken from different surgerys cases of several patients. Due to the long duration of each intervention and the high resolution and bit rate of the videos, the whole surgery is divided in several videos, resulting in an overall file count of 1,276 videos. Due to the sheer size of the video archive, we employed temporal subsampling and extracted five frames a second for indexing, all in all 593,446 frames. Average linear search time for combining three retrieval models – *color and edge directivity descriptor* (CEDD), *color correlogram*, and *pyramid histogram of oriented gradients* (PHOG) – was 30 seconds. Note that for this proof of concept we did not employ

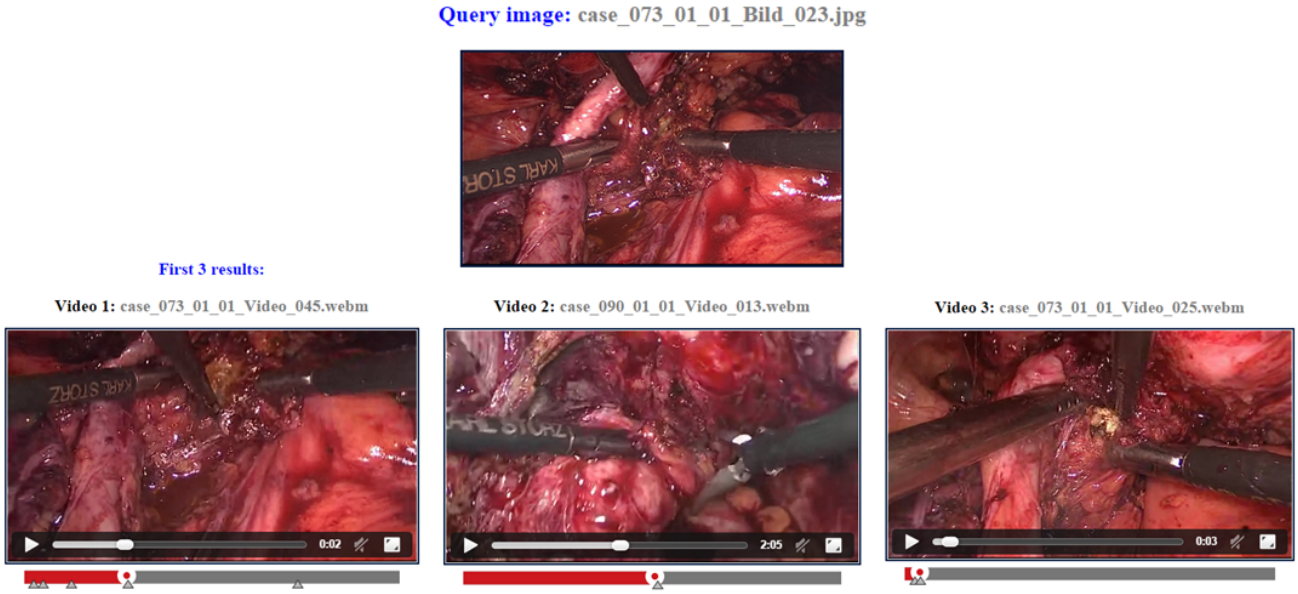

Fig. 3. Screenshots of the result presentation showing the three top videos and the query image. All results are presented in HTML5 and can be viewed in recent browsers supporting HTML5 videos and JavaScript. Best matching frames are indicated by triangles in the red and gray time line below the video player.

indexing strategies like hashing, metric indexes or clustering to speed up searching.

For our experiments, we used 600 queries based on shots captured by the surgeons, as presented in Section IV. Our experiments were twofold. First, we investigated the potential of each query to retrieve the video of the procedure where the query shot had been captured from. A quantitative metric was computed by comparing the retrieved videos with the ground truth. As our user interface only displays the top three ranked results, our study focused in the precision at positions 1, 2, and 3.

As a second qualitative evaluation was ran with a *thinking aloud test* [26]. We created an interactive web page (cp. Fig.3) featuring ten different surgery cases, and for each of them, the query shots available for search. The three search approaches were blindly labeled as search engine A (for sum of ranks fusion of global features), search engine B (for sum of scores fusion of global features) and search engine C (for the use of SIMPLE based local features). This was, we avoided any bias of the subjects towards any of the three approaches.

