UNIVERSIDAD AUTONOMA DE MADRID

ESCUELA POLITECNICA SUPERIOR

SUMMARIZATION OF SURVEILLANCE VIDEOS BASED ON VISUAL ACTIVITY

-PROYECTO DE FIN DE CARRERA-

Virginia Fernández Arguedas

JUNIO 2008

SUMMARIZATION OF SURVEILLANCE VIDEOS BASED ON VISUAL ACTIVITY

-PROYECTO DE FIN DE CARRERA-

AUTOR: Virginia Fernández Arguedas
TUTOR: Ebroul Izquierdo
PONENTE: José María Martínez Sánchez

Grupo de Tratamiento de Imágenes
Dpto. de Ingeniería Informática
Escuela Politécnica Superior
Universidad Autónoma de Madrid
Junio de 2008

-PROYECTO DE FIN DE CARRERA-

Título: *Summarization of surveillance videos based on visual activity*

Autor: Dª Virginia Fernández Arguedas

Tutor: D. Ebroul Izquierdo

Tribunal:

Presidente: Jesús Bescós Cano

Vocal: Javier Ortega García

Vocal secretario: José María Martínez Sánchez

Fecha de lectura:

Calificación:

Abstract:

Nowadays, surveillance systems are all over the world. The enormous growth on the number of deployed CCTV cameras and, consequently, the huge growth of the number of recorded hours of video sequences are generating the necessity of a method that reduces the amount of this information. Summarization appears as a possible solution to these problems.

This master thesis proposes two different methods to summarize surveillance video sequences based on motion activity: thresholding between following frames and thresholding between a background and the frames. The former consists on an estimation of the energy of the difference between following frames. Considering temporal redundancy between following frames exists, only the frames whose differential energy is bigger than a fixed threshold would be saved in the summary. Given that they will be the only ones that would add some relevant information. While the latter, firstly, calculates a reconstructed background in order to consider only the relevant information (the foreground) in the frames. Later it will also calculate the energy of the difference (between the reconstructed background and the current frame). Only those frames whose differential energy overpasses the fixed threshold will be saved, as they hold relevant information. Both approaches depend on a fixed threshold, its definition can vary the quality of the summaries as well as the amount of relevant information they would contain. These approaches try to summarize videos automatically in order to reduce the necessary resources (to store or check information) like for example bit rate. The amount of redundant information in a surveillance video sequence usually is really considerably compared with the amount of useful information. In order to summary, these approaches try to take advantage of this special feature of surveillance videos.

Keywords:

Video summarization, CCTV, surveillance video, video, background reconstruction.

Resumen:

Actualmente, los sistemas de seguridad están repartidos alrededor de todo el mundo. El enorme desarrollo y expansión de las cámaras CCTV y, en consecuencia, el gran aumento del número de horas de video grabadas generan la necesidad del desarrollo de un método que reduzca dicha cantidad de información. La sumarización aparece como una posible solución a estos problemas.

Este Proyecto de Fin de Carrera propone dos métodos distintos para sumarizar videos de seguridad basados en el movimiento: aplicar un umbral entre imágenes contiguas y aplicar un umbral entre una imagen y un background calculado. El primer método consiste en el cálculo de la energía de la diferencia entre imágenes contiguas. Si tenemos en cuenta la redundancia temporal existente entre imágenes vecinas, solo las imágenes cuya energía de la diferencia sobrepase un umbral se guardaran en el resumen. Dado que dichas imágenes son las únicas que añadirían información relevante. Por otro lado, el segundo método reconstruye un background genérico para la secuencia de video, de modo que sólo se tenga que considerar la parte de las imágenes que contenga información relevante. Para ello calculamos la diferencia entre cada imagen y el background reconstruido, así solo analizaremos los cambios de la parte de la imagen que va a ir variando. Después hallaremos la energía de la diferencia entre una imagen y el background reconstruido. Únicamente, aquellas imágenes cuya energía de la diferencia sobrepase el umbral fijado serán guardadas porque contienen información relevante. Ambos métodos dependen del umbral fijado, el valor determinado hará variar la calidad de los resúmenes, así como la cantidad de información relevante que se graba. Estos métodos tratan de resumir videos automáticamente para reducir la cantidad de recursos necesarios, como por ejemplo la tasa binaria. Por último, la cantidad de información redundante en un video de seguridad normalmente es mucho mayor que la cantidad de información útil. Ésta es la característica que tratan de explotar todos los métodos de sumarización de videos de seguridad.

Palabras clave:
Sumarización de videos, CCTV, videos de seguridad, video, reconstrucción de background.

Este proyecto ha sido parcialmente financiado por la cátedra Infoglobal-UAM para “Nuevas tecnologías de video aplicadas a sistemas de seguridad”

Acknowledgements

I would like to thank Prof. Ebroul Izquierdo, from Queen Mary University of London, for giving me this opportunity, assisting and advising me during my period in London. I will be ever thankful to him for this opportunity. I also would like to express my greatest gratitude to Prof. José María Martínez Sánchez not only for his always helpful advices, but also for his encouragement regarding my PFC in London.

Many thanks to all my friends...to all the friends I made at QMUL, at UAM or the ones I met by chance during my life. For all your contributions and friendship, thanks!

I would also like to thank all my professors and mates that have been helping me along the degree, especially to Jesús Bescos, that always was there to help or to give all of us a good advice.

I am especially grateful to my family that has always supported me in everything I wanted to do. Their support and encouragement along my degree has been really important for me.

Finalmente, muchas gracias a todos por enseñarme lo que sé, darme lo que tengo y quererme y apoyarme sin límites.

Virginia Fernández Arguedas
Junio 2008

CONTENTS

Figures

Tables

Table 11: Most important multimedia compression formats accepted in FFMPEG

1 Introduction

1.1 Motivation

Nowadays everybody is really concerned about security. Recent events, crime, terrorism, … have increased people's worries. This increment has caused the logical appearance of some questions: who is safe, where we are safe and what gives us this security. The recent events have determined that these questions, which in normal situations would not be difficult to answer or we would not have even wondered about them, seem to be complex and of a great importance.

The rise of concern in people's mind generates the necessity of an increment of security, above all in big cities or really crowded places as airports, train stations, bus stations, shop moles… Governments want their citizens to feel safe. For that reason, the number of security systems has increased extremely fast, creating a huge net of CCTV systems all around the world. Consequently, an enormous amount of data is generated everyday all over the world. The quantity is so big that there are not enough available resources to analyse it. So, usually the information is not checked unless something important happens.

Millions of surveillance cameras have been installed, producing, each of them, a big data amount. Moreover, the bigger part of this data is useless. It has no interesting information in it. Usually it makes no sense to take care of it, so it can be removed without any relevant information loss.

When something happens in a city, an airport, a train station, in summary, in a crowded place, the amount of data that must be analysed is so big that takes too long time to check it all. Furthermore, it would be a waste of time and a waste of resources to check the whole produced data amount. That is why summarization appears as the solution.

The summarization objective is to reduce the data amount. It can be based on different paradigms and it can follow different methods to do it. However, the outcome must be the same, a shorter version of the video sequence. Giving as a result a version that can be analysed in a suitable amount of time.

1.2 Objetives

The main objective of this work is to develop a method to summarize surveillance videos. The chosen criterion to do that is the visual activity. As soon as something moves in the image the program will detect some visual activity. The first aim of the method is to determine if this motion is enough important or not to keep the image in the summary.

The required result is a video sequence, shorter than the original one, which shows all the important events that have taken place in the whole original video sequence.

In order to summarize, first, what an important event is must be defined. Along this project, we will consider an important event to all movements in images that implies a "big" change in it. We will try to reject all the frames that change for different reasons than motion.

1.3 Organization of the report

The report is composed by the following chapters**:**

- **Chapter 1:** introduction, objectives and motivation of the work.
- **Chapter 2:** it is the state of the art. It includes descriptions of surveillance videos, colour spaces ... This chapter also includes a description of summarization, its objectives, different kinds, classification, methods... Moreover, in this chapter surveillance is described and the necessity of summarization in surveillance.
- **Chapter 3:** in this chapter, the methods developed to summarize surveillance video sequences are described and explained.

- **Chapter 4:** in this chapter, the methods previously described are analysed and tested. Some comparisons are also calculated.
- **Chapter 5:** conclusions obtained during the development of the methods. Besides, some possible lines for future work.
- **Appendix A:** tutorial to install the developed software
- **Appendix B:** tutorial for the programmer
- **Appendix C:** Compression
- **Appendix D:** CCTV
- **Appendix E:** Colour spaces and techniques to store images.

2 State of the art

2.1 Analysis of surveillance video sequences

Nowadays, the interest in surveillance in public, military and commercial scenarios is growing up due to the increasing demand of security. For that reason, thousands of video cameras can be found at public places as public transport, banks, and airports ... in every country. This expansion of surveillance cameras all over the world has been possible thanks to the technological development and it has provoked the decrease of the surveillance system prices.

Once the economical problem has been solved another problem appears. This big amount of surveillance cameras produces a huge amount of information which is difficult to work with. As a result, now, the problem is the lack of time to check all this information and also the lack of staff that is necessary to do it.

Previously the information was analyzed and checked by an operator. This person was responsible of taking care of some cameras at the same time working thus during a long period of time. If we take into account that the information included in these videos is significant. Moreover, it can include images of restricted areas. A lack of attention could provoke the appearance of a dangerous situation.

For all these reasons, the development of a system that reduces this amount of information is needed. Furthermore, it is one of the main issues related with security.

The mentioned system must be composed basically by two kinds of techniques: scalability and usability [1]. Another technique that is also needed to handle the amount of information is summarization. The main advantage of summarization is that it reduces the longitude of the surveillance video. As a result, the storage capacity of every system increases.

The main objective of summarization is to skim through the content and view it in differing detail depending on the preference [2]. So “the need for video summarization originates primarily from a viewing time constraint” [3].

However, a shorter version of the original video sequence is desirable in a number of applications, especially when storage, bandwidth and/or power are limited [3], not only when the storage capacity needs to be optimized [4].

2.2 Video monitoring

Video monitoring started as a simple method of black and white video feeds from remote cameras to a central monitoring location. This central location used to be attended by people and the analog recording was performed by analog technologies. This model was the basic start of the current Closed Circuit Television (CCTV).

CCTV became really popular and started offering collection surveillance by analog cameras connected in closed network. This network was composed by multiplexing controllers, monitor TVs and video recorders. All these elements were connected by coaxial cables.

The technical limitations of video surveillance were collection and processing of monitored information. These problems required the adaptation to new technologies in order to match the growing demand for video surveillance.

Historically, video quality was in low resolution as the coaxial cables were limited in bandwidth and distance reach. Moreover, video was stored as an analog signal on magnetic tape. Magnetic tapes had many operational problems, like constant tape changing, cumbersome information retrieval and very limited remote access. However, when hard drives entered in the market as a digital-format replacement for analog video, it became feasible to begin storing and using video in digital form.

Closed Circuit TV (CCTV) is a powerful tool used in a diverse range of applications, nowadays, especially for security. The critical locations where CCTV systems are usually placed are in: airports, train stations, military compounds and airbases and public “hotspots”. Basically, CCTV systems are located in these points because they can be potential threats and requires situational awareness to the possible effective response of the relevant enforcement agencies.

Nowadays, the vast majority of imagery from CCTV security systems is merely recorded for future analysis. To be used in case that an event has occurred. Even then, the analysis may take hours or days to complete.

The difficulties in extracting information from CCTV data arise from a number of factors [5]:

- Events of interest often involve dynamic interactions.
- They are largely application specific.
- They are often hard to define and distinguish from events of no interest.
- Vast quantities of raw data are involved.
- There can be great variations in scenes because of lightings changes. This can occur not only in outdoor scenes because of changing sun angles, time of day and weather, but also indoor scenes with some natural lighting.

It must also be taken into account that the CCTV images are invariably noisy and cameras are poorly maintained. Moreover, CCTV cameras often suffer small amounts vibration through the mounting, combined with image noise and wind effects on trees etc, which causes some features to appear to jitter in position. So usually the image quality is not the best one that can be expected.

2.3 Characteristics of the videos

The characteristics of a video can be classified in <u>three</u> different groups:

- <u>Dominant</u>: these are the characteristics that make the video more powerful than other types of data. A viewer can identify them but they are qualitative characteristics so an algorithm will not be able to appreciate them.

- <u>Native</u>: these are the characteristics that are natural on videos. The kind of characteristics that are own of the videos and they do not appear as a result of the association of the video with other types of data.

- <u>Assigned</u>: these are those characteristics seen of the designer of the video data model. They include the attributes of the high-level video contents, which are identified during the design process, indicate the characteristics assigned to the video data.

Nowadays, “all video analysis make exhaustive use of the assigned characteristics; many [6] but practically none have taken note of the dominant characteristics” [7].

2.3.1 Classification of characteristics

As it has been shown the most important **<u>high level</u>** characteristics of a video are classified in three groups, each group describes a different nature in the characteristic.

Before describing each characteristic, a video object must be defined. A video object is a sequence of video frames that have their own attributes and attribute values describing the content.

2.3.1.1 Dominant characteristics

- Richness of information content: each video object contains a meaningful scene. If it is analyzed, we can gain a lot of information about the theme, which implies that our knowledge about the topic will increase.

- Rapid context switching: the context of a scene is formed by the idea, situation, events and other background information related to it. It duty is to help a viewer to understand the scene. An important characteristic of videos is the rapid change in context; it can be appreciated by a viewer but not by an algorithm.

- Spontaneous reaction content: the action and the consequent reactions of some elements of a video object make the video presentation lively. It would be easy for a viewer to know if an action that is taking place in the video is a reaction (provoked by a different action) or not. However, it would be impossible for an algorithm to realize it.

These characteristics make the video data more powerful than other kind of data as text data or even voice data.

2.3.1.2 Native characteristics

- Video data is multidimensional: the contents of a conventional text data are one-dimensional (for example: integers, strings, real numbers, etc). Whereas one-dimensional variables are not enough to work with video data which is more complex. This complexity comes from the fact that video data depends on other dimensions, as temporal and spatial, that add more information to the video.

- Video object can be complex objects that are composed by smaller individual objects: a video object consists of many video elements which are of interest to the users. This characteristic provokes the necessity of a hierarchy among video elements. A consequence of this characteristic is that the analysis of the video object must be done in a low level analysis, which means that must be done element by element.
- The accurate representation of the extracted contents may not always be possible: due to the size or the quality of the image sometimes, the contents cannot be represented properly; they are substituted by similar shape (transcoded form). A solution to recognize the objects in their transcoded form is to uses annotations.

The native characteristics should influence the video analysis algorithm.

2.3.1.3 Assigned characteristics

There are four different kind of assigned characteristics associated with the video object at its elements level [8]:

- Purpose: the explanation for the presence of the video element in the data model.

- Defining property: the observable characteristics of the video element can be inherited by all its specialized characteristics.

- Effect on the information system: the observable effect of the action of the video elements on other video elements in other category classes.

- Behavioural pattern: the observable pattern of the behaviours of different video element categories.

2.4 Surveillance videos

Video surveillance, understood as the simplest model, began with a simple closed circuit television monitoring. Early, in 1965, it started being used by the police in public places and these cameras were controlled by officers at all times. At the present time, people are being watched by surveillance cameras almost everywhere. Moreover, surveillance cameras are in places like shopping malls, sporting events, schools and places of employment

Video surveillance systems are one of the main sources of information and security thanks to their wide-spread and increasing presence in all countries. However, the adopted closed circuit devices used in the surveillance systems are often affected by poor quality.

In some cases, even if the images have a poor quality, they can give useful information.

The images and video sequences that come from video surveillance systems need to be digitalized in order to be processed by dedicated software to enhance features useful for its analysis. Generally, when these images are analysed, the existing corruptions tends to disappear or at least they are reduced.

It must be said that the corruption in the images can be regarded as a result of the processing of the images, their storing or simply because of the bad quality of the cameras used for the surveillance systems. Finally, these corruptions put limits to the surveillance systems.

When we work with surveillance videos or images we must deal with some common problems coming from the use of poor quality devices in the surveillance systems, as for example the followings[9]:

- Low resolution of the images (that usually implies the need to increase the size of the interesting details).
- Lack of contrast.
- Different types of noise or disturbances.
- Blurring caused by motion or lack of focus.
- Geometric distortions (that limits the reconstruction of the objects inside the image).

As it has been mentioned before, surveillance systems are basically installed in crowded places or strategic places, for example, in shopping malls, airports, ... Typically these places are huge and they must be covered by lots of surveillance video cameras what produce a large amount of information. In order to obtain mainly useful information, the surveillance system must be robust, above all for avoiding false alarms. False alarms are called to events that do not imply any danger or worry. For instance, in terms of surveillance video systems we can consider possible false alarms the following events [10]:

- Moving trees.
- Rain.

- Small camera motion (caused by some external-system event).
- Varying illumination conditions.

Video surveillance systems can be divided into three main categories [11]:

- Operator-controlled video surveillance.
- Basic automated video surveillance.
- Smart video surveillance.

Operator-controlled video surveillance [12] [13]:

A basic CCTV video surveillance system can be considered as an operator-controlled video surveillance system. It, basically, consists on a collection of video cameras (mounted in fixed positions) that cover a circumscribed area defined by the fields of view of the used video cameras. Later, the video streams are transmitted to a central location where the video stream will be displayed on one or several video monitors and it will be recorded. The main characteristic of this surveillance video system is that all the displayed information will be observed by the person in charge. This person will then determine if there is going on an activity that requires a response.

Basic automated video surveillance:

Automated video surveillance systems try to reduce the burden on the user by employing video motion detectors to determine where there is motion in a given scene.

As a result, this kind of systems can remove some false alarms without the necessity of someone controlling the video. However, the problem is that a large amount of information related to motion is not relevant, so it can be removed. In order to remove this information and to extract the relevant information from the video sequence a smart video surveillance system is necessary.

Smart video surveillance:

This kind of surveillance systems achieves more than motion detection. Its common objectives are to detect, classify, track, localize and interpret behaviours of objects of interest in the environment. Many of the systems also classify activity by analyzing object actions and object interactions.

Usually the typical objective of these surveillance systems is to interpret activity in real time and present a clear picture of the activity to the user.

Surveillance systems have gradually become more independent, this is to say they barely require some human intervention. Consequently the path for full automation of this process is being paved. This is the idea over which the previous classification was based on. Another classification could be in which commercial applications each system could be used.

2.5 Summarization

Nowadays, video summarization is one of the most promising techniques. Its objective is to create a short version of the original video sequence or a subset of key frames which contains as much information as possible as the original video sequence. Moreover, a video summary can have varying amount of detail depending on the requirements in a specific problem.

The ideal summary presents the most interesting and important aspects of the video sequences with the minimal redundancy. For that reason, all the techniques applied to summarize videos must be focused on looking for the existing redundancy (temporal and spatial redundancy). Moreover, the aim of these techniques is to remove all the existing redundancy without loosing any important information (not redundant data).

Actually, the most promising techniques are not summarizing techniques if not automatic summarizing techniques. Their objective is to create a short version or subset of key frames which contains as much information as possible from the original video sequence. Nowadays, the research in summarization area is based on automatic summarization because manual intervention is usually tedious and it usually causes delays. So if an automatic summarization technique is applied, some information about the content of a large video will be provided faster.

The properties of a video summary depend usually on the application domain, the characteristics of the sequences to be summarized and the purpose of the summary. Generally, summarization techniques try to eliminate redundant or similar frames. For example, sometimes in order to remove redundant frames, the summarization technique keeps one similar frame which represents a set of frames similar to the key frame (the kept frame).

One of the disadvantages of the summarization techniques is that in order to reduce the amount of information it removes some frames. This fact takes place taking into account the amount of existing redundant information. As a result, the output video sequence probably has lost the continuity that existed in the original video. So, basically, the outcome will seem like a fast-forward video with fast changes between following frames. There are some techniques to avoid this fact as the one in [14].

Types of video summaries:

The techniques for automatic video summarization can be classified basically in two approaches: static storyboard summary and dynamic video skimming. The main difference between these approaches is that the former is a collection of static key frames of video shots, while the latter is a shorter version of a video composed of a series of selected video clips. [15].

In one hand, static storyboard allows non-linear browsing of video content by sacrificing the temporal evolution of a video.

In the other hand, dynamic video skimming preserves the time-evolving nature of a video by linearly and continuously browsing certain portions of video content depending on a given time length. The conclusion is that the skim is the smallest comprehensible video representation of the original segment. [16]

In contrast, if we think about what a summary is going to content, then we can have a different summary's classification. Then we will distinguish between: video skim, video highlights and multimedia video summary.

"Video skim" is a temporally condensed form of the video stream that preferably preserves the most important information. It is a set of short video sequences composed of automatically selected portions of the original video. During the literature, some authors have used terms like "preview" and "trailer" as synonyms of "video skim". However, there is a slight difference among them. A trailer has a commercial perspective, so it contains special highlights with a purpose of attracting audience. For instance, a preview conveys key aspects of a program to allow users to quickly see what it is about.

"Video highlights" is a form of summary that aims at including the most important events in the video.

Finally, "Multimedia video summary" is a collection of audio, visual and text segments that preserve the essence and the structure of the underlying video.

Video is a rich medium and it can be classified into a large number of types and genres, for example, TV programs, produced film, home videos, educational videos, multimedia presentations, etc. Each type of video has a typical usage and specific characteristics.

<u>Classification of the summarization methods:</u>

A possible classification for the summarization approaches would be [17]:

- Usage.
- Content.
- Method.

Usage:

Video summarization appeared to solve a typical users' problem, how to handle video more efficiently (for example, to use fewer resources by doing the same). There are three related aspects to the manner in which summaries are used: selection, consumption and production.

Firstly, video summaries can have a "selection" application. It will provide a condensed and descriptive form of a vide content that will help users to select among a large collection of items or video sequences.

Secondly, summaries can also be generated for more efficient consumption of content. This application tries to improve the use of store, time, and other resources by reducing the duration of the video sequence.

Thirdly, summaries can be created from scratch or by reusing existing material. The former is a kind of producing new summaries called "creation", while the latter is called "repurposing".

The usage of a summary depends on the goal users set in performing a specific task: known and unknown intention. A user with a known goal in consuming video summaries will have high requirements. Meanwhile, a user with an unknown goal will not pose strict requirements. According with this, we can classify the summaries in two groups. The first one, it will include summaries that will satisfy specific users, while the second one will content these summaries that will produce less specific results.

The usage of video and summaries changes drastically depending on the context. Professionals have completely different requirements than home users or typical consumers.

Content:

Current summarization techniques can be classified based on the characteristics of the content they can process.

- Application domain:

  Automatic video summarization techniques can be classified depending on their specialization or applicability to specific domains. The most common application domains are surveillance, television, home videos, video presentations,

- Genre:

  Genre- dependent summarization techniques exploit specific structures and properties of a particular genre. The most typical genres are: news, sport, sitcoms, talk shows and feature films.

  A few systems claim to be genre- independent and provide video summaries regardless of the type of program.

- Levels of editing and structure:

  Content is produced at different levels of editing and contains different levels of structure.

Method:

Another sub classification can be defined with respect to the source of information that are considered (video, audio, text, ...), to the type of algorithm that is employed, to the output that is produced and if the summary fulfil the objectives or not. According to this, the method criteria can be classified in: source, algorithm, output and personalization.