We asked participants to investigate and compare the results of the different search engines and to give us feedback upon their quality and their usefulness. To allow participants to investigate subtle and non-obvious differences between the different search engines, we encouraged them to open multiple tabs in the web browser and compared the results by switching between them. We asked the users to test which of the three search engines satisfies the users needs, and which of them gives subjectively better results by mining ie. more accurate or broader. It was up to them to decide if the search engines returned what seemed natural to the users. It was up to the users to pick several of the queries and investigate possible results. In that sense it was a heuristic evaluation asking experts on the overall performance. The test subjects had been working in the field of computer science focusing on retrieval and analysis of endoscopic videos for several years. The participants were asked to voice their thoughts throughout the tests and the tests have been recorded on video (cp. Fig. 4). After the tests we reviewed and transcribed the interview recordings and test sessions. Based on the transcripts and the notes taken we discussed the results and concluded on the test.

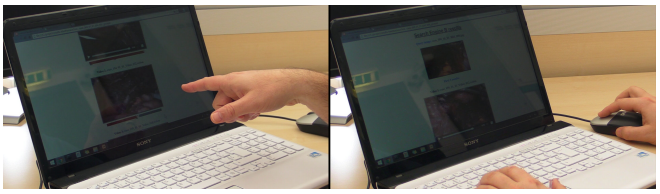

Fig. 4. Still frames from the thinking aloud test recordings. Test participants pointed out and explained the utility of particular results.

A. Experimental results

Based on the whole set of queries, our tests have shown that for 470 out of 600 (78.3%) of the queries, the source video was at the first position of the result list. In 84.2% of the queries the source video was among the top three positions for the *sum of ranks* approach, a very similar figure was obtained also for the the *sum of scores*. Local SIMPLE descriptor led to slightly better results, as in 79.8% of the queries the source video was in the first place, while in 84.6% of the queries the matching video was the first three videos (cp. Table I).

TABLE I
RESULTS OF THE TESTS ON WHERE THAT ACTUAL VIDEO CAN BE FOUND IN THE RESULTS. THE FIRST TWO COLUMNS GIVE THE TWO DIFFERENT TESTED FEATURE FUSION APPROACHES, THE THIRD ONE GIVES THE RESULTS ON THE USE OF THE SIMPLE-CEDD DESCRIPTORS.

	Sum of Ranks	**Sum of Scores**	**SIMPLE-CEDD**
Precision @ 1	470	471	479
Precision @ 2	21	20	21
Precision @ 3	14	15	8

This indicates that the subsampling of five frames per second is enough for the used dataset to yield meaningful results. Note at that point that the shots are not necessarily in the video frames as they were taken from the live videos, so the ground truth at hand is more on a semantic level than mimicking a near duplicate task.

In the second experiment – the thinking aloud test – users in general expected to see the same background in several shots within the videos, which are similar to the query image. The participants choose the query image based on their intuition of what would result interesting, ie. they were driven by their own curiosity. They were driven by many reasons, as for example the simplicity of the background with specific organs on it, or specific movements of the surgeons as for instance cut tissue. Other reasons are a specific background, ie. bloody or damaged tissue, or a specific event using different instruments, which lets the user relate to a specific part of the procedure. Based on the overal state of tissue seen in the scene, ie. if it has been cut or cauterized, users know a rough time point within the surgery from the video. It gives them an orientation about the specific moment of the intervention, ie. they know whether the video is from in the beginning, during or the end of the procedure. After choosing a query image, the participants were expecting to see directly videos showing similar interventions. Due to the length of the videos, users consider an useful tool in the application when the results are marked in the time line; it allows them to find the right moment without the need to watch the whole video.

As an overall impression, for the search engines A and B, which are the sum of ranks and sum of scores fusion of global features, user commented they are good approaches showing in the top results the most relevant shots within the videos. However, in many cases the videos with higher ranks in the results show content which is semantically dissimilar by for instance featuring a different organ, instrument or background. For search engine C, which is based on the SIMPLE local features, users agreed it is the search engine that fits better when searching for semantically similar content. This technique also tends to retrieve fewer hits, which is (i) less confusing for the user and (ii) users need fewer steps to reach the right time point.