3 Design and development

3.1 Objective. Summarization of surveillance videos

In this chapter, we will explain the two approaches (but there are three sections… the fourth in another appendix) that have been developed in order to summarize surveillance videos. The main objective of this section is to explain how they work and which features of the surveillance videos they take into account to reduce the amount of redundant information existing in the original video sequence.

3.2 Methods

3.2.1 Thresholding: difference between following frames

Theoretical bases of this method:

As we have seen before, there is a big redundancy in images that must be considered in order to compress them (when we want to reduce the bit rate).

Firstly, the spatial redundancy must be taken into account; this is the redundancy that appears between parts of the same image. For example, sometimes the image has large parts that are homogeneous.

Secondly, temporal redundancy must be considered. This last kind of redundancy appears between following frames. That means that there is a part of information that is repeated in following frames. This part usually is the background of the image which is quite strange that it changes. In fact, it is even more atypical if we take into account that we are working with surveillance cameras that strangely will change their location. Things that can cause a change in the background, causing problems in this algorithm (because the algorithm would think that these changes are important events) could be [18]:

- Time of the day: gradual changes in ambient illumination or in the exposure settings of the camera will alter the appearance of the background. These changes should have no effect on what is considered background or foreground.

- Light switch: sudden changes in illumination alter the appearance of the background. However, these changes should have no effect on the differentiation between background and foreground.

- Waving trees: sometimes the background may be in constant motion, but it should not be confused with foreground.

- Camouflage: A foreground object may appear similar to the background, but it should still be segmented as foreground.

- Bootstrapping: Background subtraction modules typically require a training period over which they acquire a model of the background. A training period absent of foreground objects is not available in some environments. All pixels in frames after the training period should nevertheless be labeled correctly as foreground or background.

- Moved objects: Background objects may move. Moved objects should not be considered part of the foreground.

- Sleeping/waking person: A foreground object may become completely motionless. It should always remain part of the foreground. If the background subtraction algorithm is adaptive, then there is the possibility that the background will subsume a sleeping person; upon waking, parts of the waking person that do not exhibit visible change will remain in the background. Again, all of such a “waking” object should be marked foreground.

- Shadows: Foreground objects often cast shadows on the background. Cast shadows should have no effect on background/foreground labeling.

Definition and explanation of the algorithm

Once it has been determined what kind of data can be removed without any information loss. The algorithm must be described according with these reasons.

The first method that has been implemented is the threshold applied to following frames. This method should take advantage of the temporal redundancy existing between following frames. Reducing this redundancy the bit rate will be reduced as well. Consequently a compressed image will be obtained without any information loss.

During the development of this algorithm, it has been assumed that the surveillance cameras are deployed pointing to a fixed place, because this is one of the most common feature of surveillance video cameras. Consequently, one of the basic attribute of these surveillance videos is that the background is almost constant, or shows very smooth differences due to changes in the illumination during long periods of time (or other weather change). This consideration, however, implies that adjacent frames usually include a huge quantity of redundant information. The presence of this redundancy (temporal redundancy) arises as a very valuable property to distinguish more relevant information through a comprehensive study of the changes between frames, and it is closely related with the main target in video summarization.

The target in video summarization is to present, in a synthetic way the content of the video, while preserving the essential message of the original. Formally, a video summary is a kind of video that keeps the relevant parts of a whole video. This notion is quite subjective and it depends on the criteria is going to be used. In this program the criteria will be the value of energy, which is calculated from the difference between adjacent frames.

The objective of this program is to summarize surveillance videos by removing the really similar following frames (this is done by reducing the temporal redundancy). This program will keep the information identified as high-value data in the video stream and it will present a short version made up of these portions of high-value information while the less important parts of the video are removed.

The algorithm comprises three steps that can be grouped into two different stages. In the first stage, the “difference frame” and its energy are calculated for each frame. In the second stage, the frames are classified and the reference frame is refreshed.

Algorithm steps are described in the following:

Step1: *Calculation of the “difference frame”.*

The “difference frame” is calculated for every frame as the difference between the current frame and the reference frame and taking the absolute value.

Step2: *Calculation of the energy of the “difference frame”.*

Once the “difference frame” has been calculated, its energy value is founded using the following expression, where x(i,j) is the differential intensity of the pixels belonging to the “difference frame”:

$$\frac{\sum_{i=0}^{N-1}\sum_{j=0}^{M-1} x(i,j)}{N*M}$$

Step3: *Classify the frames using a threshold.*

In this step the frames are classified in two groups using a fixed threshold. One of the groups is composed by frames whose energy value (previously calculated) exceeds the threshold value while the other is formed by those frames whose energy value do not exceed the threshold. When one of the energy values exceeds the threshold, its corresponding frame becomes the reference frame for the next iteration. As a result, the program only takes into account those frames that experiment a great variation compared with the previous reference.

The frame classification procedure is conducted as it is shown in the next figure:

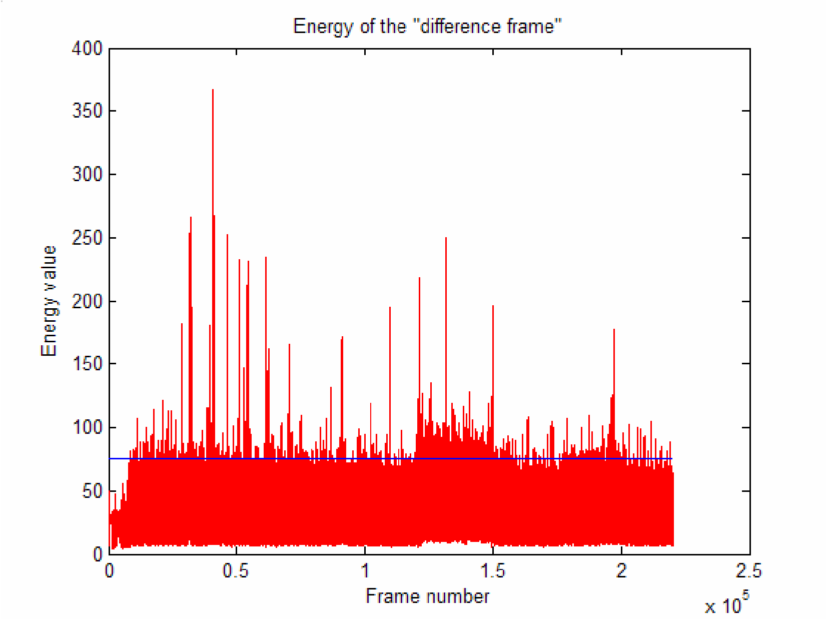


Figure 1: energy of the difference between following frames values and thresholding

At the end the program designed to evaluate the performance of this algorithm will return two files. The first one contains all the energy values of every difference frames. The second one and most important is the file that contains the group of frames whose "difference frame" has exceeded the fixed threshold. This group is the summary of the surveillance video for that threshold.

The number of frames that would compose the summarized video sequence will depend on the chosen threshold value. If it is really high the number of frames will be small, but the frames will be more representative. Meanwhile if the threshold value is too small, too many frames will overpass it and the whole process will be useless. For these reasons, the selection of the threshold value must be done with a lot of care.

This algorithm has been described in the paper called "Event detection and clustering for surveillance video summarization". [19].

3.2.2 Construction of the background: background reconstruction

Theoretical bases of this method:

The main purpose of this algorithm/method is to create a new background, a kind of average background from the video sequence or from only a part of the video sequence. The basic idea is to create a new background from the surveillance video sequence that would make the “program of summarization by thresholding” more stable to external factors as light changes, defects on the image, quality of the video, …. One idea to increase the stability is to obtain a background that is included in all the frames of a whole video sequence (or only included in a significant number of frames of the video sequence). This is the purpose of this program, to keep the most frequent background.

One of the main characteristics of a surveillance video sequence is that the background is usually quite constant. This kind of systems consists on a stationary camera (one or more stationary cameras) with a fixed focal length. Therefore, the background that appears in the different frames of the surveillance sequence is motionless.

Due to the existence of this characteristic in the surveillance videos, the background can be removed from the different frames. So in each image we would have only the useful information. Moreover, it is this information which must be kept on mind to know if the frame is relevant or not.

In this program the assumption is that the background would be the most often observed part over the surveillance sequence. According to this assumption, basically, this program makes a background reconstruction based on “Pixel intensity classification (PIC)”.

Definition and explanation of the algorithm

Considering that the background must be composed by these parts of the image that are the most visible in the surveillance video sequence, we will look for the pixels that appear with the maximum frequency in the image sequence.

If the most frequent pixels are looked for, that means that we are considering that the intensity of the background is stable for a long period. So we will choose these pixels that appear in the longest stable interval in order to take part in the background image.

In this program it is assumed that the foreground is going to change (it is not going to be stable for a long period of time), so it will not cause any confusion between background and foreground. In case that the foreground is stable for a long time, the program can consider the foreground as part as the background which will be a bad conclusion. However, even if this takes place, it will not cause any problem in the summarization algorithm. Because this foreground will not be representative as it is stable and it must not be included in the summary.

The algorithm comprises four steps grouped into two stages. In the first stage, the intensity stable intervals are located. In the second stage, the average intensity value of each intensity stable interval is calculated. Moreover, during the second stage the intervals with similar intensity values will be joined together in the same group.

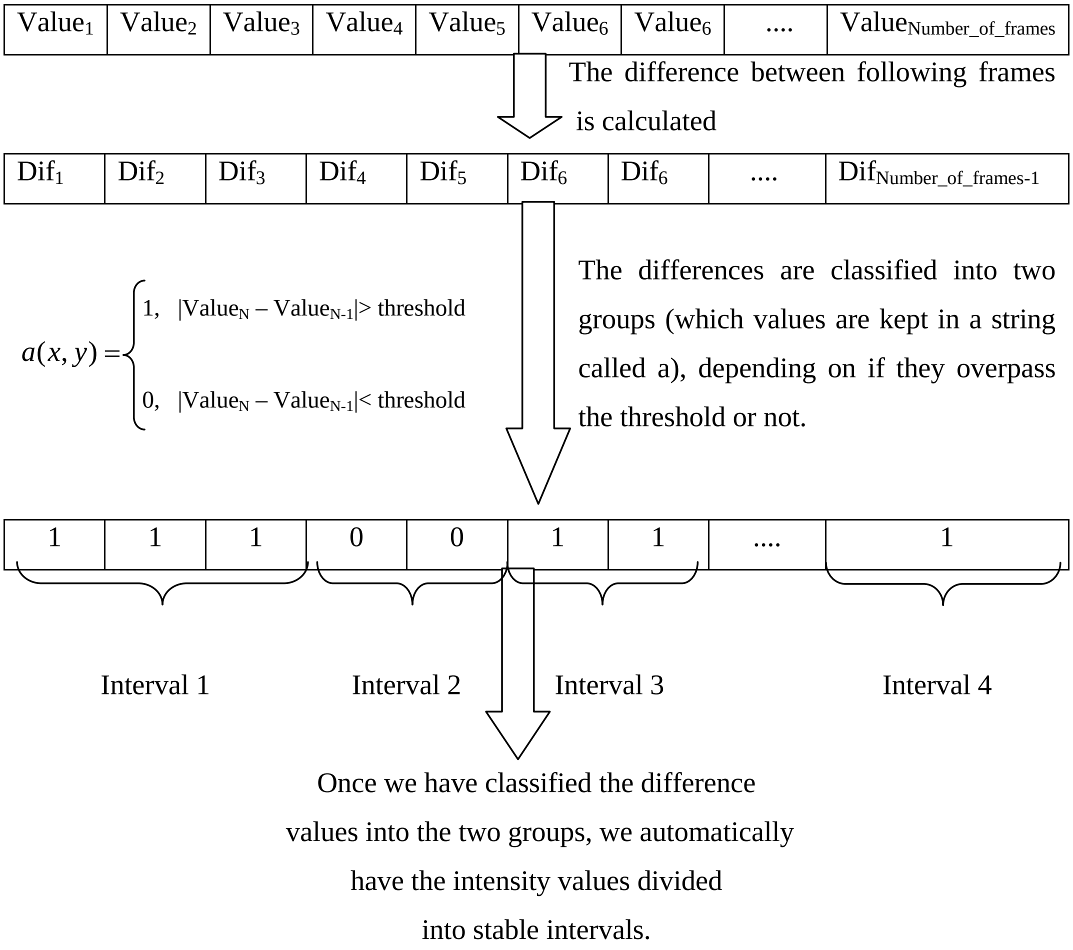
The main objective of this program is to look for the most frequent intensity value to each pixel, because it will be selected to the background frame.

This background image is worked out by calculating the average of all the frames of the surveillance video. The process followed in order to obtain this background frame (reference frame) is composed by the next steps:

Step1: *Classify of the intensity stable intervals.*

The idea consists on analysing the intensity value of each pixel along the time. For that reason every selected intensity value for a pixel is independent from the others pixel's values and it is also analysed independently.

During this step, all the intensity values of a pixel along the time are saved into a string (in c++). When all these values are available, the difference between following intensity values will be calculated. If this difference is bigger than a fixed threshold, the intensity value will receive a flag with value 1, while if it is smaller it will received the flag with value 0. This step consists only in classifying the difference between adjacent intensity values (adjacent on time) of the same pixel. In order to classify the differences into stable intervals, we must consider if the adjacent "difference values" are really similar or not.



Value_1	Value_2	Value_3	Value_4	Value_5	Value_6	Value_6	....	$\text{Value}_{\text{Number\_of\_frames}}$

The difference between following frames is calculated

Dif_1	Dif_2	Dif_3	Dif_4	Dif_5	Dif_6	Dif_6	....	$\text{Dif}_{\text{Number\_of\_frames-1}}$

$$a(x,y) = \begin{cases} 1, & |\text{Value}_N - \text{Value}_{N-1}| > \text{threshold} \\ 0, & |\text{Value}_N - \text{Value}_{N-1}| < \text{threshold} \end{cases}$$

The differences are classified into two groups (which values are kept in a string called a), depending on if they overpass the threshold or not.

1	1	1	0	0	1	1	....	1

Once we have classified the difference values into the two groups, we automatically have the intensity values divided into stable intervals.

Figure 2: process of classification of the energy of the difference values into intervals

The aim of this step is to classify the intensity values of each pixel (independently) in stable intervals along the time. So, at the end of this step, each pixel's intensity value is part of an interval. The common feature of all the pixels from the same interval is that all of them have more or less the same intensity value (because the difference between following intensity values must be less than a fixed threshold).

Step 2: *calculate average intensity of each intensity stable interval.*

The main aim of this step consists on calculating the average intensity value for every single stable interval. As a result, we will have as many average intensity values as the number of different existing intervals for the current pixel.

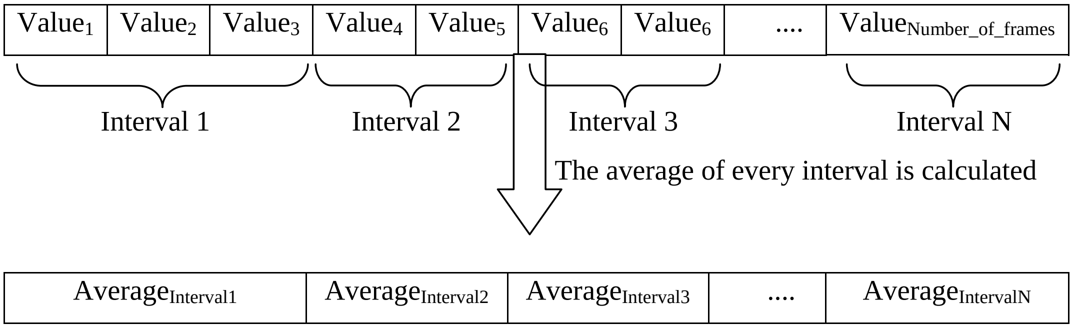
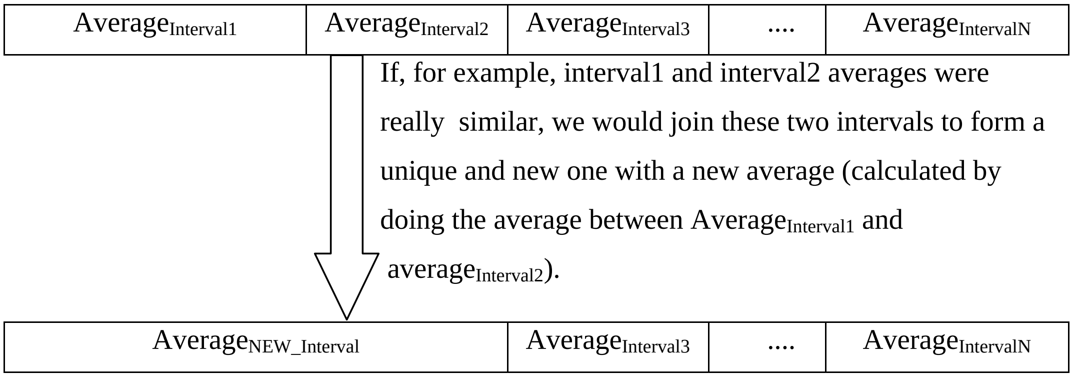


Figure 3: Estimation of the average intensity of each intensity stable interval

Step 3: *classify intensity stable intervals with close average intensity values as the same group named intensity homogenous interval, and tale count of pixels of intensity homogenous interval.*

Once the intensity averages for each interval are already calculated, the next step would be to join stable intervals that have really similar averages. In order to do that, we must determine where the limit between is similar or not. We will do it by establishing another threshold.

If two or more intervals have a similar average, they will be joined together to form a unique interval. After joining the intervals, we must work out the real average for the new interval which will be the average of all the averages of the joining intervals.



$\text{Average}_{\text{Interval1}}$	$\text{Average}_{\text{Interval2}}$	$\text{Average}_{\text{Interval3}}$	....	$\text{Average}_{\text{IntervalN}}$

If, for example, interval1 and interval2 averages were really similar, we would join these two intervals to form a unique and new one with a new average (calculated by doing the average between $\text{Average}_{\text{Interval1}}$ and $\text{average}_{\text{Interval2}}$).

$\text{Average}_{\text{NEW\_Interval}}$	$\text{Average}_{\text{Interval3}}$	....	$\text{Average}_{\text{IntervalN}}$

Figure 4: union of intervals with similar averages and new estimation of the interval average

Step 4: *choose intensity value with the maximum pixel number of intensity homogenous interval as background intensity value.*

Once all the intervals of similar intensity average have been joined, the number of pixels that compose each interval must be counted. So, at the end, we select the interval that is composed for a bigger group of pixels as the proper value for the background frame.

Once all these steps have taken place, we obtain the next background frame: (also a comparison with a "slide show" may help to see the "quality of the results")

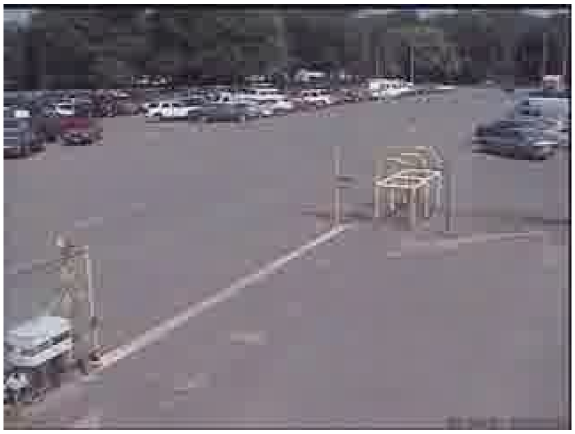


Figure 5: new background calculated applying background reconstruction

It must be taken into account that this “background frame” has been calculated using only 100 frames from the whole surveillance video sequence. The whole sequence could have been used but it would have taken longer time and the background could not have been so clear because the temporal dependence among frames loose its accuracy with time.

As future work it can be proposed a program which calculate a different background frame (using this method) every 100 frames, more or less, in order to adapt the background to the sequence and to obtain better results.

This process has been described previously in a paper called *“A background reconstruction algorithm based on pixel intensity classification in remote video surveillance system” [20]*.

The goal of this method is to calculate the “background frame” so, once we have calculated, it is necessary to use another program that calculate the summary using this new frame. This program is described in the following section.

3.2.3 Thresholding: difference between each frame and a calculated background

Theoretical bases of this method:

This method, as well as the first one, is based on the redundancy (spatial and temporal redundancy) among frames in order to compress reduce the amount of existing data and, for instance, the bit rate.

Spatial redundancy is important in terms of compression algorithms. However, to develop a summarization system we must focus on reducing the temporal redundancy. The temporal redundancy is really big between following frames, but it is still big among frames close in time. If there is some movement in the image, it will take some frames in happening. Surveillance video cameras do not change their location and this would be the only way of having a change of plane that would cause a change strict change in the image (not a progressive change).

Definition and explanation of the algorithm:

The theoretical basis on which this algorithm is based are the same that defined the first method: surveillance video cameras are located in a fixed position, temporal redundancy allows to remove some data (redundant data) without a loss of information, background is almost constant during the video sequence, changes in the video are progressive and smooth, … All these features allows to remove frames that only has redundant data without damaging the video sequence and without any loss of information.

The biggest distinction between this method and the first one is that in this program the temporal redundancy is going to be used in long term, so we are not going to calculate only if the following frames are similar or not. Now the difference is going to be calculated between each frame of the surveillance video sequence and the built background (this background has been built applying the previous section method). As a result, we are not going to save only some independent frames that suppose a big change in the video sequence. With this algorithm, all the frames that take part in a progressive change in the video sequence will be saved in the summary sequence. For that reason, we will not see only one frame related with this change; we will see the whole sequence that represents this change.

Algorithm steps are described in the following:

Step 1: *Procurement of the background frame.*

The main objective of this step is to load the background frame which is going to be used as a reference image. This aforementioned background frame was calculated previously using the background reconstruction method.

This step takes place only once during the whole program.

Step 2: *Calculation of the difference*

At the beginning of this step, the difference frame will be worked out. It will be calculated by subtracting the value of the current frame of the video sequence to the background frame's intensity value of the same pixel.

Step 3: *Calculation of the energy of the "difference frame".*

Once the "difference frame" has been calculated, its energy value is found using the following expression, where x(i,j) is the differential intensity value of the pixels belonging to the "difference frame" (calculated in the previous section):

$$\frac{\sum_{i=0}^{N-1}\sum_{j=0}^{M-1} x(i,j)}{N * M}$$

Step 4: *Classify the frames using a threshold.*

In this step the frames are classified in two groups using a fixed threshold. One of the groups is composed by frames whose energy value (previously calculated) exceeds the threshold value. While the other is formed by those frames whose energy value does not exceed the threshold. When one of the energy values exceeds the threshold, its corresponding frame is saved into the summary video sequence.

3.2.4 FFMPEG

In this PFC, FFMPEG has been used to convert videos from the compressed format (MPEG) to an uncompressed format. Any uncompressed format could be used. Although the one that was chosen was YUV4:2:0.

After applying FFMPEG to the compressed video sequences, we obtain an uncompressed video sequence. Afterwards it will be used to apply all the programs/methods that have been explained before.

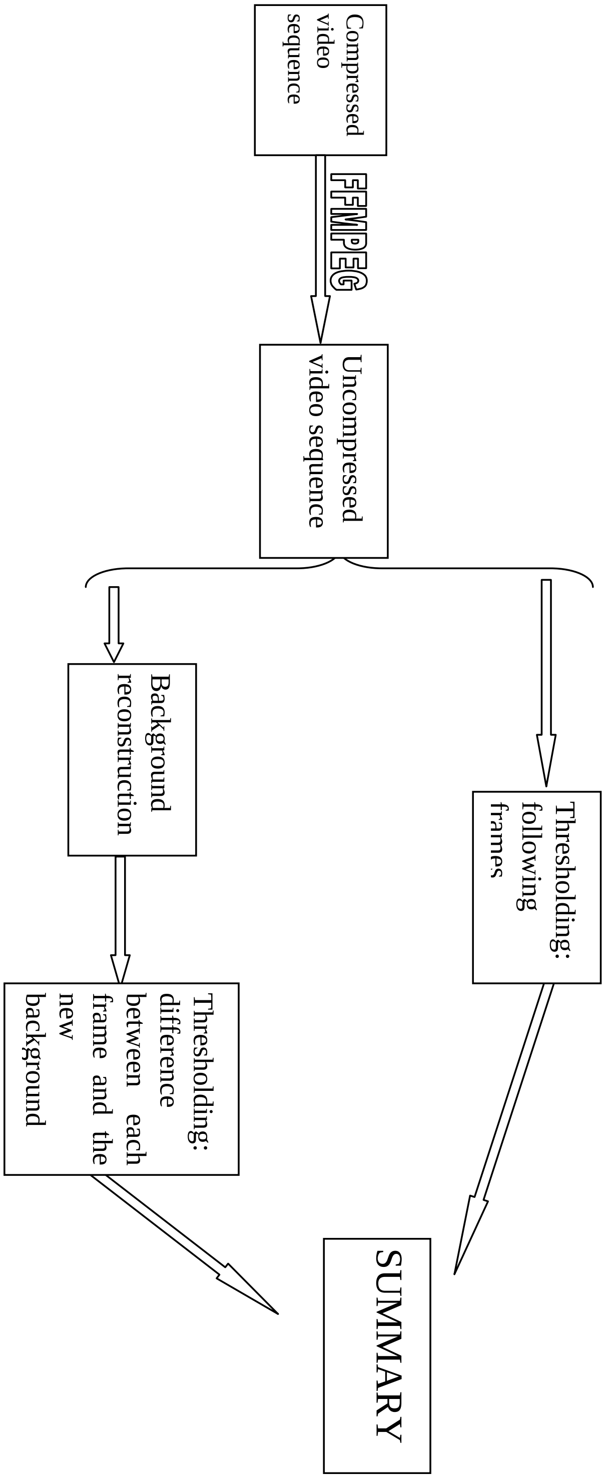

Compressed video sequence
FFMPEG
Uncompressed video sequence
Background reconstruction
Thresholding: following frames
Thresholding: difference between each frame and the new background
SUMMARY

Figure 6: Scheme of the whole process to generate a summary using the explained methods froma a .mpg video sequence

4 Testing and results

The objective of this project is to develop some methods that produce summaries from the original video sequences. In order to test the methods, some surveillance videos have been used. The efficiency of both methods will be checked.

<u>Thresholding: difference between following frames</u>

The only parameter that can vary in this program is the threshold. In order to test how it works and how many frames will compose the summary depending on the threshold value, we have given it three different values.

This method has been tested with nine different video sequences, obtaining the following results:

Number of frames in the summary					
		Total number of frames	**Threshold**		
			1000	**1500**	**1800**
Video sequences	IRIT01	45772	3432	1405	1060
	IRIT02	45745	2803	1300	1021
	IRIT03	40386	3472	1874	1226
	IRIT04	40306	2346	963	719
	IRIT05	35726	2894	1624	838
	IRIT06	35560	2088	872	662
	IRIT07	50550	4052	2255	1344
	IRIT08	50532	3240	1244	876
	IRIT09	51085	3382	1446	691

Table 1: Number of frames that compose the summaries applying the method called “Thresholding: difference between following frames”

Using the values founded previously we have evaluated the method several times. The results of these evaluations are given in the table 1, and can be interpreted graphically using figure 7 and figure 8.

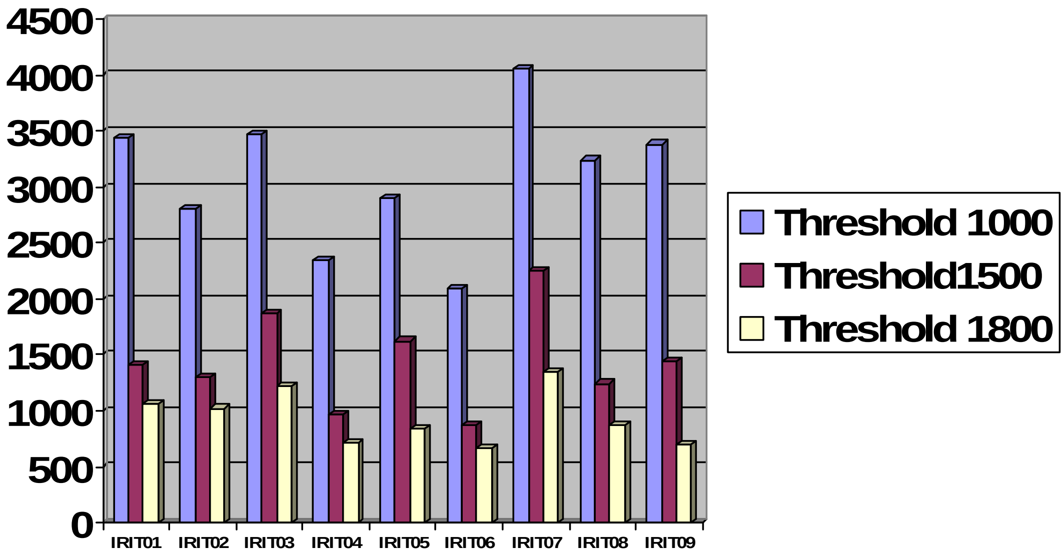


Figure 7: Comparison of the number of frames that compose the summary depending on the selected value for the threshold using the method called "Thresholding: difference between following frames"

However, even if this graphic compares the obtained results for the nine videos by applying "Thresholding: difference between following frames", we have not considered that the original surveillance video sequences do not have the same number of frames. So we cannot use this graphic to compare the results.

In order to compare the results obtained by applying this method to the nine videos sequences, we need to calculate numbers that can be compared. Consequently, all the numbers of frames that compose the summaries will be divided by the total number of frames of the original surveillance video sequence. In that way, we will obtain relative numbers that can be compared.

As a result, the next graphic compares all the relative numbers obtained to the different video sequences:

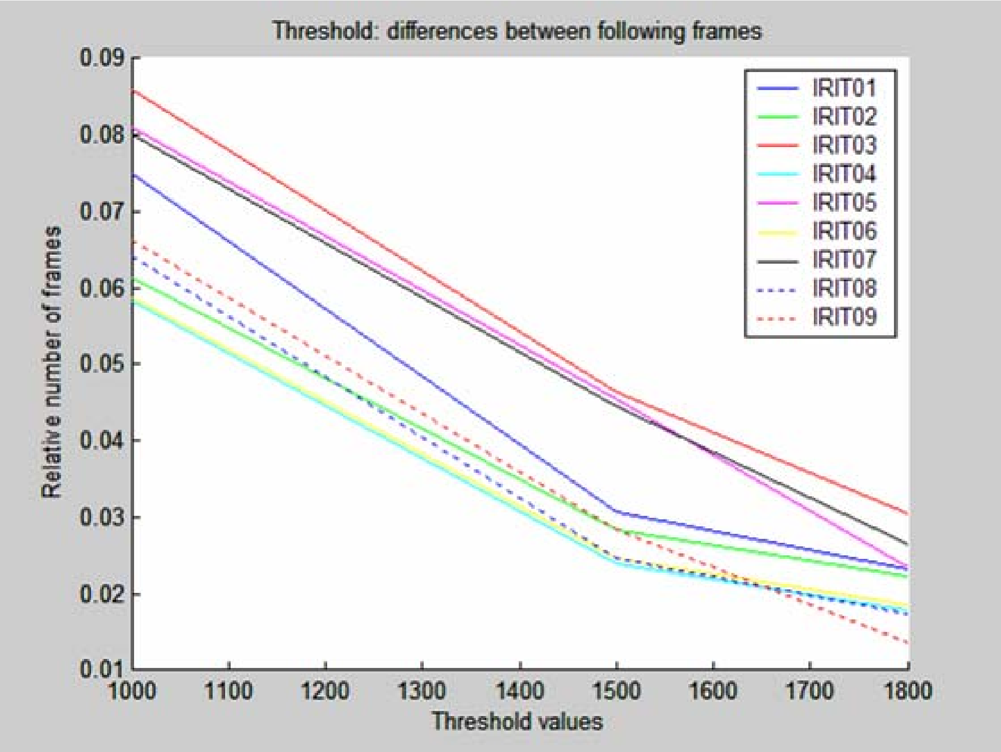


Figure 8: Evolution of the grade of the summarization (number of frames that compose the summary) depending on the threshold value when the used method is "Thresholding: difference between following frames"

As we can observe in the figure, there are two different situations. The first situation happens, for example to the video sequences called IRIT05 and IRIT07, the second threshold reduces the number of frames in the summary but the third one reduces this amount even more. The second situation consist on the fact that the second threshold obtains really good results and then the third one reduces the amount of frames in the summary but not so much, this happens to video sequences called IRIT01 and IRIT02.

Background reconstruction

In order to test this program, and evaluate how constant are surveillance video sequences. We have applied this method to some video sequences using different number of frames used to calculate them.

The chosen values of the number of frames are 100, 200, 300 y 1000. The objective was to discover if the background of the original video sequence varies frequently.

The obtained results by applying the program with the before mentioned values we have obtained exactly the same background. As shown in the next figure:

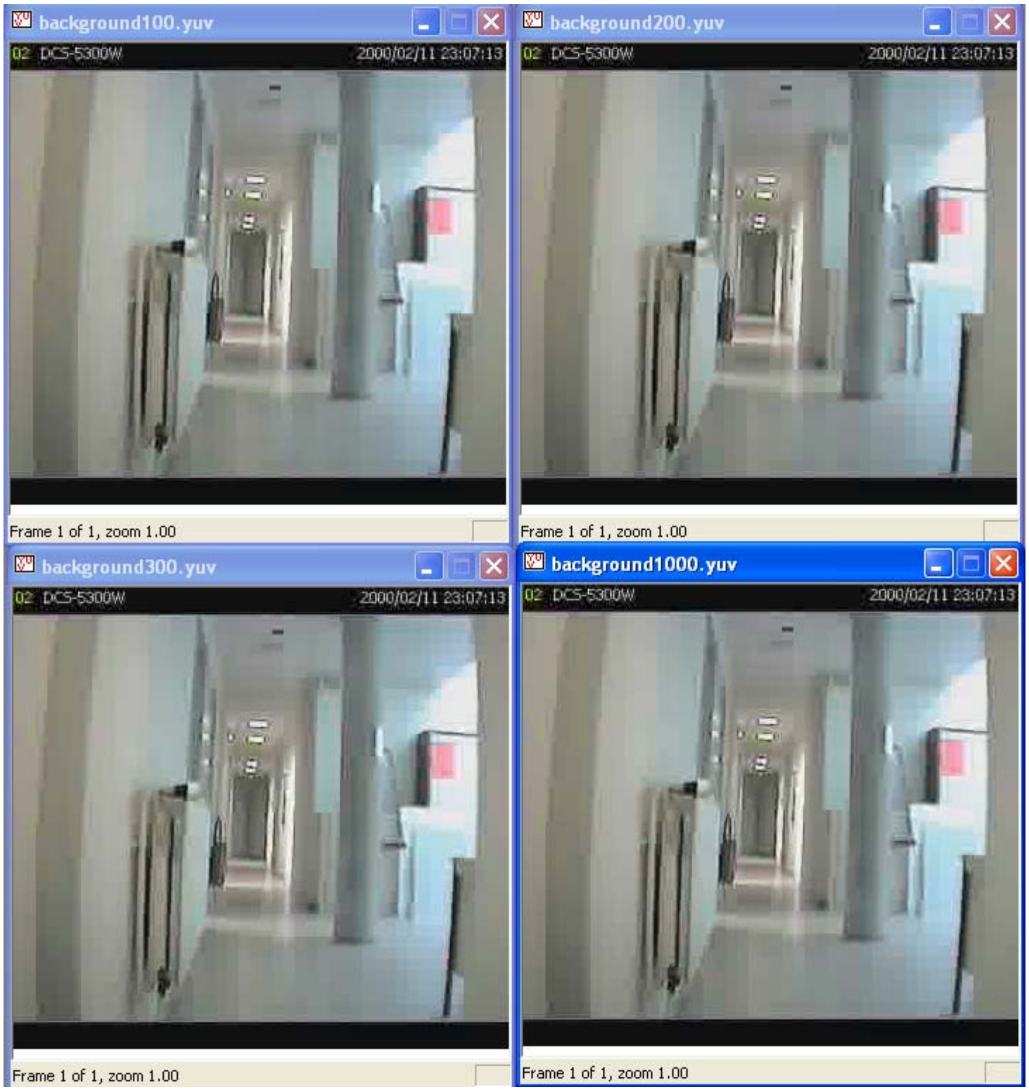


Figure 9: Comparison among the different backgrounds obtained applying the method called background reconstruction and using four different number of frames to calculate them. Using the method called "Thresholding: differences between each frame and a calculated background"

As the background is not going to change at least during one thousand frames, we will use this background in the next method.

Thresholding: difference between each frame and a calculated background

The only parameter that can vary in this program is the threshold. In order to test how it works and how many frames will compose the summary depending on the threshold value, we have given it three different values: 1000, 1500 and 1800.

This method has been tested with seven different video sequences, obtaining the following results:

Number of frames in the summary					
		Total number of frames	**Threshold**		
			1000	**1500**	**1800**
Video sequences	IRIT01	45772	260	52	44
	IRIT02	45745	2552	1109	486
	IRIT03	40386	1622	571	67
	IRIT04	40306	1228	533	333
	IRIT05	35726	366	16	0
	IRIT06	35560	2939	2035	1817
	IRIT07	50550	348	11	0
	IRIT08	50532	1930	597	369
	IRIT09	51085	136	15	0

Table 2: Number of frames that compose the summaries applying the method called "Thresholding: difference between each frame and a calculated background"

Using the values founded previously we have evaluated the method several times. The results of these evaluations are given in the table 2, and can be interpreted graphically using figure 10 and figure 11.

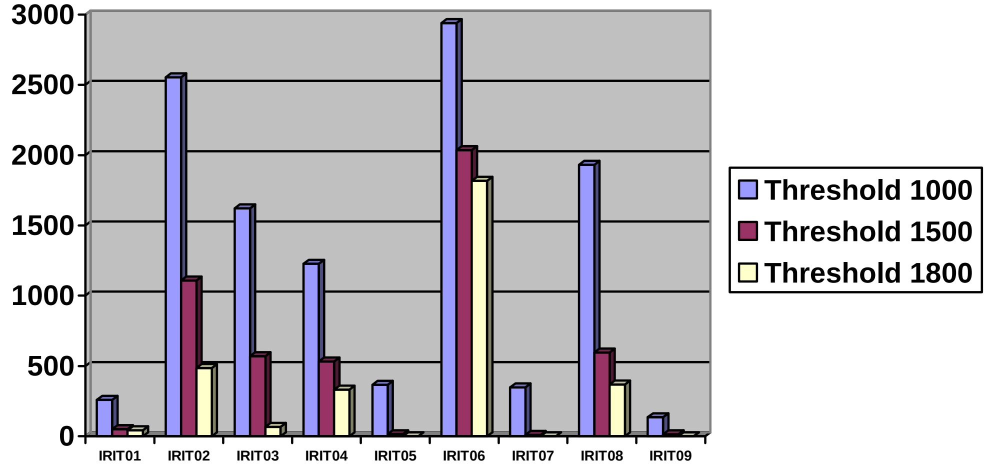


Figure 10: Evolution of the grade of the summarization (number of frames that compose the summary) depending on the threshold value, shen the used method is "Thresholding: difference between each frame and a calculated background"

In this method, we have the same problem that we had in the method "Tresholding: difference between following frames". We need to calculate the relative values to be able to compare how the methods behave in each video sequence. As shown in the next figure:

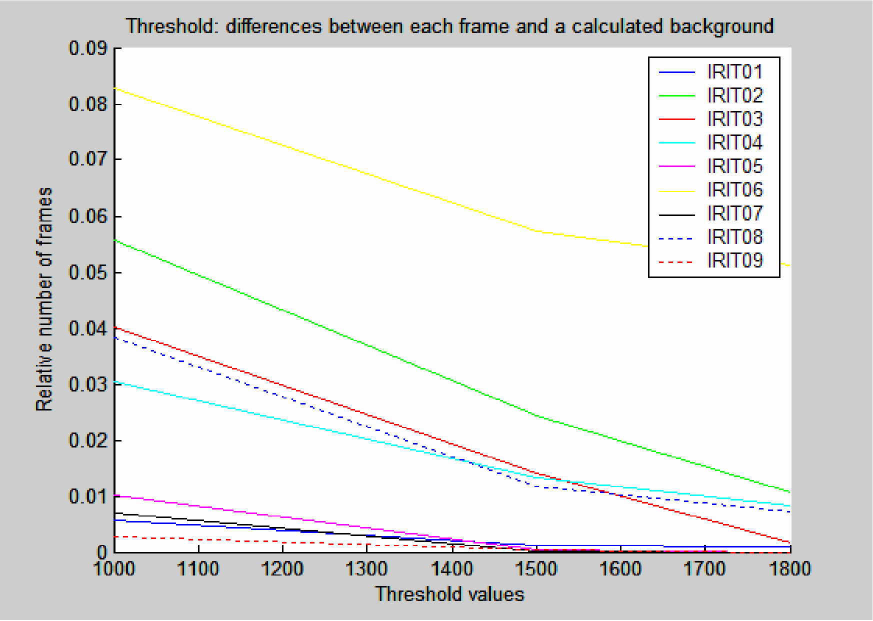


Figure 11: Evolution of the grade of the summarization (number of frames that compose the summary) depending on the threshold value when the used method is "Thresholding: difference between each frame and a calculated background"

As happened in the other “thresholding” method, there are the same two different situations. However, in this method the difference is not so obvious. Moreover, all the video sequences have a good behaviour with the second threshold.

Summarization limits:

There are two things that can put limits to the summaries, they are the bit rate and the threshold value. The number of frames that will compose the summary will depend on the bit rate we want to transmit and vice versa.

For that reason, it would be reasonable that we could fix or the bit rate, that would determine the number of frames that is allowed to be transmitted. Or to fix the threshold value that would determine the number of frames that would be kept.

In these programs, we only allow to fix the threshold value. We will consider as future work to improve the programs in order to allow fixing or the threshold value or the desirable bit rate.

Temporal efficiency:

Another issue that must be taken into account is the temporal efficiency of each method. In order to compare both, their performance must be considered as a relevant feature.

In the following tables we will observe the necessary time of each program to calculate a summary so we will be able to compare their temporal efficiencies.

Firstly, we have calculated the time we will need to simulate the method called “Thresholding between following frames” with different video sequences:

Necessary time to calculate the summary (seconds)				
		Threshold		
		1000	1500	1800
Video sequences	IRIT01	2082	2580	2219
	IRIT02	2404	2597	2337
	IRIT03	1821	2076	1209
	IRIT04	1569	2078	2048
	IRIT05	1579	1136	1107
	IRIT06	1309	1765	1553
	IRIT07	2937	2348	2629
	IRIT08	2899	2440	3302
	IRIT09	2828	3023	2856

Table 3: Time necessary to calculate the summaries applying the method called "Thresholding: difference between following frames"

Once, the first method has been applied, we calculate how long would take to reconstruct the background. It depends on the number of frames we will consider to do it. For that reason, we have measured the duration with two different amounts, firstly, one hundred frames, and secondly, one thousand frames.

Necessary time to calculate the background (seconds)			
		Number of frames	
		100	1000
Video sequences	IRIT01	144	1400
	IRIT02	179	1490
	IRIT03	167	1520
	IRIT04	165	1529
	IRIT05	119	1453
	IRIT06	153	1441
	IRIT07	105	1447
	IRIT08	176	1595
	IRIT09	169	1499

Table 4: Time necessary to calculate the average background applying the background reconstruction method previously described in the Design and Development Section

Obviously, it will be necessary longer time to calculate the background when more frames are taken into account. So, we must consider that (as we have seen before) the reconstructed background is exactly the same if we use one hundred frames as if we use one thousand frames. For that reason, we will use the one hundred frames background to calculate the summary by using the second method "Thresholding: difference between each frame and a calculated background".

The necessary time to calculate the background using the second method and a one hundred frames background is:

Necessary time to calculate the summary (seconds)				
		Threshold		
		1000	**1500**	**1800**
Video sequences	IRIT01	1859	1834	1939
	IRIT02	2073	2119	2042
	IRIT03	1442	1249	1439
	IRIT04	1303	923	1350
	IRIT05	802	819	1080
	IRIT06	812	1429	1647
	IRIT07	1655	1880	1760
	IRIT08	1869	1569	1610
	IRIT09	1506	1978	2018

Table 5: Time necessary to calculate the summary applying the method called "Thresholding: difference between each frame and a calculated background"

If we consider that the time to calculate the background must include the time spent in calculating the reconstructed background, then the time spent in calculating the summary using the method called "Thresholding: difference between each frame and a calculated background" will be:

Necessary time to calculate the summary (seconds)				
		Threshold		
		1000	1500	1800
Video sequences	IRIT01	2003	1978	2083
	IRIT02	2252	2298	2221
	IRIT03	1609	1416	1606
	IRIT04	1468	1088	1515
	IRIT05	921	938	1199
	IRIT06	965	1582	1800
	IRIT07	1760	1985	1865
	IRIT08	2045	1745	1786
	IRIT09	1675	2147	2187

Table 6: Time necessary to calculate the summary applying the method called "Thresholding: difference between each frame and a calculated background" including the time spent in calculating the general background used in it

The next step in this project will be to make a comparison between both methods to obtain some conclusions related with the generated results. In order to establish some patterns that would give us some clues to decide in which situation each method is better, we have to compare the efficiency of both methods in time and in grade of summarization (number of frames will compose the summary).

Video sequences	Total number of frames	Number of frames of the summaries								
		Threshold = 1000			Threshold = 1500			Threshold = 1800		
		First method	Second method	Difference	First method	Second method	Difference	First method	Second method	Difference
IRIT01	45772	3432	260	3172	1405	52	1353	1060	44	1016
IRIT02	45745	2803	2552	251	1300	1109	191	1021	486	740
IRIT03	40386	3472	1622	1850	1874	571	1303	1226	67	1159
IRIT04	40306	2346	1228	1118	963	533	430	719	333	386
IRIT05	35726	2894	366	2528	1624	16	1608	838	0	838
IRIT06	35560	2088	2939	- 851	872	2035	- 1163	662	1817	- 1155
IRIT07	50550	4052	348	3704	2255	11	2244	1344	0	1344
IRIT08	50532	3240	1930	1310	1244	597	647	876	369	507
IRIT09	51085	3382	136	3246	1446	15	1431	691	0	691

Table 7: Comparison between the number of frames that compose the summaries depending on which method is applied to calculate it (“Thresholding: difference between following frames” or “Thresholding: difference between each frame and a calculated background”)

Video sequence s	Threshold = 1000			Threshold = 1500			Threshold = 1800		
	First method	Second method	Difference	First method	Second method	Difference	First method	Second method	Difference
IRIT01	2082	2003	79	2580	1978	602	2219	2083	136
IRIT02	2404	2252	152	2597	2298	299	2337	2221	116
IRIT03	1821	1609	212	2076	1416	660	1209	1606	- 397
IRIT04	1569	1468	101	2078	1088	990	2048	1515	533
IRIT05	1579	921	658	1136	938	198	1107	1198	- 91
IRIT06	1309	965	344	1765	1582	183	1553	1800	- 247
IRIT07	2937	1760	1177	2348	1985	363	2629	1865	764
IRIT08	2899	2045	854	2440	1745	695	3302	1786	1516
IRIT09	2828	1675	1153	3023	2147	876	2856	2187	669

Table 7: Comparison between the time spent in calculating each summary depending on which method is applied (“Thresholding: difference between following frames” or “Thresholding: difference between each frame and a calculated background”)

If we analyse both tables, we realize that usually the second method (“Thresholding: difference between each frame and a calculated background”) obtains better results than the first method (“Thresholding: difference between following frames”). Usually, it temporal efficiency is better and it level of summarization, too.