As we indicated above, the dataset employed in this research is 33 hours approximately. As we are indexing only 54 procedures, it is difficult to provide semantically similar in higher ranks of the result list. Users consider search engine C a good approach because it only shows videos which contain real similarities with the query image, without showing false shots in the last positions. The participants indicate that this application is a good approach in order to re-find the video where the query image belong within the whole, eventually huge, data set. Mostly, this result appears in the first top video on the list. They consider this a useful tool to the doctors, who day by day record a huge amount of data which is difficult to access and retrieve ad hoc when needed.

VI. CONCLUSION

In this paper we presented a novel application for re-finding shots within endoscopic video streams, which is based on a real world use case from laporoscopic surgery. In our experiments we were able to find the shots in the respective videos within the first three results. A small study with two expert users also indicates that such a tool is of value for the everyday work routine of a surgeon. The methods employed, however, can be used in a number of scenarios. One obvious approach is video hyperlinking, ie. to find visually similar scenes in different video streams, and therefore, allowing for non-linear video browsing. Another interesting experiment would be to employ this approach to ad-hoc search within surgery procedures. Surgeons may take a shot and search the database for similar situations. Next steps in this project are a user study involving multiple surgeons, a large scale evaluation on our test data set including 600 shots. For deployment in real life, however, we have to investigate indexing strategies which allow for faster search time. We further aim at reducing the number of frames to be indexed by an automated method of frame selection for indexing.

ACKNOWLEDGEMENTS

This work was supported by Lakeside Labs GmbH, Klagenfurt, Austria and funding from the European Regional Development Fund and the Carinthian Economic Promotion Fund (KWF) under grant KWF-20214/25557/37319. It has also been developed in the framework of the project TEC2013-43935-R, financed by the Spanish Ministerio de Economa y Competitividad and the European Regional Development Fund (ERDF).

REFERENCES

[1] C.-H. Wei, C.-T. Li, and R. Wilson, "A content-based approach to medical image database retrieval," *Database Modeling for Industrial Data Management: Emerging Technologies and Applications. Idea Group, Hershey*, pp. 258–291, 2006.

[2] H. Müller, N. Michoux, D. Bandon, and A. Geissbuhler, "A review of content-based image retrieval systems in medical applications - clinical benefits and future directions," *International journal of medical informatics*, vol. 73, no. 1, pp. 1–23, 2004.

[3] A. Kumar, J. Kim, W. Cai, M. Fulham, and D. Feng, "Content-based medical image retrieval: A survey of applications to multidimensional and multimodality data," *Journal of digital imaging*, vol. 26, no. 6, pp. 1025–1039, 2013.

[4] H. Müller, P. Clough, T. Deselaers, and B. Caputo, *ImageCLEF: Experimental Evaluation in Visual Information Retrieval*, 1st ed. Springer Publishing Company, Incorporated, 2010.

[5] A. G. S. de Herrera, J. Kalpathy-Cramer, D. D. Fushman, S. Antani, and H. Müller, "Overview of the imageclef 2013 medical tasks," *Working notes of CLEF*, vol. 2013, pp. 1–15, 2013.

[6] A. G. S. de Herrera, R. Schaer, D. Markonis, and H. Müller, "Comparing fusion techniques for the imageclef 2013 medical case retrieval task," *Computerized Medical Imaging and Graphics*, vol. 39, pp. 46–54, 2015.

[7] O. Ozturkmenoglu, N. M. Ceylan, and A. Alpkocak, "Demir at imageclefmed 2013: The effects of modality classification to information retrieval," *Working Notes of CLEF*, 2013.

[8] G. Quellec, M. Lamard, G. Cazuguel, B. Cochener, and C. Roux, "Wavelet optimization for content-based image retrieval in medical databases," *Medical image analysis*, vol. 14, no. 2, pp. 227–241, 2010.