The only video sequence where the temporal efficiency as well as the level of summarization is better in the first method is with the video sequence called “IRIT06”. In order to conclude why this happens we must analyse the advantages of each method:

- The first method analyses how the frames evolve in time. As a result it only compares following frames. Comparing frames that are neighbours, we can detect when the image experiment an important change (bigger or smaller depending on the threshold). It has to be considered that this method compares following frames so, the previous image can be classified as a changeable background.

  For that reason, if the video sequence is static or the changes in it are really small and slow, this method will not consider them. And the frames which represent these changes will not be kept in the summary.

  However, if the video sequence keeps fast movements, a lot of frames will be kept. Because the energy of the difference between following frames will be really big due to the big change in the frames caused by the high speed in the events.

  Another issue that must be considered is that the number of frames that will be kept in the summary can also affect to the time efficiency. The time spent in saving the relevant frames must be considered and if the number of saved frames increased the necessary time to keep them will increase, too.

- The second method analyses the evolution in time of the video sequence. However and compared with the first method, the second method is more focused on the evolution of the foreground of the video sequence.

  For that reason, if the video sequence is static or the changes in it are really slow, they will cause the same effects as if the changes are faster. Given that it compares every single frame with a calculated background (an average background), so what really matters is if the change in the foreground is big or not.

  For example, if some object appears quickly in the image, what this method will analyse is the background (that in this case is the fast object). As a result, if the object is big enough (depending on the fixed threshold) then the program will consider this frame as relevant, and it will be saved in the summary.

  Generally, this method keeps fewer frames than the first method, except when the sequence is slow, and then the first method is more efficient. The reason is that the first method does not keep following frames that are really similar (it takes advantage of the temporal redundancy), while the second method keep every frames which foreground causes a big change in the differential energy value.

In order to compare visually the results obtained by both methods, we can use the following links (all of them have been generated using the surveillance video sequence called IRIT02 and the value of the threshold is 1800):

Thresholding: differences between following frames.

frames_IRIT02_1800.yuv

Thresholding: differences between each frame and a calculated background.

background_IRIT02_1800.yuv

In order to be able to see these summaries we have to introduce some parameters:
Width=352
Length=328
Format= YUV 4:2:0
Header skip=0

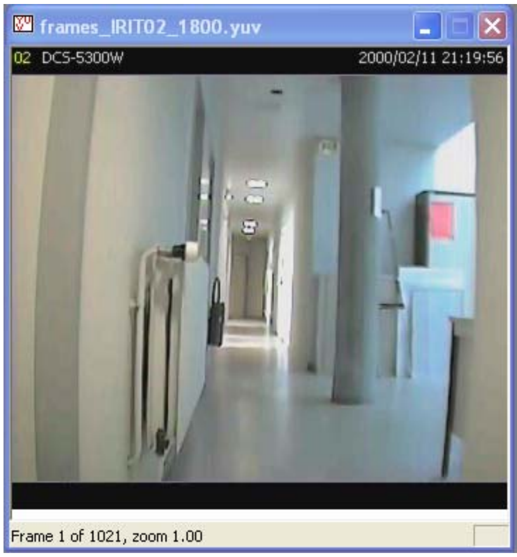
Analysis of results:

In order to compare both summarization methods, we will show some of frames that compose the resulting summaries of the same surveillance video sequence by applying both proposed methods. On the left column we will see the frames that compose the summary generated by the method called “Thresholding: difference between following frames. While on the right column we will find the frames that compose the summary generated by the method called “Thresholding: difference between each frame and a calculated background”.

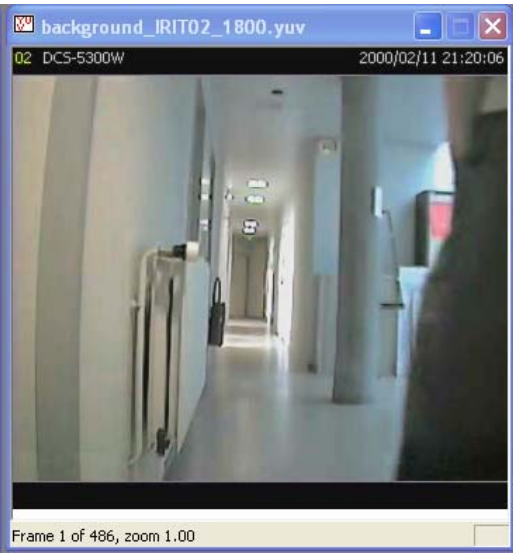
Firstly, the ten first frames will be shown to compare how the two methods start the summary. Later only relevant frames will be shown in order to establish how both methods work and why these frames have been chosen.

Following the first ten frames are shown:

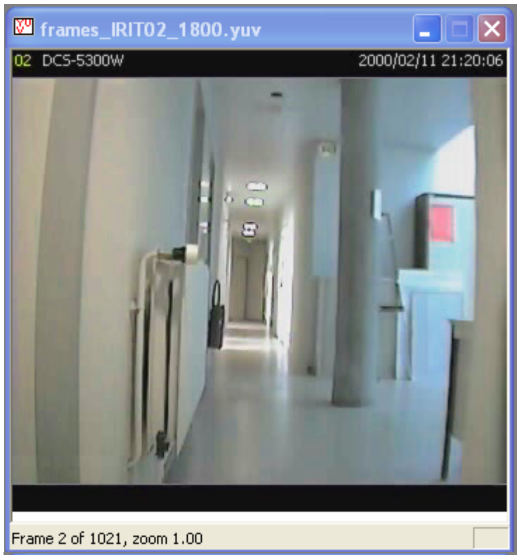
Deleted: ¶

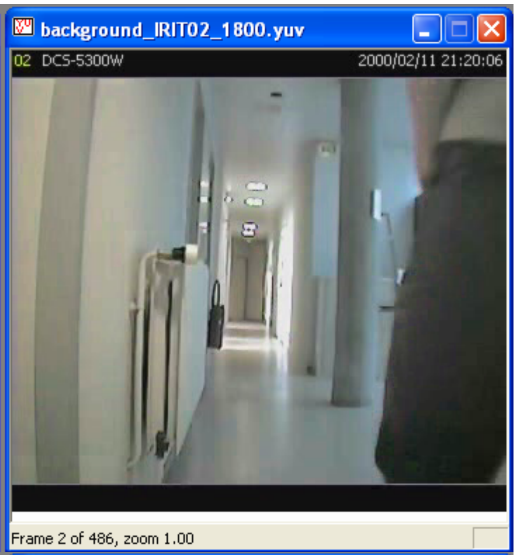


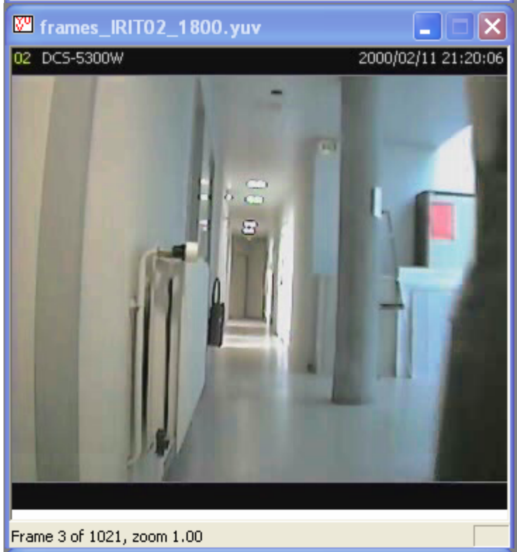
Frame 1 of 1021, zoom 1.00

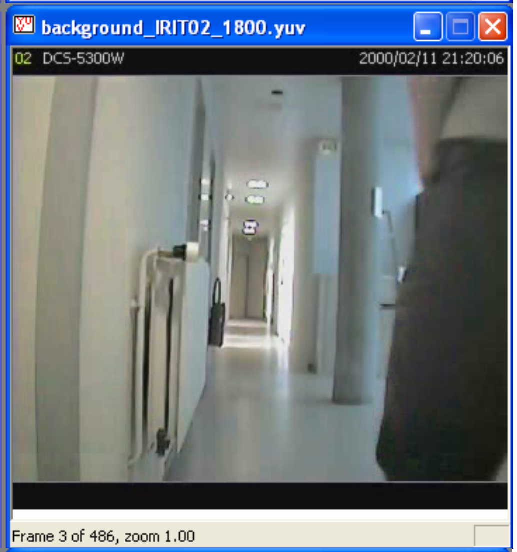


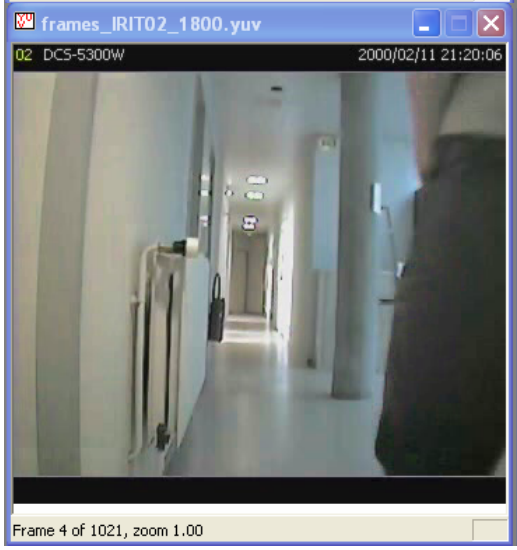
Frame 1 of 486, zoom 1.00

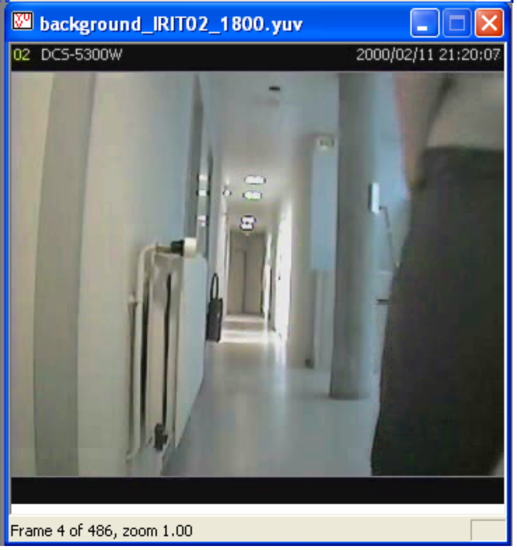

frames_IRIT02_1800.yuv
02 DCS-5300W
2000/02/11 21:20:06
Frame 2 of 1021, zoom 1.00

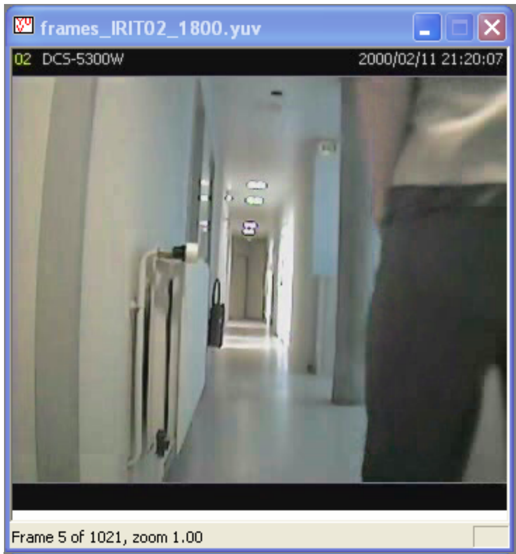

background_IRIT02_1800.yuv
02 DCS-5300W
2000/02/11 21:20:06
Frame 2 of 486, zoom 1.00

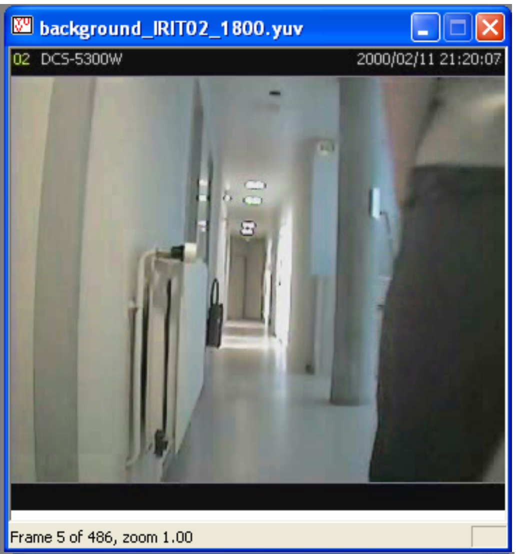

frames_IRIT02_1800.yuv
02 DCS-5300W
2000/02/11 21:20:06
Frame 3 of 1021, zoom 1.00

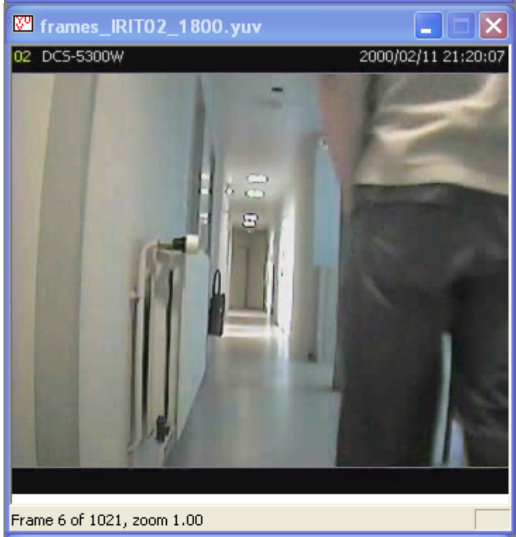

background_IRIT02_1800.yuv
02 DCS-5300W
2000/02/11 21:20:06
Frame 3 of 486, zoom 1.00

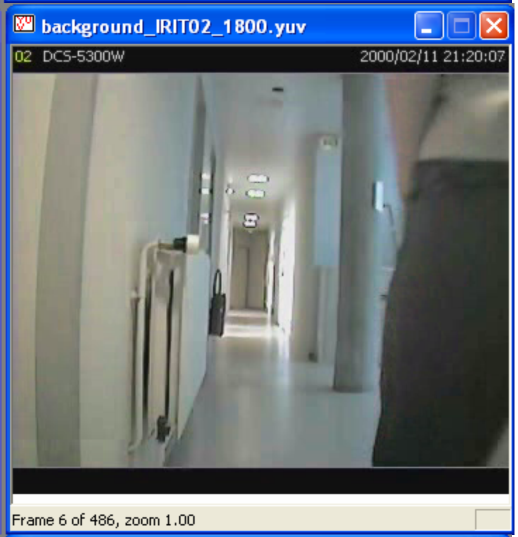

frames_IRIT02_1800.yuv
02 DCS-5300W
2000/02/11 21:20:06
Frame 4 of 1021, zoom 1.00

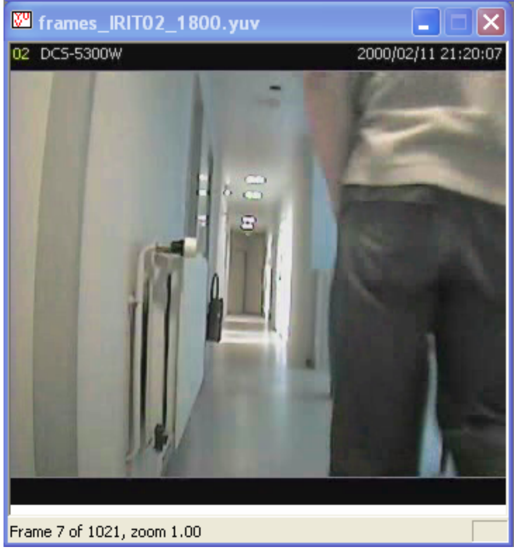

background_IRIT02_1800.yuv
02 DCS-5300W
2000/02/11 21:20:07
Frame 4 of 486, zoom 1.00

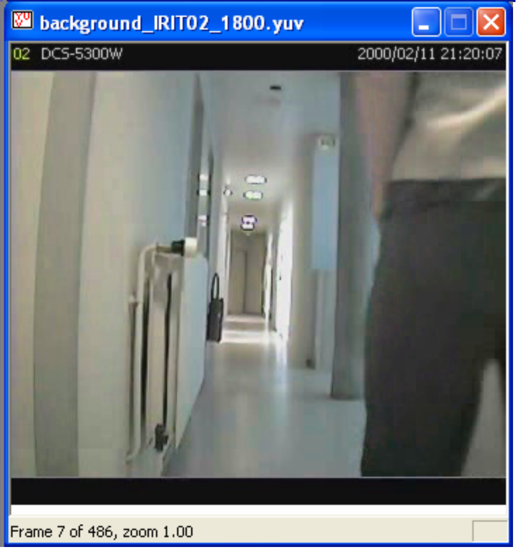

frames_IRIT02_1800.yuv
02 DCS-5300W
2000/02/11 21:20:07
Frame 5 of 1021, zoom 1.00

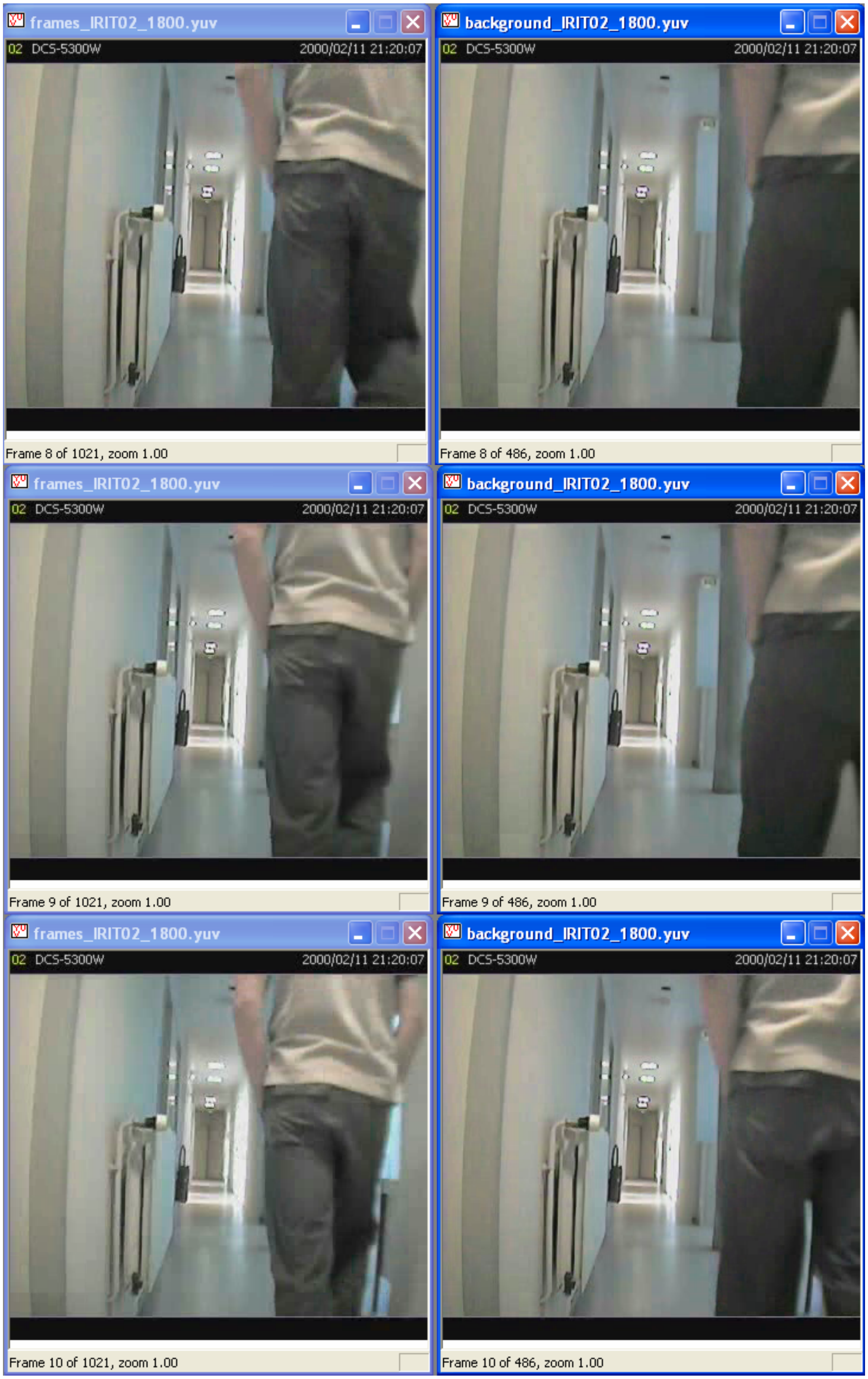

background_IRIT02_1800.yuv
02 DCS-5300W
2000/02/11 21:20:07
Frame 5 of 486, zoom 1.00


frames_IRIT02_1800.yuv
02 DCS-5300W
2000/02/11 21:20:07
Frame 6 of 1021, zoom 1.00


background_IRIT02_1800.yuv
02 DCS-5300W
2000/02/11 21:20:07
Frame 6 of 486, zoom 1.00


frames_IRIT02_1800.yuv
02 DCS-5300W
2000/02/11 21:20:07
Frame 7 of 1021, zoom 1.00


background_IRIT02_1800.yuv
02 DCS-5300W
2000/02/11 21:20:07
Frame 7 of 486, zoom 1.00



Frame 8 of 1021, zoom 1.00

Frame 8 of 486, zoom 1.00

Frame 9 of 1021, zoom 1.00

Frame 9 of 486, zoom 1.00

Frame 10 of 1021, zoom 1.00

Frame 10 of 486, zoom 1.00

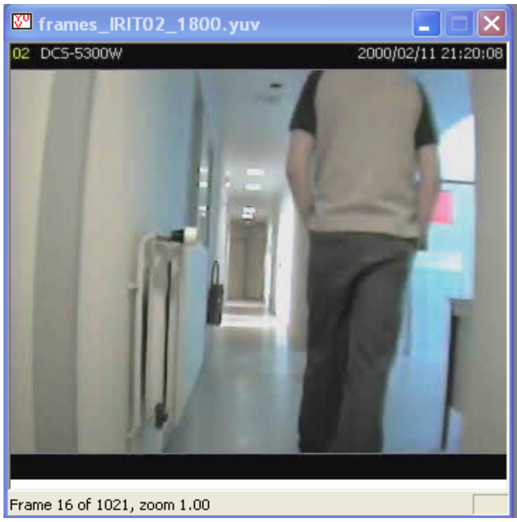
Figure 12: Ten first frames of the summaries obtained by applying both summarization methods

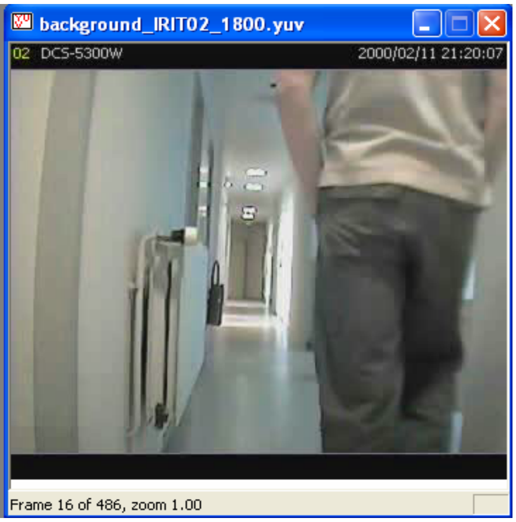
As it can be observed in the frames, the first method ("Thresholding: difference between following frames") is composed by 1021 frames while the second method ("Thresholding: difference between each frame and a calculated background") has generated a 486 frames summary.