[9] M. M. Rahman, S. K. Antani, and G. R. Thoma, "A learning-based similarity fusion and filtering approach for biomedical image retrieval using svm classification and relevance feedback," *Information Technology in Biomedicine, IEEE Transactions on*, vol. 15, no. 4, pp. 640–646, 2011.

[10] J. Kalpathy-Cramer and W. Hersh, "Multimodal medical image retrieval: image categorization to improve search precision," in *Proceedings of the international conference on Multimedia information retrieval*. ACM, 2010, pp. 165–174.

[11] A. Mourão, F. Martins, and J. Magalhães, "Multimodal medical information retrieval with unsupervised rank fusion," *Computerized Medical Imaging and Graphics*, vol. 39, pp. 35–45, 2015.

[12] C. E. Kahn Jr and C. Thao, "Goldminer: a radiology image search engine," *American Journal of Roentgenology*, vol. 188, no. 6, pp. 1475–1478, 2007.

[13] J. Fan, H. Luo, and A. K. Elmagarmid, "Concept-oriented indexing of video databases: toward semantic sensitive retrieval and browsing," *Image Processing, IEEE Transactions on*, vol. 13, no. 7, pp. 974–992, 2004.

[14] G. Quellec, M. Lamard, G. Cazuguel, Z. Droueche, C. Roux, and B. Cochener, "Real-time retrieval of similar videos with application to computer-aided retinal surgery," in *Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE*. IEEE, 2011, pp. 4465–4468.

[15] S. A. Chatzichristofis and Y. S. Boutalis, "CEDD: color and edge directivity descriptor: A compact descriptor for image indexing and retrieval," in *Computer Vision Systems, 6th International Conference, ICVS 2008, Santorini, Greece, May 12-15, 2008, Proceedings*, 2008, pp. 312–322.

[16] J. Huang, R. Kumar, M. Mitra, W. Zhu, and R. Zabih, "Image indexing using color correlograms," in *1997 Conference on Computer Vision and Pattern Recognition (CVPR '97), June 17-19, 1997, San Juan, Puerto Rico*, 1997, pp. 762–768.

[17] A. Bosch, A. Zisserman, and X. Munoz, "Representing shape with a spatial pyramid kernel," in *Proceedings of the 6th ACM international conference on Image and video retrieval*. ACM, 2007, pp. 401–408.

[18] C. Iakovidou, N. Anagnostopoulos, A. C. Kapoutsis, Y. Boutalis, and S. A. Chatzichristofis, "Searching images with MPEG-7 (& mpeg-7-like) powered localized descriptors: the SIMPLE answer to effective content based image retrieval," in *12th International Workshop on Content-Based Multimedia Indexing (CBMI)*. IEEE, 2014, pp. 1–6.

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

[20] J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos," in *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*. IEEE, 2003, pp. 1470–1477.

[21] H. J. Escalante, C. A. Hérnadez, L. E. Sucar, and M. Montes, "Late fusion of heterogeneous methods for multimedia image retrieval," in *Proceedings of the 1st ACM international conference on Multimedia information retrieval*. ACM, 2008, pp. 172–179.

[22] C. G. Snoek, M. Worring, and A. W. Smeulders, "Early versus late fusion in semantic video analysis," in *Proceedings of the 13th annual ACM international conference on Multimedia*. ACM, 2005, pp. 399–402.

[23] H. J. Escalante, C. A. Hérnadez, L. E. Sucar, and M. Montes, "Late fusion of heterogeneous methods for multimedia image retrieval," in *Proceedings of the 1st ACM international conference on Multimedia information retrieval*. ACM, 2008, pp. 172–179.

[24] K. Mc Donald and A. F. Smeaton, "A comparison of score, rank and probability-based fusion methods for video shot retrieval," in *Image and video retrieval*. Springer, 2005, pp. 61–70.

[25] M. Lux, "LIRE: Open source image retrieval in java," in *Proceedings of the 21st ACM international conference on Multimedia*. ACM, 2013, pp. 843–846.

[26] T. Boren and J. Ramey, "Thinking aloud: Reconciling theory and practice," *Professional Communication, IEEE Transactions on*, vol. 43, no. 3, pp. 261–278, 2000.

Bibliography

[1] A. Araujo, J. Chaves, D. Chen, R. Angst, and B. Girod. Stanford I2V: A News Video Dataset for Query-by-Image Experiments. In *Proc. ACM Multimedia Systems*, 2015.

[2] Andre Araujo, David Chen, Peter Vajda, and Bernd Girod. Real-time query-by-image video search system. In *Proceedings of the ACM International Conference on Multimedia*, MM '14, pages 723–724, New York, NY, USA, 2014. ACM.

[3] Jim Bankoski. Intro to webm. In *Proceedings of the 21st International Workshop on Network and Operating Systems Support for Digital Audio and Video*, NOSSDAV '11, pages 1–2, New York, NY, USA, 2011. ACM.

[4] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. In *Computer vision–ECCV 2006*, pages 404–417. Springer, 2006.

[5] Ted Boren and Judith Ramey. Thinking aloud: Reconciling theory and practice. *Professional Communication, IEEE Transactions on*, 43(3):261–278, 2000.

[6] Anna Bosch, Andrew Zisserman, and Xavier Munoz. Representing shape with a spatial pyramid kernel. In *Proceedings of the 6th ACM international conference on Image and video retrieval*, pages 401–408. ACM, 2007.

[7] Savvas A. Chatzichristofis and Yiannis S. Boutalis. CEDD: color and edge directivity descriptor: A compact descriptor for image indexing and retrieval. In *Computer Vision Systems, 6th International Conference, ICVS 2008, Santorini, Greece, May 12-15, 2008, Proceedings*, pages 312–322, 2008.

[8] A García Seco de Herrera, Jayashree Kalpathy-Cramer, D Demner Fushman, Sameer Antani, and Henning Müller. Overview of the imageclef 2013 medical tasks. *Working notes of CLEF*, 2013:1–15, 2013.

[9] Alba G Seco de Herrera, Roger Schaer, Dimitrios Markonis, and Henning Müller. Comparing fusion techniques for the imageclef 2013 medical case retrieval task. *Computerized Medical Imaging and Graphics*, 39:46–54, 2015.

[10] Alberto Del Bimbo. *Visual information retrieval*. Morgan and Kaufmann, 1999.

[11] Hugo Jair Escalante, Carlos A Hérnadez, Luis Enrique Sucar, and Manuel Montes. Late fusion of heterogeneous methods for multimedia image retrieval. In *Proceedings of the 1st ACM international conference on Multimedia information retrieval*, pages 172–179. ACM, 2008.

[12] Jianping Fan, Hangzai Luo, and Ahmed K Elmagarmid. Concept-oriented indexing of video databases: toward semantic sensitive retrieval and browsing. *Image Processing, IEEE Transactions on*, 13(7):974–992, 2004.

[13] Matthew B. Hoy. Html5: A new standard for the web. *Medical Reference Services Quarterly*, 30(1):50–55, 2011. PMID: 21271452.

[14] Jing Huang, Ravi Kumar, Mandar Mitra, Wei-Jing Zhu, and Ramin Zabih. Image indexing using color correlograms. In *1997 Conference on Computer Vision and Pattern Recognition (CVPR '97), June 17-19, 1997, San Juan, Puerto Rico*, pages 762–768, 1997.

[15] Chryssanthi Iakovidou, Nektarios Anagnostopoulos, Athanasios Ch Kapoutsis, Yiannis Boutalis, and Savvas A Chatzichristofis. Searching images with MPEG-7 (& mpeg-7-like) powered localized descriptors: the SIMPLE answer to effective content based image retrieval. In *12th International Workshop on Content-Based Multimedia Indexing (CBMI)*, pages 1–6. IEEE, 2014.

[16] Charles E Kahn Jr and Cheng Thao. Goldminer: a radiology image search engine. *American Journal of Roentgenology*, 188(6):1475–1478, 2007.