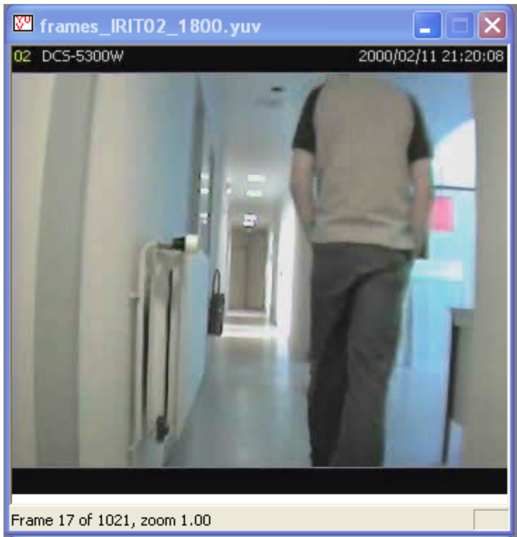
The first method analyses how the frames evolve in time. For that reason if a video sequence is quite static, it will keep only some frames, the key frames. The second method, however, is focused on the background and it considers only the images which foreground is big (because they will cause a big change in the frame as well as in the energy of the difference). So, it does not matter if the image is slow or static, because the second method only considers the change in the frame caused by the foreground.

For example, in the first ten frames, the first method takes them more separated in time than the second one due to the size of the foreground (it is big, the man is close to the camera). The first method only takes some frames, the key ones, not a lot because the video sequence is quite slow. However, the second method saves more frames, because the foreground is big and causes a big change in the frame and also in the energy of the difference.

In the next video sequence, we can observe that as soon as the foreground becomes smaller, the second method does not save the frames while the first one does it, because the change of the frame compared with the following one is relevant (the result will be a longer summary by applying the first method than applying the second one).

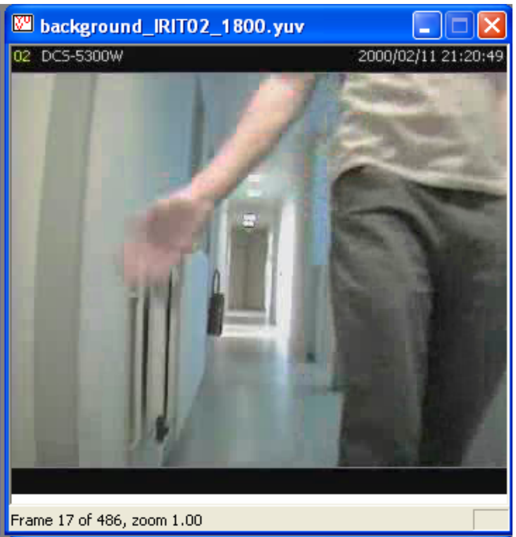





frames_IRIT02_1800.yuv

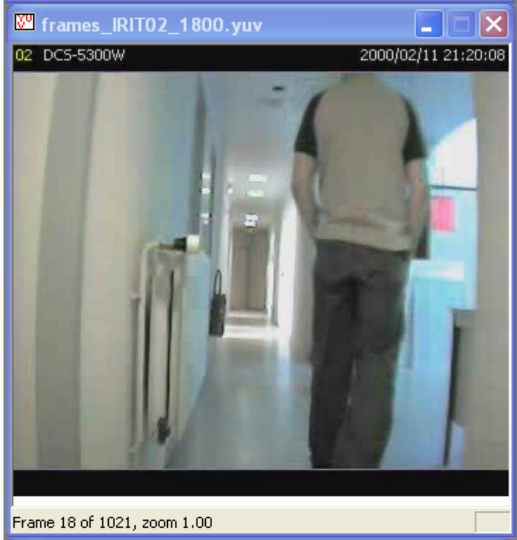
02 DCS-5300W 2000/02/11 21:20:08

Frame 17 of 1021, zoom 1.00


background_IRIT02_1800.yuv

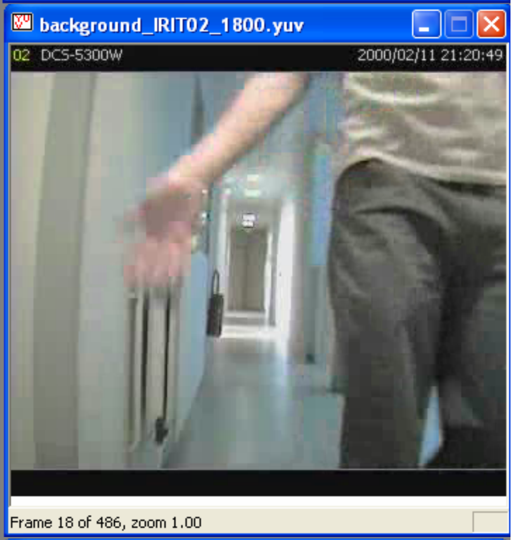
02 DCS-5300W 2000/02/11 21:20:49

Frame 17 of 486, zoom 1.00


frames_IRIT02_1800.yuv

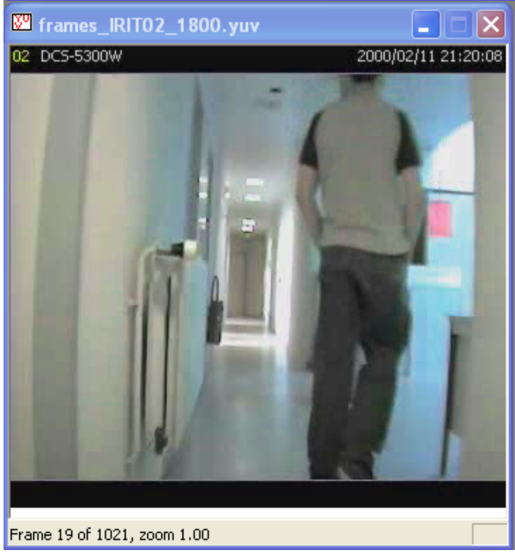
02 DCS-5300W 2000/02/11 21:20:08

Frame 18 of 1021, zoom 1.00


background_IRIT02_1800.yuv

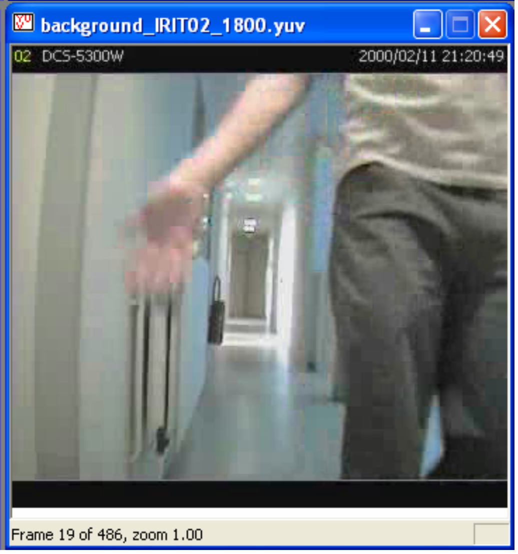
02 DCS-5300W 2000/02/11 21:20:49

Frame 18 of 486, zoom 1.00


frames_IRIT02_1800.yuv

02 DCS-5300W 2000/02/11 21:20:08

Frame 19 of 1021, zoom 1.00


background_IRIT02_1800.yuv

02 DCS-5300W 2000/02/11 21:20:49

Frame 19 of 486, zoom 1.00

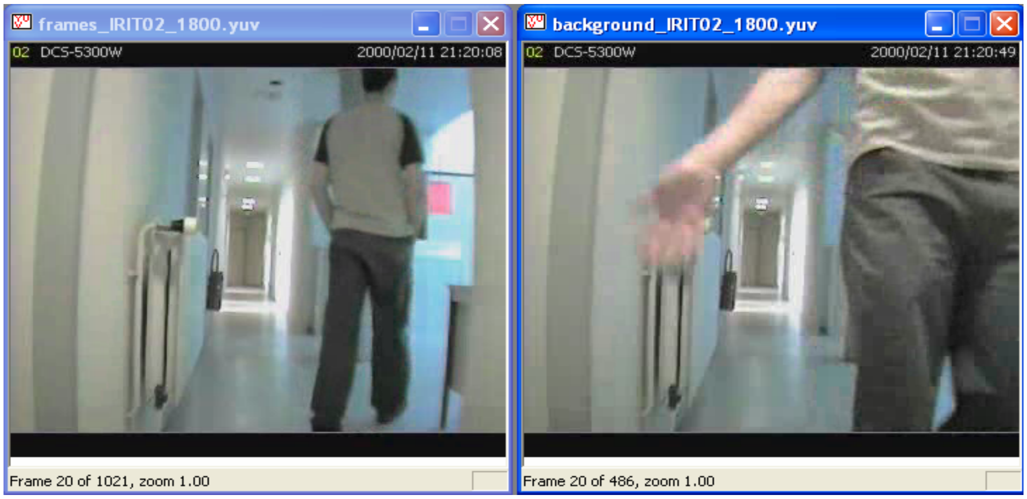


Figure 13: Sequence from the summaries I.

In the next sequence (composed by two frames) we can observe how the second method does not save any frame from 21:20:49 to 21:25:14.

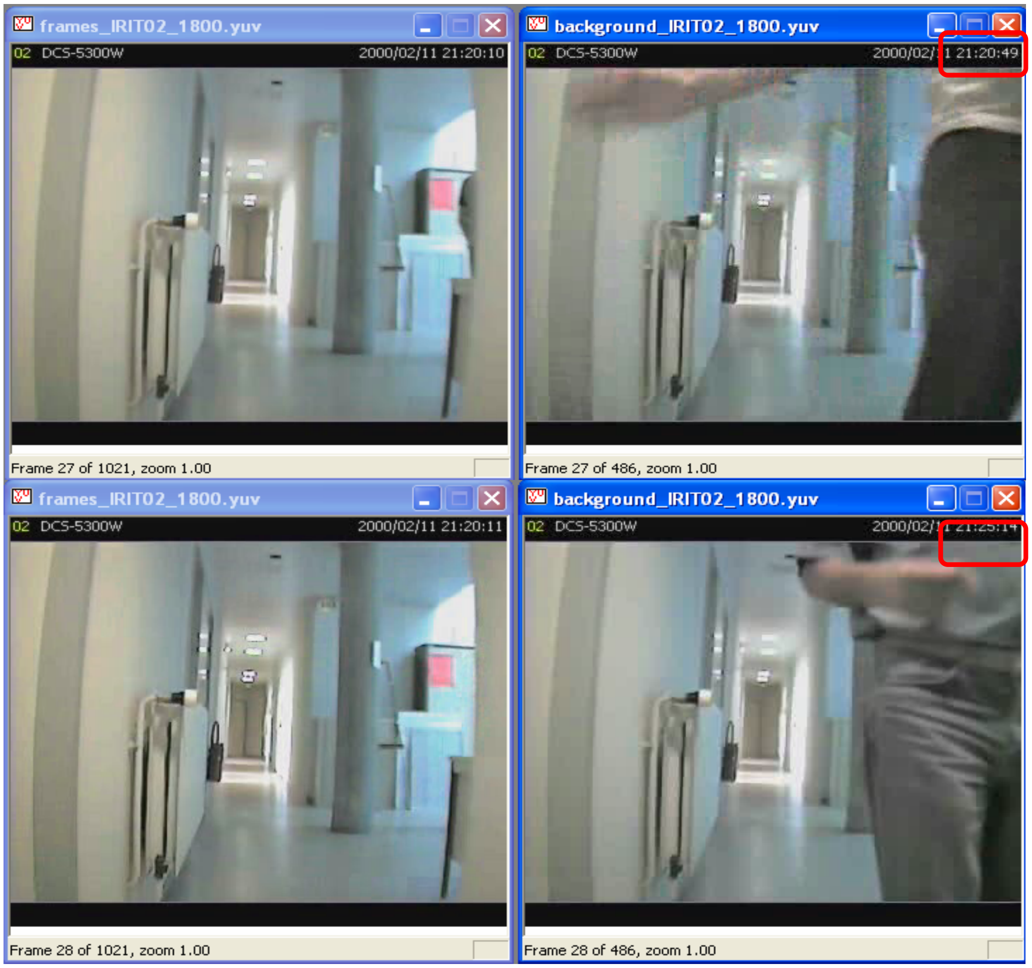


Figure 14: Sequence from the summaries II

However, as we can observe in the next image, the first method saves an event during this period.

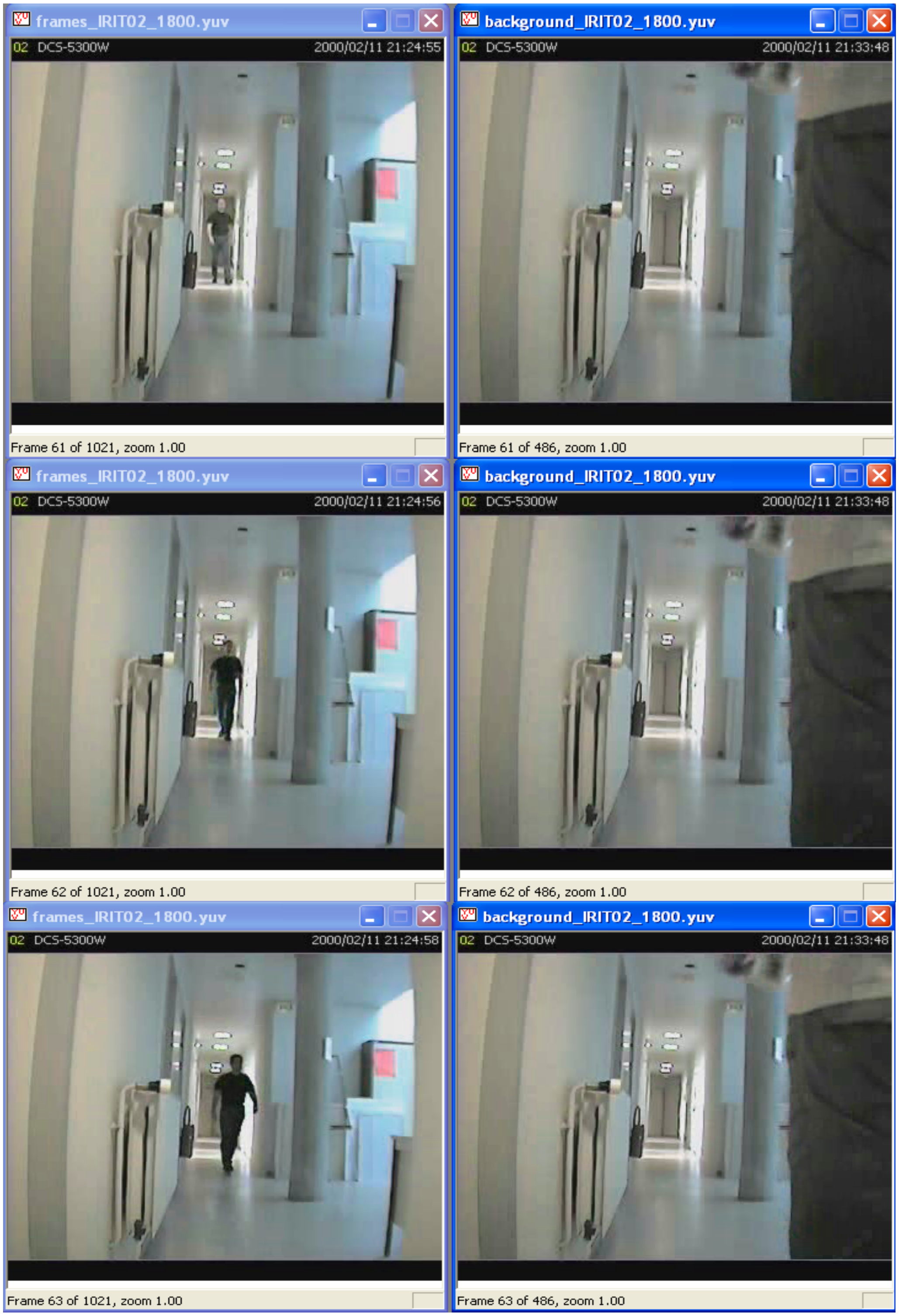


Figure 15: Sequence from the summaries III

Basically, these frames have not been considered by the second method because the foreground does not cause a big change in the energy of the difference due to the size of the foreground. As a result, the summary produced by the second method does not retrieve this event. While the first method retrieves it because this event causes a relevant change between following frames.

Moreover, in this video sequence the second method only saves the events that take place close to the camera. This happens because the threshold value is quite high (1800) so no small foregrounds will be taken into account. As a result, the second-method summary will lose some important information, while the first-method will be longer but it will take care more about details.

To conclude, I would like to say that all the tests have been performed in a PC with the next characteristics:

Microsoft XP
Pentium 4
CPU 3,40GHz
1GB RAM

5 Conclusions and future work

5.1 Conclusions

Along this project two summarization methods have been developed. The first one considers more the temporal evolution between frames, while the second one is more focused in how the foreground changes. They are two different approaches to summarize surveillance video sequences and both of them make good use of different properties of the surveillance video sequences. The former considers that surveillance video sequences usually do not experience fast changes. The latter considers that surveillance video sequences generally do not change their background.

The objective of this final project was to develop some methods to summarize surveillance sequences. In order to do it, we have considered the special features of the surveillance video sequences. Now, that both methods have been implemented and tested. We must consider the necessities of surveillance systems. The main objective of surveillance systems is to observe all the relevant events that have taken place. Moreover, the objective of a summarization system is to reduce the amount of information that must be checked. If we consider both objectives, we will observe that the necessity of a summarization system of surveillance video sequences is to reduce the amount of data that must be checked, without loosing any information relevant events.

Both systems fulfill this objective. Nevertheless the efficiency must be considered, too. The first method ("Thresholding: difference between following frames") is slower and the generated summaries are bigger. However, we can be sure that there is not going to loose any information about events that have taken place. While the second method is faster and its summaries are shorter, but it can consider that some events are not important because they are small (in size, compared with the whole image size).

To sum up, if the surveillance system is more worried about speed, the second method is more advisable. While if the surveillance system is looking for details then the first one is more recommended.

Finally, I would like to remark the practical utility of the first implemented method. As a practical example, this method, thresholding differences between following frames, has been used to obtain the results exposed in a recent paper called “Event Detection and Clustering for Surveillance Video Summarization”. This paper has been published in the conference Wiamis 2008 (Workshop on Image Analysis for Multimedia Interactive Services).

5.2 *Future work*

Due to the fact that this topic is prominent nowadays, there are many possible improvements and many suggestions, too. As future work we can consider the analysis and comparisons with different methods to summarize surveillance videos, as well as some suggestions and improvements of the developed methods.

According to my opinion the most important lines for future work would be:

- In the second method, “Thresholding: difference between each frame and a calculated background”, mainly, the difference between each frame and a previously calculated background is worked out in order to determine if the frame has change enough to take part into the summary.

  In order to calculate the new background, the number of frames used must be determined. In the before explained experiments, the number of frames varied between: 100, 200, 300 and 1000.

  However, these numbers do not represent the whole video sequence. For that reason, as future work we could propose to program a method that calculates a new background periodically. In a video sequence the background can change along the time although in surveillance videos it does not change quickly.

  In order to reduce the temporal redundancy, we could calculate these new periodic backgrounds that would adjust easily to the video sequence. For that reason, the calculated background must change periodically. As it can be observed in the section called “Testing and Results”, in surveillance video sequences the background varies slowly. Moreover, in our video sequences we could observe that the background does not suffer almost any change in one thousand frames. Even though we must take care of the duration of the intervals, because if it is small, a long time will be consumed to calculate the new background.

Applying this method, the temporal redundancy is taken into account and it would be easier to reduce the number of useless frames in order to build the summary.

- Another improvement could be a program composed by both previous methods. Instead of comparing all the frames with the new calculated background, it can be subtracted to all the frames. That way only the foregrounds are going to be compared now. So, as we have all the “useful information” (the foregrounds) separated from the useless one, now, we can compare what is really changing in the image. In order to do that, we will compare following foregrounds. Afterwards, we will apply a threshold to decide if the change is important or if the change is only due to the bad quality of the video sequence.

- Another enhancement would be to be able to fix or the threshold value or the bit rate we want to transmit. Instead of only fixing the threshold value.

- In the video sequence, it would be a necessary to analyse how these algorithms work with special events that could change the luminance or chrominance values of the frames, like for example, some changes in lighting. It will be necessary a benchmark with that includes special situations.

References

[1] “Intelligent distributed surveillance systems: a review”, Valera, M. and Velastin, S.A. IEE Proceedings - Vision, Image and Signal Processing, 152(2), pp. 192-204. ISSN (online) 1350-245X. 2006.

[2] “Architecture for video summarization services over home networks and the internet”, Sam Shipman, Regunathan Radhakrishan and Ajay Divakaran. Mitsubishi Electric Research Labs, Cambridge, MA, USA. 0-7803-9459-3/06/$20.00 ©2006 IEEE.

[3] “Rate-Distortion Optimal Video Summary Generation”. Zhu Li, Member, IEEE, Guido M. Schuster, Member, IEEE, Aggelos K. Katsaggelos, Fellow, IEEE, and Bhavan Gandhi, Member, IEEE. IEEE TRANSACTIONS ON IMAGE PROCESSING, VOL. 14, NO. 10, OCTOBER 2005.

[4] “Rate-Distortion Optimal Video Summary Generation”. Zhu Li, Member, IEEE, Guido M. Schuster, Member, IEEE, Aggelos K. Katsaggelos, Fellow, IEEE, and Bhavan Gandhi, Member, IEEE. IEEE TRANSACTIONS ON IMAGE PROCESSING, VOL. 14, NO. 10, OCTOBER 2005.]

[5] “Video motion processing for event detection and other applications”, R J Evans, E L Brassington, C Stennett, Roke Manor Research Ltd, United Kingdom.

[6] A. Yoshitaka, et al., “Knowledge-Assisted Content-Based Retrieval for Multimedia Databases”, IEEE Multimedia, pp. 12-21, Winter 1994.

[7] P.N. Sridharan and S. Raman, “Chracteristics of video data for signal analysis”, Proceedings of ICSP ’96, pp. 1954-1957.

[8] M. Flavin, “Fundamental concepts of information modeling”, Yourdon Press, 1981.

[9] “A forensic image processing environment for investigation of surveillance video”, M. Jerian, S. Paolino, F. Cervelli, S. Carrato, A. Mattei, L. Garofano, DEEI University of Trieste, ScienceDirect.

[10] “Performance Evaluation of a Real Time Video Surveillance System”, S. Muller-Schneiders, T. Jager, H.S. Loos, W. Niem, R. Bosch, Corporate Research Advance Engineering Multimedia, Telematic and Surround Sensing Systems, Hildesheim – Germany.

[11] “Classificatin of Smart Video Surveillance Systems for Commercial Applications”, M.H. Sedky, M. Moniri, C.C. Chibelushi.

[12] "Classificatin of Smart Video Surveillance Systems for Commercial Applications", M.H. Sedky, M. Moniri, C.C. Chibelushi.

[13] "Toward Efficient Collaborative Classification For Distributed Video Surveillance", C.P. Diehl, PhD Thesis, Carnegie Mellon University.

[14] "Video Summarization by Curve Simplification", D.DeMenthon, V.Kobla and D.Doermann,ACM MM98 (1998).

[15] "Video Summarization and Scene Detection by Graph Modelling", Chong-Wah Ngo, Yu-Fei Ma and Hong-Jiang Zhang.

[16] "Video Skimming for Quick Browsing based on Audio and Image Characterization", Michael A. Smith and Takeo Kanade.

[17] "Video summarization: Methods and landscape", Mauro Barbieri, Latitha Agnihotri and Nevenka Dimitrova, Philip Research.

[18] "The Ultimate Futility of Background Subtraction", John Krumm, Kentaro Toyama, Barry Brumitt, Brian Meyers.

[19] "Event detection and clustering for surveillance video summarization", Urosh Damnjanovic, Virginia Fernández, Ebroul Izquierdo, José María Martínez, Wiammis 2008.

[20] "A background reconstruction algorithm based on pixel intensity classification in remote video surveillance system".

[21] "The MPEG Handbook, MPEG-1 MPEG-2 MPEG-3", John Watkinson.

[22] "MPEG Digital Video-Coding Standards", Dr Thomas Sikora, IEEE Signal Processing Magazine, September 1997.

[23] "The MPEG Handbook, MPEG-1 MPEG-2 MPEG-3", John Watkinson.

[24] "CIE Publication N. 17.4 – 1987, *International Lighting Vocabulary*, Definition 845-03-25".

[25] "A Review of RGB colour spaces", Danny Pascale.

Appendixes

A Programmer's tutorial

The purpose of this section is to provide a guide to the future user that will make him easier to use the developed methods. For that reason, all the input and output parameters of all the programs will be detailed, as well as, the program used to visualize the resulting summaries.

- Thresholding: difference between following frames.

  Input arguments:

  - Name of the video sequence with the extension (it only accepts YUV files).
  - Video sequence (it must be inside the folder that contains the script or redirect it).
  - Value of the threshold.

  Output arguments:

  - Summary (It will be a YUV file).

- Construction of the background: Background reconstruction.

  Input arguments:

  - Name of the video sequence with the extension (it only accepts YUV files).
  - Video sequence (it must be inside the folder that contains the script or redirect it).
  - Number of frames that will be used in order to calculate the background.

  Output arguments:

  - New background frame (It will be a YUV file and its name will be "Background.yuv").

- Thresholding: difference between each frame and a calculated background.

  Input arguments:
  - Name of the video sequence with the extension (it only accepts YUV files).
  - Video sequence (it must be inside the folder that contains the script or redirect it).
  - Value of the threshold.
  - Background frame (It must be inside the folder that contains the script or it must be redirect it). This frame was calculated in the program called "Background reconstruction". It must receive the name "Background. yuv".

  Output arguments:
  - Summary (It will be a YUV file).
  - A text file that includes all the energy values of the difference between each frame and the calculated background.
  - A text file that includes the number of all the saved frames included in the summary. It will help to know the position of these frames in the whole video sequence.

During this project I have used SeqView program in order to visualize the YUV files. It will not allow visualizing long videos. However, it works properly with videos around 512MB. Another important issue related with SeqView is that in order to visualize the video sequence, first, the following values must be inserted:

- Size of the frames that compose the video sequence, width and height.
- Compression format of the video sequence (The compression format that has been used all along this project is YUV 4:2:0).
- Header skip will take null value.

B Compression

All the existing algorithms of video compression can be classified into two groups by the Shanon’s theory. The first group would be compression without loss (lossless) and the second group would be compression with loss (lossy).

- Compression without loss (Lossless):

The main objective of these algorithms is to reconstruct the information without losing any information. The consequence of this demand is that the compression rate is lower than other of other compression algorithms. Even though the compression rate would be really close to H, that is the compression rate defined by Shanon. Where H is defined by the following equation:

$$H = \log_2 m \text{ (bits / character)}$$

Where m is the total number of possible characters and/or symbols.

- Compression with loss (Lossy):

The main characteristic of these algorithms is that they reduce the amount of bits by deleting those bits that are unnecessary. This kind of algorithms is used in images where all the colours and tones are not always necessary, and they require a huge amount of storage capacity.

Distortion is the main point of analysis in these algorithms. In fact, it is a mathematical tool which shows if the compressed information is too far from the original one or not.

One point to take into account is that the recover information is not exactly the same of the original one. That means that in the compression some losses have taken place. As a result, if one of these algorithms is applied to an image, the outcome image can shows difference in some pixels or even in the whole image. That depends on the quality of the compression algorithm and in the compression rate.

Some good examples of these algorithms are:

- MP3 in the audio area.
- MPEG- 4 in the video area.

Redundancy

Redundancy is the main principle in which codification is based. A video sequence has three different kinds of redundancy in order to obtain a good compression rate. They are:

- Spatial redundancy.
- Temporal redundancy.
- Visual redundancy (Redundancy that human beings introduce psychologically).

Firstly, temporal and spatial redundancy takes place because the value of pixels is not absolutely independent one from each other. Moreover, the value of pixels is usually correlated with the value of some close pixels in time and close pixels in space. As a result, these values can be predicted some way.

Secondly, psycho visual redundancy is more related with the human eye limitations. That means that usually images have more information, that what is necessary, in order human eye can appreciate details. For example, human eye cannot appreciate easily special details or fast transitions or movements.

Redundancy is the reason why bit-rate can be reduced and why an image can be compressed. If redundancy is eliminated, bit-rate will be reduced without getting worse the quality of the image that human eye received.

There are four different types of images depending on which kind of redundancy is taking into account to codify them.

- I Images: These kinds of images are coded as if they were independent of the other images of the sequence. Furthermore, they are coded using JPEG rules. Consequently, temporal redundancy is not taken into account (intraframe compression), only spatial redundancy is important for this kind of images so the compression is small. And these kinds of images are the bigger ones.

- P Images: these images are coded using a prediction from the previous I or P image by using motion compensation. In contrast to the I images, P images to be decoded need extra information. It needs itself and the previous I or P image. During this process both spatial and temporal redundancies are considered, as a result the compression rate is bigger than the one obtained in I images.

- B Images: to coded B images it is necessary to use a previous I or P image, a next I or P image, motion compensation and motion estimation. Moreover, during the decoding process it will be necessary to use the B image, as well as the previous and next I or P image. Consequently, the process will be longer, the level of compression will be higher and the resulting image will be smaller.

- D Images: this is the last type of image. D images are intraframe images with low resolution. They are used mainly into activities that don’t require high quality rates.

It is called Group of Images (GOP) to a group of images that goes from an I image to the next I image.

Types of images

The transmit ion or storage of video sequences the necessary band-with is really big. In order to reduce this amount compression algorithms appeared.

There are two ways to compress an image:

a. Intra-frame compression.
b. Referential compression.

Intra-frame compression consists on the compression of each frame independently. The result will be a list of compressed images without any relation with the other frames of the video sequence. This is a big disadvantage because the redundancy is not taken into account. Moreover, we will not take advantage of the existent temporal redundancy between following frames. In order to solve this problem and to make the process more effective the referential compression appeared.

Referential compression is more complex than the intra-frame compression process. The process is composed by two steps:

a. A reference image is chosen.
b. Only the difference with the reference image is going to be compressed.

If the video sequence is quite long, the final compressed image can be distort as a result of the following modifications. Consequently, a reference image is introduced in the sequence in order to avoid this problem.

In MPEG compression algorithm all images are classified into three different groups depending on the compression way chosen (intra-frame or referential compression):

- Type I: every frame is compressed independently (intra-frame compression). The method used would be JPEG or a similar one.

- Type P: these frames are calculated using a previous frame.

- Type B: these frames are calculated by interpolating two different frames.

The method used to compress a video sequence is divided mainly into two different steps. Firstly, an I image is compressed independently as a key frame. Then the next P and B images are compressed using referential image compression. (They are compressed in that way in order to reduce the amount of stored information.).

The referential compression method usually is divided into the following steps:

- Each different frame is divided into “macro block”(the most common dimensions are 8x8 and 16x16).
- Each macro block has an associated position reference in the reference image.
- A motion vector is calculated for every macro block. It is used to calculate its position in the frame using only the reference image.
- In order to decide the similarity between macro blocks, one algorithm has been created
- If the algorithm establishes that the macro block has not changed (it is exactly the same as this macro block in the previous frame), then this macro block is not going to be coded.
- If the difference between the macro block of different frames is too big, then the macro blocks are going to be coded independently from the other images (either previous or next ones).

As it has shown before there are two kind of images that use the same type of compression, B or P images. The main difference between them lies in the fact that B reference images are coded from two different images that have been previously decoded, a previous and a next one. While a P reference images only depends on one previously decoded image.

The decodification process is more complex than the codification process. The main reason is because the order is really important in the decoding process. In order not to have any problem during the decodification process all images are going to be organized into groups called Group of Images (GOP). A GOP always starts with an I image and finish in the next I image.

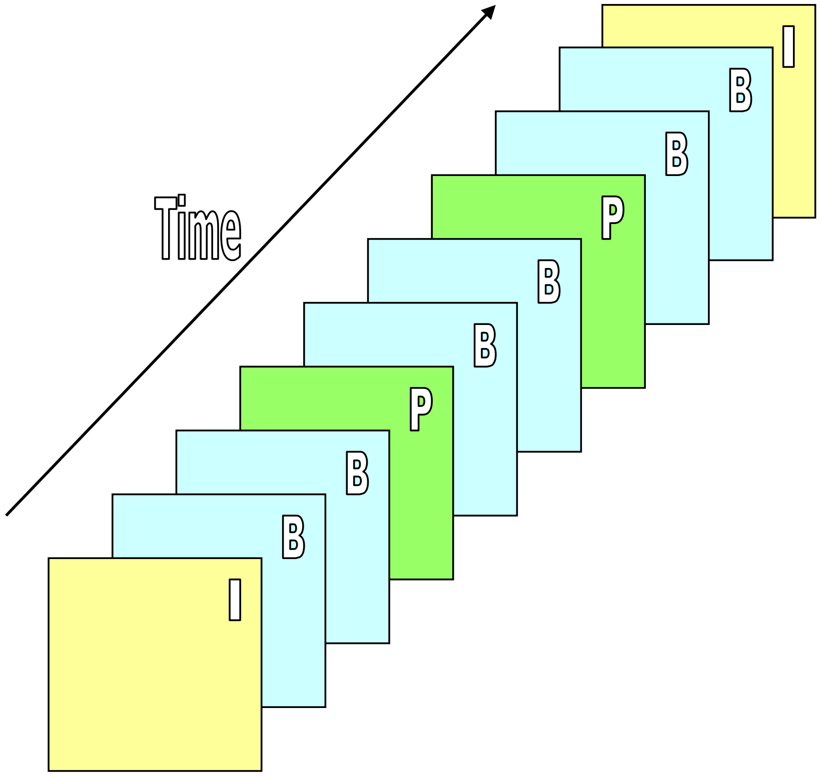


Figure 16: Example of a Group of Images (GOP)

The previous figure shows the typical structure of a Group of Images. It has a fixed organization in order to follow a decoding scheme. The proper order to decode a GOP is:

1) I images are decoded. They are independent from the others, as a result they must be the first ones in being decoded.
2) P images are decoded. They have been coded using I images as references. That is why they must wait to be decoded.
3) B images are decoded by using the I and P images that are already decoded. They are the last ones in being decoded because they have to wait until all the other information is decoded. They used the decoded information to decode this kind of images by interpoling two previous I or P decoded images.

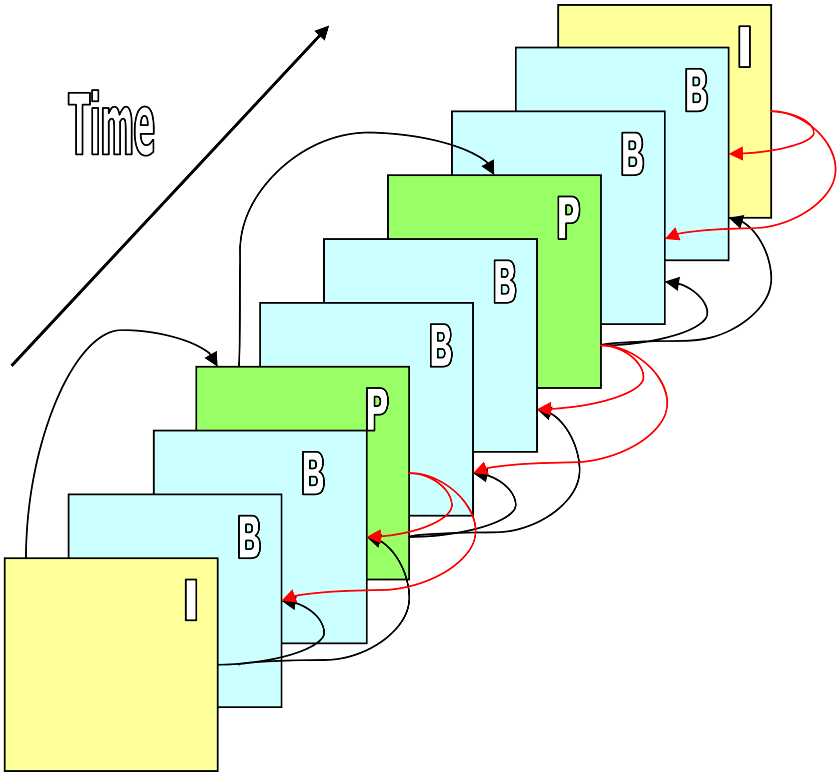


Figure 17: Scheme for decodification of a GOP

In these sequences, I images are the points where a decodification can always start. Moreover, I images have a small rate of compression and they usually needs more space than P or B images to be stored.

In contrast to I images, P or B images use a referential compression so its compression rates are higher. Although they have to wait for a decoded I or P image to be decoded.

B images reduce the needed storage space. Furthermore, they also produce a intermediate image between the two referential images used to code it. Consequently, the error rate is reduced by using B predictions more than using P predictions.

Group of images

A Group of Images (GOP) is the smallest part of a video sequence that allows a compression process to start. A GOP follows a fixed scheme that fulfills the following rules:

- It always starts with an I image.
- It is usually composed by 12 frames.
- It is formed by some I, P and B frames.

Due to this organization, if a mistake takes place into the prediction of one of the images, it will be transmitted to other of the images of the same GOP. P and B images use some other images of the same GOP to be compressed and decompressed, as a result if one mistake takes place into the GOP it will transmitted.

- If the mistake takes place in an I image → B and P images will be affected by it.
- If the mistake takes place in a P or B image → only B and images will be affected by it.

The only kind of images that is never going to be affected by other frames' mistake will be I image, because it does not depend on any other image to be compressed or decompressed.

If we consider that to create a P image it is necessary to use an I image. And that to create a B image it is necessary to use the I and the P image. We can distinguish between two kinds of visualizations of the GOP:

1. Visualization order→ how the user sees the GOP.
2. Order in the bit stream→it is the order in which the images are created.

Considering that the user will visualize the images of a GOP in an order that differs of the order in which the images are created, we must analyze how the user will see them:

Visualization order

1	2	3	4	5	6	7	8	9	10	11	12	13
I	B	B	P	B	B	P	B	B	P	B	B	I

Bit stream order (transmission):

1	4	2	3	7	5	6	10	8	9	13	11	12
I	P	B	B	P	B	B	P	B	B	I	B	B

Figure 18: Visualizations of a GOP

Predictive coding

Predictive coding is commonly used in video transmission because it allows a reduction of the transmission bit-rate. Generally between following frames there is only one part of the new frame that is different from the previous one. Motion compensation is a part of a predictive process.

When an object moves in a video sequence, its motion can be measured and this information can be use to predict the content of some next frames in the video sequence.

Common types of prediction:

Nowadays, two different kinds of prediction exist:

a. Inter-prediction: it consists on predict a new frame from a previous one. The most important idea of the inter-prediction is that only the prediction error will be coded.
b. Intra-prediction: this method consists on predicting a current block from previously coded blocks in the same frame.

These two methods are really interesting in transmission because both of them will reduce the transmission bit-rate.

In inter-prediction method only the prediction error will be coded. Firstly, the prediction error will be coded using DCT method. And secondly, we must emphasize that prediction errors have smaller energy than the original pixel values, so they can be coded with fewer bits.

The main difference between inter-prediction and intra-prediction is that inter-prediction exploits temporal redundancy while intra-prediction does it with spatial redundancy.

Finally, it must be stressed that those regions of a frame that cannot be predicted will be coded directly using DCT.

Motion compensation:

Considering the fact that following frames have a lot of common information (temporal redundancy), it would be a good idea to try to reduce this huge amount of unnecessary information. In MPEG and other codification processes, the method used to reduce the temporal redundancy is “motion compensation”. The basic idea of this method is to eliminate the temporal redundancy existing between following frames, which are part of the same sequence, in order to obtain a higher compression rate by estimating the motion between following frames.

It is important to remark that this technique obtains good results with highly correlated images (highly correlated images are images that do not present abrupt changes between them).

It must take into account that when a part of an image changes, there are two possible reasons. The first one is to consider that an object, which was in the image, has moved. The second reason is that the camera that was storing the frames has changed its position and as a result the stored frame is different from the previous one. This second reason would include a big change in the current frame, not only a change in some pixels.

(As this project is related with surveillance video cameras, we will consider that the video camera will not change its position, even though there are some existing surveillance video cameras that can do it).

Motion compensation algorithm, firstly, looks for the macro block (of the current frame that is going to be coded) in the reference frame. If this macro block appears in the reference frame, the motion vector will be coded. If this macro block does not appear in the reference frame, we will look for the most similar macro block in the reference frame and its motion vector will be coded. In case of this macro block does not appear and there is not a similar one, the whole macro block will be coded. This last case will take advantage of the spatial redundancy but not of the temporal redundancy. After this process, the error-frame will be calculated.

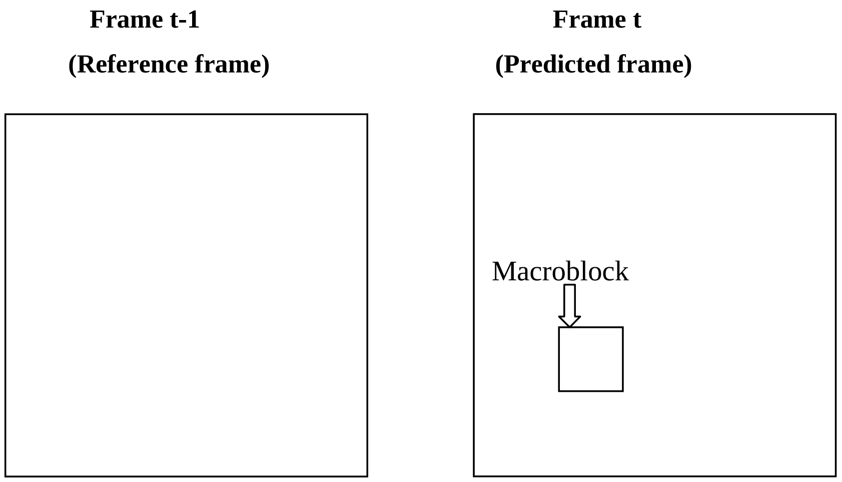


Figure 19: Reference frame and predicted frame

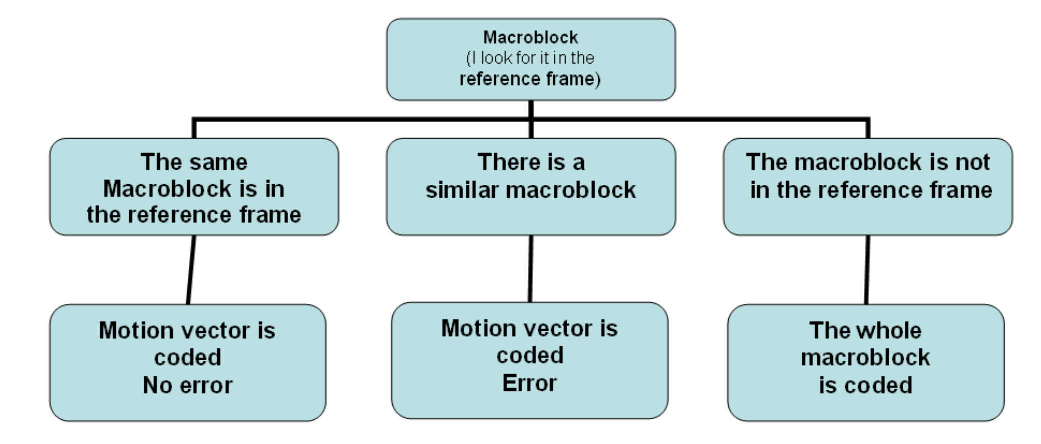


Figure 20: Scheme of motion compensation

This process would be more effective if it was applied to each single pixel instead of being applied in each macro block. However, the process would be too costly (according to time and resources).

<u>Encoder based in motion compensation:</u>

There are some methods that reduce the temporal redundancy. Even though all of them follow the same sequence of steps:

1. Each frame is divided into blocks; the most common sizes are 8x8 and 16x16.
2. A prediction is calculated to each block of the frame. It is calculated looking for where the blocks where in the previous frame (reference image). So to calculate this prediction motion compensation is used.
3. We subtract the prediction to the block of the current frame.
4. A DCT is applied to the difference between the prediction and the current block.
5. The DCT coefficients are quantified. As a result the components with high frequencies are eliminated. Then the quantified DCT coefficients are coded usually with a non-uniform coder.
6. If there is not a previous frame or there is not a similar block in the reference image, the whole block will be coded instead of coding the difference between the prediction and the current block.

Usually the blocks are going to join forming a big block called macro block. Generally macro blocks are used instead of using blocks to reduce the computational cost. If the blocks are bigger then fewer predictions must be calculated.

In fact, the whole process is applied to macro blocks (instead of blocks). The motion vector of each element of the macro block is calculated. So for each element of the macro block we are going to store:

- Outcome of the DCT and the quantification to each block of the difference between the prediction and the current block (macro block).
- Motion vectors of each single element that take part in the macro block.
- Before coding the outcomes, a filter is applied to the DCT coefficients in order to eliminate the high frequency elements and in order to increase the compression rate.

MPEG

MPEG is an acronym for the Moving Pictures Experts Group which was formed by the ISO (International Standards Organization) to set standards for audio and video compression and transmission[21].

Below, the data flow is described. Data rate is reduced by a compressor (coder) located in the transmitter. The compressed data are then passed though a communication channel and returned to the original rate by an expander (decoder). The ratio between the source data rate and the channel data rate is called compression factor or coding gain.