[17] Jayashree Kalpathy-Cramer, Alba García Seco de Herrera, Dina Demner-Fushman, Sameer Antani, Steven Bedrick, and Henning Müller. Evaluating performance of biomedical image retrieval systems– an overview of the medical image retrieval task at ImageCLEF 2004–2014. *Computerized Medical Imaging and Graphics*, 2014.

[18] Jayashree Kalpathy-Cramer and William Hersh. Multimodal medical image retrieval: image categorization to improve search precision. In *Proceedings of the international conference on Multimedia information retrieval*, pages 165–174. ACM, 2010.

[19] Martha Larson, Bogdan Ionescu, Xavier Anguera, Maria Eskevich, Pavel Korshunov, Markus Schedl, Mohammad Soleymani, Georgios Petkos, Richard Sutcliffe, Jaeyoung Choi, and Gareth J.F. Jones, editors. *MediaEval 2014 Multimedia Benchmark Workshop*, volume 1263 of *CEUR Workshop Proceedings*, Oct. 2014.

[20] Mathias Lux. LIRE: Open source image retrieval in java. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 843–846. ACM, 2013.

[21] André Mourão, Flávio Martins, and João Magalhães. Multimodal medical information retrieval with unsupervised rank fusion. *Computerized Medical Imaging and Graphics*, 39:35–45, 2015.

[22] Henning Müller, Paul Clough, Thomas Deselaers, and Barbara Caputo. *ImageCLEF: Experimental Evaluation in Visual Information Retrieval*. Springer Publishing Company, Incorporated, 1st edition, 2010.

[23] Duong-Trung-Dung Nguyen, Mukesh Saini, Vu-Thanh Nguyen, and Wei Tsang Ooi. Jiku director: A mobile video mashup system. In *Proceedings of the 21st ACM International Conference on Multimedia*, MM '13, pages 477–478, New York, NY, USA, 2013. ACM.

[24] Okan Ozturkmenoglu, Nefise Meltem Ceylan, and Adil Alpkocak. Demir at imageclefmed 2013: The effects of modality classification to information retrieval. *Working Notes of CLEF*, 2013.

[25] Gwénolé Quellec, Mathieu Lamard, Guy Cazuguel, Béatrice Cochener, and Christian Roux. Wavelet optimization for content-based image retrieval in medical databases. *Medical image analysis*, 14(2):227–241, 2010.

[26] Gwénolé Quellec, Mathieu Lamard, Guy Cazuguel, Zakarya Droueche, Christian Roux, and Béatrice Cochener. Real-time retrieval of similar videos with application to computer-aided retinal surgery. In *Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE*, pages 4465–4468. IEEE, 2011.

[27] Md Mahmudur Rahman, Sameer K Antani, and George R Thoma. A learning-based similarity fusion and filtering approach for biomedical image retrieval using svm classification and relevance feedback. *Information Technology in Biomedicine, IEEE Transactions on*, 15(4):640–646, 2011.

[28] Mukesh Saini, Seshadri Padmanabha Venkatagiri, Wei Tsang Ooi, and Mun Choon Chan. The jiku mobile video dataset. In *Proceedings of the 4th ACM Multimedia Systems Conference*, MMSys '13, pages 108–113, New York, NY, USA, 2013. ACM.

[29] P. Seshadri, M. Chan, W. Ooi, and J. Chiam. On demand retrieval of crowdsourced mobile video. *Sensors Journal, IEEE*, PP(99):1–1, 2014.

[30] Thomas Sikora. The mpeg-4 video standard verification model. *Circuits and Systems for Video Technology, IEEE Transactions on*, 7(1):19–31, 1997.

[31] Josef Sivic and Andrew Zisserman. Video google: A text retrieval approach to object matching in videos. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pages 1470–1477. IEEE, 2003.

[32] Alan F. Smeaton, Paul Over, and Wessel Kraaij. Evaluation campaigns and trecvid. In *MIR '06: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*, pages 321–330, New York, NY, USA, 2006. ACM Press.

[33] Cees GM Snoek, Marcel Worring, and Arnold WM Smeulders. Early versus late fusion in semantic video analysis. In *Proceedings of the 13th annual ACM international conference on Multimedia*, pages 399–402. ACM, 2005.