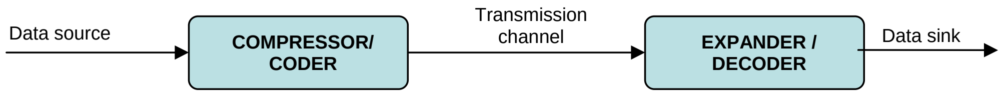


Figure 21: Scheme of transmittion

MPEG is an asymmetrical system since in this system the encoder is more complex than the decoder. The encoder needs to be algorithmic or adaptive whereas the decoder is “dump” and carries out fixed actions. This is a really useful property above all to broadcasting systems, where it exist a huge necessity of reducing the price of the decoders. As MPEG is an asymmetrical system it is possible to achieve the reduction of the number of complex encoders in exchange of the increase of the number of inexpensive decoders.

Compressor / coder	**Expander / decoder**
Algorithmic	Deterministic
(It does different things according to nature of input)	(It always does what the bit stream tells it to do)
Complex to make	Simple to make
Expensive coder	Inexpensive decoder
Few coders	Many decoders

Table 8: Comparison between coder and decoder

One of the main characteristics of MPEG system is that it is not the encoder which is standardized. The only requirement of the MPEG is that it must produce a compliant bit stream. While the way in which a decoder shall interpret the bit stream is defined.

The MPEG standards give very little information about the structure and operation of the encoder. It only requires that the bit stream follows some characteristics. Therefore, any coder constructions will meet the standard. As the encoder constructions are not revealed in the bistream, manufacturers can supply encoders using different algorithms. As a result a competition among different kinds of encoder designs will exist. This will provoke a technological evolution, the appearance of new and better designs and the existence of a huge range of encoders that differ in price and complexity. To sum up, MPEG defines the protocol of the bit stream between encoder and decoder. The decoder is defined by implication, while the encoder depends on the designer.

However, MPEG is not only a compression scheme; it also standardizes the protocol and syntax under which it is possible to combine multiplex audio data with video data to produce a digital equivalent of a television program.

MPEG has also some requirements for synchronizing.

MPEG Evolution

Modern image and video compression techniques offer the possibility to store or transmit the vast amount of data necessary to represent digital images and video in an efficient and robust way.

The main aims of international standardization of video communication systems serve to two important purposes: interoperability and economy of scale. While the final goal of video source coding is the bit-rate reduction for storage and transmission.

MPEG is a group of the ISO (International Organization fir Standardization) in charge of developing international standards for compression, decompression, processing, and coded representation of moving pictures, audio, or their combination. So far, MPEG has produced MPEG-1, MPEG-2, MPEG-4 version 1, MPEG-4 version 2 and MPEG-7.

From the beginning of 1980s a number of international video and audio standardization activities started. MPEG was established as a way to develop standards for coded representation of moving pictures, associated audio and their combination. The first standard was called MPEG-1. The next objective was to provide an appropriate video and associated audio-visual applications at substantially higher bit rates not successfully covered or envisaged by the MPEG-1 standard. MPEG-2 was given to provide higher video quality. Afterwards, PEG-4 standardization appears to address the need of universal accessibility and robustness in error-prone environments, high interactive functionality, coding of natural and synthetic data, as well as high compression efficiency [22].

Reasons why compression is necessary

Compression, bit rate reduction, data reduction and source coding are synonyms in MPEG. Their objective is to reduce the bit rate. There are some reasons that make these techniques so popular [23]:

- Compression extends the playing time of a given storage device.
- As the data amount is reduced, the storage density is reduced, too. As a result equipment can be more resistant to adverse environments and they will require less maintenance.
- In transmission systems, compression allows a reduction in bandwidth. Consequently, cost is reduced.
- If a given bandwidth is available to an uncompressed signal, compression allows faster than real-time transmission in the same bandwidth.
- If a given bandwidth is available, compression allows a better quality signal in the same bandwidth.

C CCTV

CCTV (Closed Circuit Television) is a visual surveillance technology designed for monitoring a variety of environments and activities. CCTV systems typically involve a fixed communication link between cameras and monitors.

Closed Circuit Television Cameras (CCTV) have become an important kind of crime prevention and security measure. CCTV is a measure that enables a place to be kept under surveillance remotely. Moreover, CCTV systems usually store all the images, so; finally, the system allows the post-incident analysis, which is really helpful in an investigation.

In the past decade, the use of CCTV has grown a lot, above all in UK, which is the country where CCTV systems are more wide-spread. The typical locations of CCTV cameras are usually crowded and public places like housing estates, car parks, public facilities, For instance, many central business districts in Britain are now covered by surveillance camera systems. Moreover, the video surveillance boom has been spread even inside homes. So, it can be said that the limits of CCTV are constantly extended.

Originally the purpose of the CCTV systems was to deter burglary, assault and car theft. Nowadays, the CCTV systems are used to numerous and different purposes. In the following some of them are going to be listed:

- Monitoring traffic on a bridge.
- A temporary system to carry out a traffic survey in a town centre.
- The well publicised use them at football stadiums.
- Hidden in buses to control vandalism.
- Production control in a factory.
- ...

History

In 1940, the United States´ Military, while the testing of the V2 missile, used closed circuit cameras so that they could monitor all the tests carried out in a safely way. The use of these cameras permitted a thorough investigation of the process avoiding the danger that otherwise the testing would imply. This was the first time were the CCTV systems were used.

From then on, the installation of cameras in public places became more and more popular. In 1960, United Kingdom started using CCTV as a way of controlling crowds in some public places. The use of this method of monitoring public places has been since then spreading out all around Britain until nowadays, were CCTV cameras are used in shops, buses, roads, squares, public rail stations, and other businesses.

In 1996, UK invested three quarters of the crime prevention budget on CCTV technology, diminishing thus, crime.

On the other hand, United States used its first CCTV system in 1969 in the Municipal building of New York City. This practice was also spread all over the country, but, contrary to United Kingdom, the use of CCTV was scarcely used in public spaces. Nevertheless, they started using these camera systems in stores, banks, gas stations, etc, due to the constant security threats helping thus to prevent theft.

With the passing years, CCTV systems have become easier to acquire and nowadays this method is even used in particular houses.

Definition

In its simplest form, a closed-circuit television (CCTV) system consists of a video camera, a monitor, and a recorder. More complex systems, multicamera systems, allow images to be viewed sequentially simultaneously, or on several monitors at once, depending upon the system.

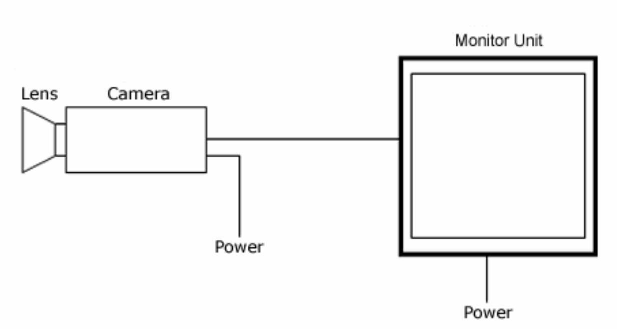


Figure 22: Scheme of the simplest CCTV system

CCTV systems can record in black and white or colour, and camera positions can be either fixed or varied by remote control to focus on activity in different locations. Zoom lenses allow either a broad view of the monitored area or selected close-ups. In addition, advances in technology enable CCTV cameras to be smaller, to use night vision, and to transmit images over the Internet.

The simplest system is a camera connected directly to a monitor by a coaxial cable, while the power that supplies the camera is provided from the monitor. This system is known as a line powered camera.

The nest development was to incorporate the outputs from four cameras into the monitor. These could be set to sequence automatically through the cameras or any camera could be held selectively. However, there were some mistakes in this system as, for example that sometimes it appeared a pause between pictures when switching. This was because the camera was powered down when not selected and it took time for the tube to heat up again. In the other hand, the advantages of this system were that it was cheap to buy and easy to install.

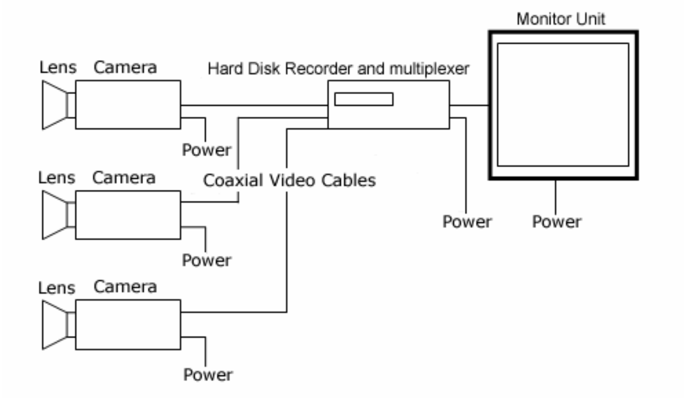


Figure 23: Scheme of a CCTV multicamera system

This is the most basic system. Of course, now there are many systems of line powered cameras.

Privacity & surveillance. Does CCTV break into people's privacy?

Some years ago and nowadays, the lack of safeness feeling into the society is causing the increment of security systems around the cities, in crowded places, banks, and airports, … This event is a result against social disorder and rising crime rates that generates the improving of the available techniques in terms of security and surveillance. The increase of surveillance systems inside the countries makes citizens feel that they are controlled or think that they are under surveillance. For that reason the dilemma appears. Does surveillance break into people's privacy?

First of all the benefits that surveillance provides must be taken into account. And if its advantages are considered then we must say that surveillance has two faces. In one hand, surveillance can be a vital tool in preventing and detecting crime. In the other hand, it breaks into people's life by saving lots of information about their daily life. As a result surveillance technologies can provide great benefits or potential threats. According to this some dilemmas related to privacy and surveillance appears.

The main point to analyse about privacy dilemmas is which kind of personal information can be used to trade (for some benefit or for convenience). Thinking about this problem lots of other dilemmas will appear as, for example, the following ones:

- Privacy as confidentiality: usually people want to keep in secret some information about themselves or their activities.
- Privacy as anonymity and privacy of identity: for different reasons, people want to keep save their identity from some actions where they are involved.
- Privacy as self-determination: our actions and behaviours must be analysed only by ourselves. Similarly, we can understand privacy as freedom to be “left alone”, this can include freedom of expression.
- Privacy as control of personal data: we might desire the right to control information about us. For example, where is it, which sees it, etc.

Each one of these dilemmas must be weighed against one or more of the following issues:

- Accountability for personal or official actions.
- The need for crime prevention and detection or for security in general.
- Public and legal standards of behaviour, which might be weighed against some personal choices.
- …

In one hand we have the possible threats that surveillance can cause related with infringing conceptions of the rights of individuals and citizens. In the other hand, surveillance has another face, it can be a source of social knowledge (a check on what citizens think is happening), a kind of security system, etc.

Nowadays, surveillance and its increase have been justified using the “war on terror” as a reason. However, its effectiveness and if it intrudes on privacy are open to scrutiny. Everyday citizens’ behaviour is more monitored and recorded in order to detect suspicious behaviours. Although sometimes surveillance is applied to other purposes that are more intrusive, as for example, the uses of traffic or congestion charge cameras.

To sum up, surveillance has two faces. The first one gives us safeness feeling, because it can provide security in cities or crowded places. However, it can be misused by breaking into people's life.

CCTV cameras. Kinds of surveillance video cameras used in CCTV systems

The objective of a video camera is to collect reflected images from objects in the environment and then to convert them into electronic signals. Cameras require more knowledge and skill to install than any other part of a CCTV system. For example, things as light sensitivity, lines of horizontal and vertical resolution, must be considered. All of these things are important because they help determine how well a camera performs in an environment.

The collected images are converted from visible light into invisible electronic signals inside a solid-state imager. These signals then are transported by one of many transmission media to the monitor, where these signals are converted back to visible light in a CCTV monitor.

Selection of cameras:

The selection criteria should take into consideration the following factors:

- Camera set should fulfil functional requirements in all demanded environmental conditions.
- It should fulfil all safety requirements ruling in the country and those specific for the application (they depends on the company and the area where the camera is used).

The main points that should be considerate when selecting CCTV cameras are: white balance for cameras, electronic iris, and long exposition to time, spectral sensitivity, external synchronization and back-up powering. Following some of these characteristics is detailed:

White balance for cameras:

White balance is the process of removing unrealistic colour casts, so that objects which appear white in persona are rendered white in your photo.

A proper camera white balance has to take into account the “colour temperature” of a light source, which refers to the relative warmth or coolness of white light.

The main problem with an incorrect “white balance” is that it can create unsightly blue, orange or even green colour casts. All these colour casts are unrealistic. Moreover, it can cause damage in portraits.

Human eyes are really good at distinguishing which colours are white or not. However, digital cameras have a great difficulty with auto white balance.

Colour temperature:

Colour temperature describes the spectrum of light which is radiated from a blackbody with that surface temperature.

(A blackbody is an object that absorbs all incidents light, not allowing it to go through and not reflecting it, too.)

Electronic iris:

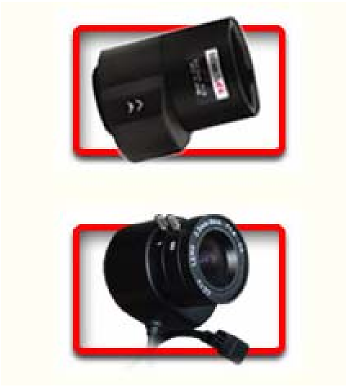
Controlling the electronic iris consists on having an optimum adjustment of the CCD output signal by controlling shutter speed. It is important in changeable lighting conditions.

Spectral sensitivity:

Humans are most sensitive at the middle wavelengths, because their spectral sensitivity falls off towards the long and short wavelengths. (It is related with the illumination type.)

<u>Types of video cameras:</u>

Security cameras can be organized into five major types:



- <u>CCTV Body cameras:</u>

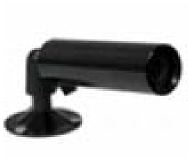
Body cameras are supplied in more conventional CCTV bodies. Moreover, body cameras are usually higher quality than bullet cameras.

Of course, there are lots of kinds of body cameras that can be classified into two groups: black and white body cameras and colour body cameras.

Figure 24: CCTV Body cameras

- <u>CCTV Bullet Cameras:</u>

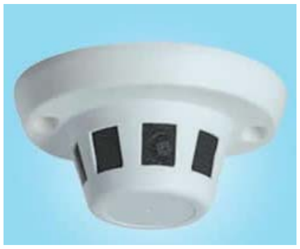
This cameras main characteristic, that gives them their name, is their shape. They are mounted in a cylindrical or “bullet” casing. There are black and white and colour bullet cameras.



They can be used indoors or outdoors. But their worst feature is that they cannot change their lenses.

Figure 25: CCTV Bullet cameras

- <u>CCTV Convert cameras:</u>



This kind of cameras is the ultimate in spy cam. Their main application is in systems that require small cameras that can be hide easily. They are used mainly indoors.

Figure 26: CCTV Convert cameras

- CCTV Dome Cameras:

CCTV Dome Cameras have a high resolution, which provide a very high quality image.

Figure 27: CCTV Dome cameras

- Infra Red CCTV Cameras:

Infra Red CCTV Cameras are used mainly in emplacement where the security at night is of high importance. It provides usually a black and white image at night, while it provides a colour image during the day.

Figure 28: Infra Red CCTV cameras

Colour vs. Black and White cameras:

Finally, in order to choose a CCTV camera the choice between colour camera and black and white must be done. This choice depends on many factors, such as environmental conditions during the whole year (seasons), size of objects in the field of view, etc.

The most important issue that must be taken into account is what the camera is for. Obviously, the colour camera will provide more detailed information than the black and white camera.

The advantages of a colour camera are mainly two. Firstly, the differences in the video sequence are noted faster, so in terms of checking surveillance video it would be a great advantage (if the check was made by a human person, not a machine). Secondly, the colour sometimes can provide more information about the situation.

In the other hand, the colour camera has also disadvantages. Firstly, their price is higher than a black and white camera. Secondly, and the most important one in terms of surveillance, in low light only the B/W (black and white) camera with high sensitivity can ensure right surveillance qualities, due to the fact that colour cameras need more light to work than B/W cameras.

As an example we can analyse a typical surveillance situation, a closed parking. Inside close parkings the level of light is low, so colour cameras would not provide the same quality image than the B/W cameras, that in this case they would be the best option.

D COLOUR SPACES AND TECHNIQUES TO STORE IMAGES

A colour model is a mathematical model that describes the different ways in which the whole range of colours can be represented as a tuple of numbers.

Colour is really subjective and personal. Trying to attribute numbers to the brain's reaction to visual stimulus is very difficult. The aim of colour spaces is to aid the process of describing colour either between people or between machines or programs.

Photoreceptor cells:

There are two kinds of photoreceptors cells in the retina: cones and rods. They have different functions and they are not evenly distributed across the retina. For instance, most of the cones are in the foves, while the rods are not located in the fovea.

The main difference between rods and cones are that cones are less sensitive to light than the rod cells. Although cones allow the perception of colour. Moreover, cones are also able to perceive finer detail and faster changes in images. Because of the fact that cones' response time to stimulus is faster than that of rods.

Basically, rod and cone cells differ in their nature and function. Rods are used primarily to see at low levels of light, whole cones are used to determine colour, depth and intensity.

Rod cells or rods of the eye can function in less intense light than any other type of photoreceptor (cone cells). They are more light-sensitive (than cones) so they are responsible for night vision. Moreover, rods have a cylindrical shape and they are concentrated at the outer edges of the retina and are used in peripherical vision.

Cone cells are somewhat shorter than rods, but wider and tapered. While rods are more numerous than cones in the retina, cones outnumber rods in the fovea.

In a human eye we can find three different kinds of cones, which differ in the spectrum of wavelengths (of photons over that they absorb). As a result, each one has different response curves that define their different response to variation in colour. An important cones' characteristic is that a single cone cannot tell colour, colour vision requires interactions of more than one type of cone. As the human eye has three kinds of cones, it is said that humans have a trichromatic vision.

The three different kinds of cones are: L-cones, M-cones and S-cones.

- L-cones respond to light of long wavelengths, peaking in the yellow region. L comes from long.
- M-cones respond mot to light of medium-wavelength, peaking at green. Its abbreviation comes from medium.
- S-cones respond to short-wavelength light, of a violet colour. It is designated S for short.

The three types have peak wavelengths near: 564-580nm, 534-545nm and 420-440 nm, respectively.

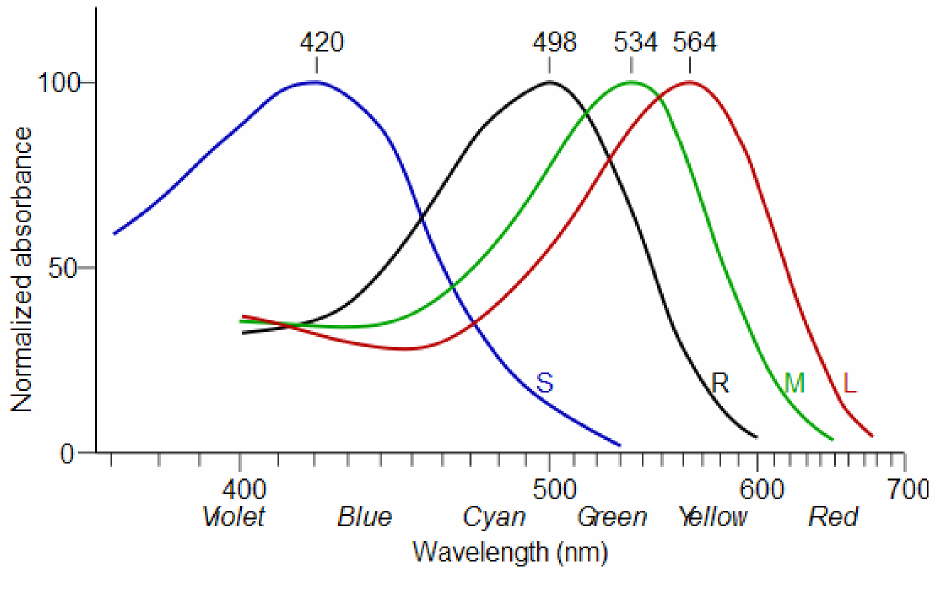


Figure 29: Spectral absorption curves of the short (S), medium (M) and long (L) wavelength pigments in human cone and rod (R) cells.

Mainly, the difference in the signals received from the three cone cell types allows the brain to perceive all possible colours.

Characteristics	Cones	Rods
Location	Mainly in the fovea	In the retina
Uses	To determine the colour	To see at low levels of light. Light-sensitive. Nocturn vision.
Kinds	3 different kinds: • L-cones • M-cones • S-cones	1 kind
Shape	"Cone" shape	Cylindrical
Perception	They perceive finer details and faster changes in the image.	They cannot perceive so finer details and less fast changes in the image.

Table 9: Characteristics of cones and rods

Characteristics that describe colours:

Colour is the brain's reaction to a specific visual stimulus. A colour can be described by measuring its spectral power distribution (the intensity of the visible electro-magnetic radiation at many discrete wavelengths). However, this leads to a large degree of redundancy. The reason why all this redundancy appears is because the retina (of the human eye) samples colours using only three broad bands, corresponding to red, green and blue light.

All the signals arriving from the sensitive cells (cones) together combined with those coming from the rods (sensitive to intensity only) give several different "sensations" of the colour. These sensations have been defined by CIE and are quoted from Hunt's book "Measuring Colour":

- Brightness: the human sensation by which an area exhibits more or less light.
- Hue: the human sensation according to which an area appears to be similar to one, or to proportions of two, of the perceived colours red, yellow, green and blue.
- Colourfulness: the human sensation according to which an area appears to exhibit more or less of its hue.
- Lightness: the sensation of an area's brightness relative to a reference white in the scene.
- Chroma: the colourfulness of an area relative to the brightness of a reference white.
- Saturation: the colourfulness of an area relative to its brightness.

The tri-chromatic theory describes how three separate lights (red, green and blue) can form any visible colour, based on the eye's use of three colour sensitive sensors (the three different types of cones).

This theory is the basis on which photography and printing operate to reproduce any colour in a scene. As well as it is how the computer colour spaces operate using three parameters to define a colour.

Colour spaces:

A colour space is a method by which a colour can be specified. As humans we may define a colour by its attributes of brightness, hue and colourfulness. However, a computer may describe a colour using the amounts of red, green and blue phosphor emission required to match a colour. While a printing press may produce a specific colour in terms of the reflectance and absorbance of cyan, magenta, yellow and black ink on the printing paper.

As a result, we can conclude that a colour can be usually defined using three co-ordinates or parameters. These parameters depend on which colour space is going to be used.

A wide range of colours can be created by the primary colours (magenta, cyan and yellow). Those colours define a colour space. There are a lot of different colour spaces; there is not a unique and universal one. Some of them are:

- The amount of magenta colour can be specified as the X axis, the amount of cyan as the Y axis and the amount of yellow as the Z axis. In that way we will have a three dimensional space in which every single colour will have a unique position.
- Another colour space would be defined by hue, saturation and brightness (as X,Y and Z axis). This colour space cannot represent all the colours.

These two colour spaces are not the most common ones. When a colour space is formally defined, the usual reference standard is CIELAB or CIEXYZ colour spaces. Which were specifically designed to describe all the colours that the average human eye can see. This is the most accurate colour space but it is too complex for everyday uses.

The ideal colour space could define all the colours (CIEXYZ), the problem is that this colour space would be too complex to use. Moreover, the average human eye cannot see all the colours so it would be a waste. As a result, the best idea is to use other colour spaces that cannot define the whole range of colours but they can represent all the colours that can be noticed by the human eye.

In this image we can observe the whole range of colours. Inside the image we can see a triangle. It represents the range of colours that human eye can notice, it is called gamut.

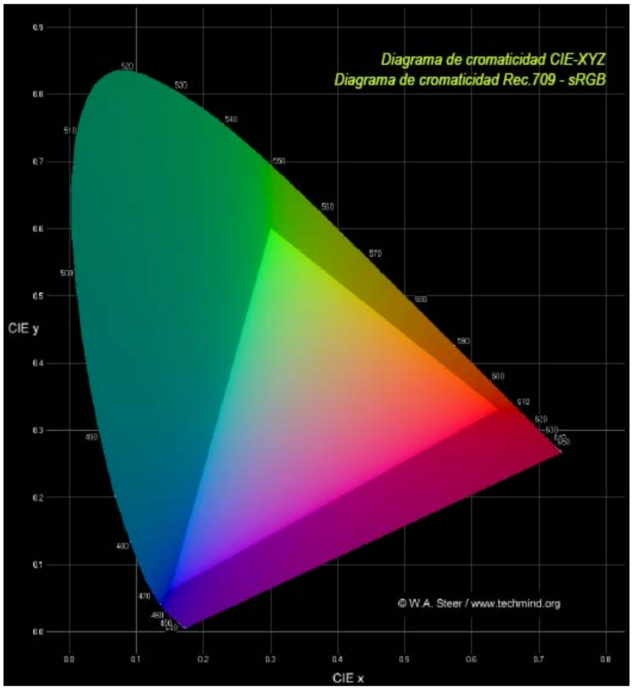


Figure 30: Chromaticity diagram CIE-XYZ

The existence of different colour spaces is due to the existence of different applications. As it has been said previously, a printing press and a computer don’t use the same colour space.

According to the CIE, a colour space is a geometric representation of colours in space, usually of three dimensions [24].

Some of these representations are made to help humans select colours, for example, the Munsell system. However, others are made to simplify data processing in machines. (The RGB system fails in this last point).

Defining a good colour space is a compromise between the availability of good primaries, the signal noise and the number of digital levels supported by the file type. There is no point in defining a very large gamut if the number of possible colours is so small that the eye will see discrete steps (banding) where uniform textures are required. A typical problem of digital systems is that sometimes the colour space gamut is much bigger than the gamut of all the output devices. This problem supposes a waste and it can be solved assigning more bits to each primary. The disadvantage is that this solution will require more computing power. However, it will minimize the banding (that appears for this problem) by down-sampling the image until obtaining a compatible image with the range of the output device.

CIE system:

CIE proposes a system that classifies colours according to the HVS (Human Visual System). Moreover, this system includes some different colour spaces that fulfill this feature. In addition, we must highlight that CIE's main objective is to allow an easy translation between the coordinates of different colour spaces.

The CIE system measured the sensitivities of the three broad bands in the eye by matching spectral colours to specific mixtures of three coloured lights. This procedure produces three values called tri-stimulus values that are unique for each colour.

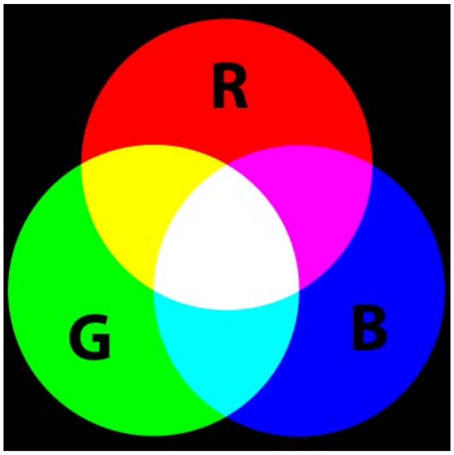
Different types of colour spaces:

There are some different colour spaces. The decision (of which colour space must be used in each moment) depends on what the colour space is for, that is to say, in which situation the colour space is going to be applied. These ones are the most common, computer related, colour spaces:

1) RGB (Red Green Blue).
2) CMY (K), Cyan Magenta Yellow (Black)
3) HSL , Hue Saturation and Lightness

4) YIQ,YUV, YCbCr, YCC (Luminance-Chrominance)
5) CIE

1) RGB(Red Green Blue)



A RGB colour space is any additive colour space based on the RGB model. The most common RGB colour space is composed by the three chromaticities of the additive primaries (red, green and blue). RGB colour space produce the gamut using those primary colours. In order to complete the specification of an RGB colour space it is necessary also a white point chromaticity and a gamma correction curve.

Figure 31: RGB colour space

RGB model consists on give a value to every primary colour. These values will vary between 0 and 255 (0 means that this primary colour is not present in the mix that compose the colour). When the number increases that means that the intensity of the colour is increasing in the mix.

In this model, a pixel is a mix of three values, each one represent the intensity of each primary colour in the final colour. For example, the most representative colours would be represented as:

	RED	GREEN	BLUE
BLACK	0	0	0
WHITE	255	255	255

Table 10: RGB Representative values

Using this system we can obtain secondary colours that are formed by mixing two primary colours and subtracting the third primary colour that has not been used in the addition.

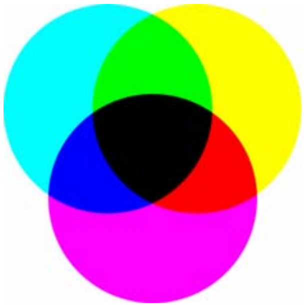
RGB model is easy to implement but it is non-linear with visual perception. RGB is a really common colour space that is used in virtually every computer system as well as television, video, etc.

The human eye is more sensitive to variations of luminance in low luminance levels than similar variation in high luminance levels. For that reason, RGB values are scaled according to this non-linear perception of the eye (using a gamma correction curve). Finally, more data triads are assigned to the lower luminance levels than to the higher ones. As a result, the RGB values' scale is close to a perceptively linear scale.

The RGB colour spaces have evolved for different reasons (for technological reasons, to fulfill professional requirements or to be adapted to a new display) generating some different and new colour spaces. Its most well-known new colour spaces are [25]:

- AdobeRGB.
- AppleRGB.
- CIE RGB.
- ColorMatch RGB.
- HDTV RGB and sRGB.
- NTSCC RGB.
- PAL/SECAM RGB.
- SGI RGB.
- Etc

2) CMY (K), Cyan Magenta Yellow (Black)



This is a subtractive model that is mainly used in printing. The main difference with RGB model is that it does not use the primary colours to form the others. The CMY(K) model uses cyan, magenta, yellow and black instead of the primary colours.

Figure 32: CMY (K) colour space

Mainly, this model is based in the light absorption. As a result, the colour of an object corresponds to the part of the light that is entering into the object and that is not absorbed by it.

The basic rules about light absorption and colour that this model follows are:

- Cyan is the opposite of red, which means that it is going to act as a filter and it is going to absorb all the red of the light.
- Magenta is the opposite to green.
- Yellow is the opposite of blue.

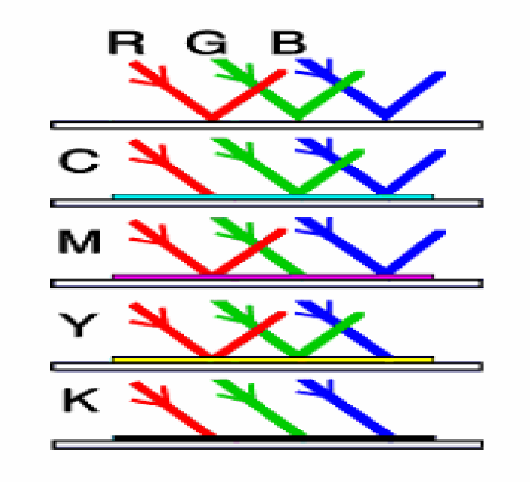
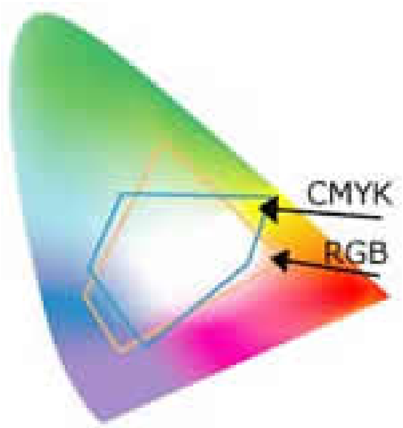


Figure 33: Scheme of colour absorption

In theory, the combination of CMY (Cyan, Magenta and Yellow) would produce black colour (that means that all the colours included in the light would be absorbed). In practice, however, this does not happen due to the light imperfections or because of the restrictions in the printing process. If we mix cyan, magenta and yellow in the same proportions, it will appear a brownish colour, not black. To solve this problem and to produce black, usually printers add black ink. That is why K appears in the name of the model.



However, if the colours produced by CMY(K) system are compared with the colours produced with RGB system, then we will realized that they are not exactly the same. Moreover, CMY(K) cannot produce the same brightness than RGB model. Other point to take into account is that the gamut produced by CMY(K) is much smaller than the RGB gamut, as we can notice in the image.

Figure 34: Comparison between RGB and CMYK gamut

If we want to translate RGB to CMY(K) or vice versa we must follow the next steps:

<u>RGB to CMY(K):</u>

Cyan=1-Red

Magenta=1-Green

Yellow=1-Blue

<u>CMY(K) to RGB:</u>

Red=1-Cyan

Green=1-Magenta

Blue=1-Yellow

These equations are only valid when the CMY(K) and RGB values are in a range between 0 and 1 (We will obtain this relative values dividing RGB or CMY(K) values between 255 that is the maximum value that the coefficients can have). In order to be able to apply this equations we must, first, normalize the RGB or CMY(K) values.

3) HSL(Hue Saturation and Lightness) and HSV(Hue Saturation and Value)

HSL and HSV are two colour spaces that are related representations of points in an RGB colour space. These two colour spaces try to describe perceptual colour relationships more accurately than RGB does, while remaining computationally method as simple as RGB.

Both HSL and HSV describe colours as points in a cylinder whole central axis ranges from black (at the bottom) to white (at the top) with neutral colours between them. The angle around the axis corresponds to "Hue", the distance from the axis corresponds to "Saturation" and the distance along the axis corresponds to "Lightness, Value or Brightness".

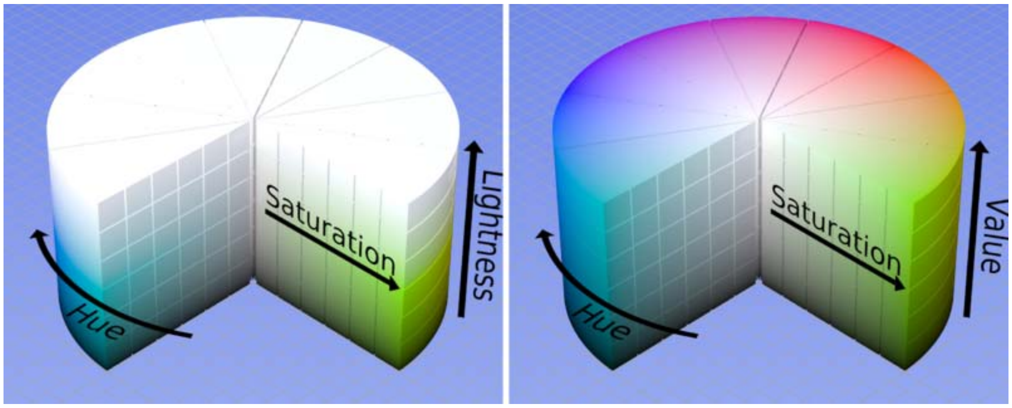


Figure 35: HSL and HSV description of colours

The two representations are really similar in purpose. Although it is important to note that "Hue" in both colour spaces refers to the same attribute, their definitions of "saturation" differ extremely. That provokes that HSV can be represented as an inverted cone of colours, while HSL is represented as a double-cone or sphere.

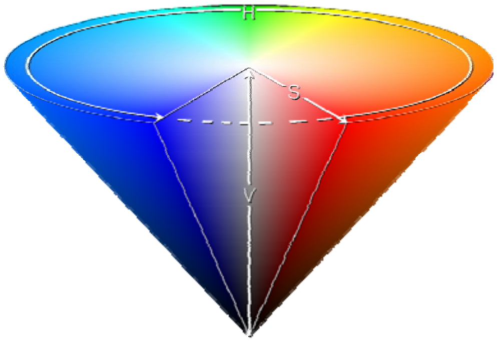


Figure 36: HSV colour space

HSL inverted cone of colours has a black point at the bottom and fully-saturated colours around the circle at the top.

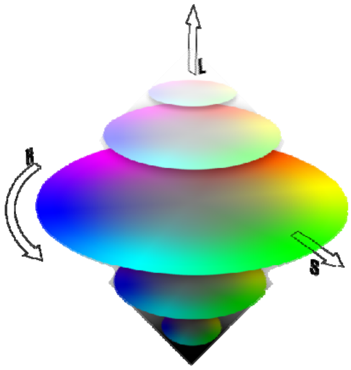


Figure 37: HSL colour space

HSL double-cone of colours has white colour at the top, black at the bottom and the fully-satured colours around of the edge of a horizontal cross-section with middle grey at its centre.

To convert RGB to HSL or HSV we will use the following equations (considering that RGB values must be in a range between 0 and 1 and that the angle will have a value between 0 and 360):

$$h = \begin{cases} 0 & \text{if } \max = \min \\ 60^\circ \times \frac{g-b}{\max - \min} + 0^\circ, & \text{if } \max = r \text{ and } g \geq b \\ 60^\circ \times \frac{g-b}{\max - \min} + 360^\circ, & \text{if } \max = r \text{ and } g < b \\ 60^\circ \times \frac{b-r}{\max - \min} + 120^\circ, & \text{if } \max = g \\ 60^\circ \times \frac{r-g}{\max - \min} + 240^\circ, & \text{if } \max = b \end{cases}$$

As it has been said before, the hue value is the same for HSL as for HSV, but the saturation value is going to be radically different so, it will have different equation that defines it.

Saturation equation for HSL model:

$$s = \begin{cases} 0 & \text{if } \max = \min \\ \frac{\max - \min}{\max + \min} = \frac{\max - \min}{2l}, & \text{if } l \leq \frac{1}{2} \\ \frac{\max - \min}{2 - (\max + \min)} = \frac{\max - \min}{2 - 2l}, & \text{if } l > \frac{1}{2} \end{cases}$$

$$l = \tfrac{1}{2}(\max + \min)$$

Saturation equation for HSV model:

$$s = \begin{cases} 0, & \text{if } \max = 0 \\ \frac{\max - \min}{\max} = 1 - \frac{\min}{\max}, & \text{otherwise} \end{cases}$$

$$v = \max$$

To convert HSL or HSV in RGB is quite complex.

Equations that must be followed to convert HSL in RGB:

$$q = \begin{cases} l \times (1 + s), & \text{if } l < \frac{1}{2} \\ l + s - (l \times s), & \text{if } l \geq \frac{1}{2} \end{cases}$$

$$p = 2 \times l - q$$

$$h_k = \frac{h}{360}$$

$$t_R = h_k + \frac{1}{3}$$

$$t_G = h_k$$

$$t_B = h_k - \frac{1}{3}$$

$$\text{if } t_C < 0 \rightarrow t_C = t_C + 1.0 \quad \text{for each } C \in \{R, G, B\}$$

$$\text{if } t_C > 1 \rightarrow t_C = t_C - 1.0 \quad \text{for each } C \in \{R, G, B\}$$

Once we have all these values, we can obtain every single colour component by applying the next equation:

$$Color_C = \begin{cases} p + ((q - p) \times 6 \times t_C), & \text{if } t_C < \frac{1}{6} \\ q, & \text{if } \frac{1}{6} \leq t_C < \frac{1}{2} \\ p + \left((q - p) \times 6 \times (\frac{2}{3} - t_C)\right), & \text{if } \frac{1}{2} \leq t_C < \frac{2}{3} \\ p, & \text{otherwise} \end{cases}$$

$$\text{for each } C \in \{R, G, B\}$$

Equations that must be followed to convert HSV to RGB:

$$h_i = \left\lfloor \frac{h}{60} \right\rfloor \mod 6$$

$$f = \frac{h}{60} - \left\lfloor \frac{h}{60} \right\rfloor$$

$$p = v \times (1 - s)$$

$$q = v \times (1 - f \times s)$$

$$t = v \times (1 - (1 - f) \times s)$$

Once we have calculated the previous values, we can obtain the triad by applying the next equation:

$$(r, g, b) = \begin{cases} (v, t, p), & \text{if } h_i = 0 \\ (q, v, p), & \text{if } h_i = 1 \\ (p, v, t), & \text{if } h_i = 2 \\ (p, q, v), & \text{if } h_i = 3 \\ (t, p, v), & \text{if } h_i = 4 \\ (v, p, q), & \text{if } h_i = 5 \end{cases}$$

4) YIQ, YUV, YcbCr, YCC (Luminance - Chrominance)

YUV system defines a colour space in terms of three components, one luminance and two chrominances. YUV model is used in the TV broadcasting systems, PAL and NTSC that are the standards all over the world.

YUV colour space is a bit unusual. The Y component determines the brightness of the colour (referred to as luminance or luma), while the U and V components –or Cb and Cr components- determine the colour itself (they are the chrominances or chroma).

Each one of the components of a single pixel in the YUV system (the luminance and the two chrominances) must be inside some limits. For instance, the Y component ranges from 0 to 1 (or to 0 to 255 in digital formats), while the U and V components range from -0.5 to 0.5 (or -128 to 127 in signed digital formats, while it would be from 0 to 255 in unsigned digital formats). Some standards further limit the ranges so the out-of bounds values indicate special information like synchronization.

One important aspect of this system is that you can remove all the information about the U and V components, and you still have the image. Although there will not be any information about colour (because you are only removing the colour information). However, the image in a grey scale will be still available without any damage. This feature is a great advantage because it allows reducing the bit rate by reducing the bit rate used to the chrominances. So, finally, we can adapt the image to the required bit rate, without destroying severely the image quality.

The human system is less sensitive to the position and motion of colour than luminance. For that reason, bandwidth can be optimized by storing more luminance detail than colour detail. At normal viewing distances, there is no perceptible loss incurred by sampling the colour detail at a lower rate.

In order to reduce the image bit rate, we can apply different samplings to the chrominances. Depending on which sampling (compression format) has been used we will distinguish among different models.

There are many different YUV formats, among them there are some that presents huge differences between them and others that are quite similar. The main aspect that must be taking into consideration is the sub sampling.

Sub sampling:

Sub sampling is a technique used to reduce the information related with colour in the image without damaging the luminance values.

Firstly, a TV image is composed by 625 horizontal lines (525 in NTSC model) in order to compress the image sub sampling will be used. The sub sampling will be applied vertically, because the image is divided into lines.

The resulting image models after sub sampling has been applied are:

- 4:4:4: the same sampling frequency is used to the three components. The luminances, as well as the chrominances do not loose information. Because the data is not sampled.

- 4:2:2: the luminance is not sampled while the chrominances are sub sampled horizontally. In this case, the chrominances sampling frequency will be half than the luminance sampling frequency.

- 4:1:1: the luminance component is not sampled. Although the chrominace components are sampled horizontally in rate 4:1. That means that chrominances sampling frequency is four times smaller than luminance sampling frequency. This is not a really common method.

- 4:2:0: the luminance is not sampled while the chrominances are sub sampled horizontally and vertically with a rate 2:1.

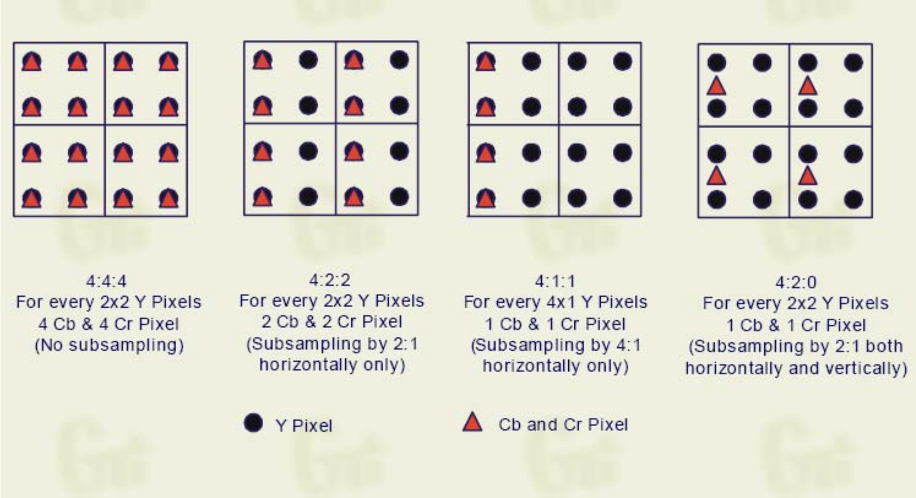


Figure 38: Sub sampling scheme

The most common sub sampling models are 4:1:1 and 4:2:0. Furthermore, they are applied in some compression models, like for example JPEG and MPEG.

<u>The most common YUV formats:</u>

Once that the sub sampling models have been explained, we can determine different YUV formats depending on which sub sampling model has been applied to the chrominance components. In fact, the existing YUV formats are:

- <u>Grey scale image.</u>
  To obtain this image, the chrominance components must be completely removed (As a result, the compressed image will not contain any information about colour). For instance, this is the maximum compression format.

The original image (with dimensions 4x4) would be:

Y_{00}	Y_{01}	Y_{02}	Y_{03}
Y_{10}	Y_{11}	Y_{12}	Y_{13}
Y_{20}	Y_{21}	Y_{22}	Y_{23}
Y_{30}	Y_{31}	Y_{23}	Y_{33}

U_{00}	U_{01}	U_{02}	U_{03}
U_{10}	U_{11}	U_{12}	U_{13}
U_{20}	U_{21}	U_{22}	U_{23}
U_{30}	U_{31}	U_{23}	U_{33}

V_{00}	V_{01}	V_{02}	V_{03}
V_{10}	V_{11}	V_{12}	V_{13}
V_{20}	V_{21}	V_{22}	V_{23}
V_{30}	V_{31}	V_{23}	V_{33}

Figure 39: Components of an image without being sub sampled

However, the real location of all the values is:

Y_{00}	U_{00}	V_{00}	Y_{01}	U_{01}	V_{01}	Y_{02}	U_{02}	V_{02}	Y_{03}	U_{03}	V_{03}
Y_{10}	U_{10}	V_{10}	Y_{11}	U_{11}	V_{11}	Y_{12}	U_{12}	V_{12}	Y_{13}	U_{13}	V_{13}
Y_{20}	U_{20}	V_{20}	Y_{21}	U_{21}	V_{21}	Y_{22}	U_{22}	V_{22}	Y_{23}	U_{23}	V_{23}
Y_{30}	U_{30}	V_{30}	Y_{31}	U_{31}	V_{31}	Y_{23}	U_{23}	V_{23}	Y_{33}	U_{33}	V_{33}

Figure 40: Real appearance of the image and its components

If we apply sub sampling to the image in order to obtain the grey scale image we will have:

Y_{00}	Y_{01}	Y_{02}	Y_{03}
Y_{10}	Y_{11}	Y_{12}	Y_{13}
Y_{20}	Y_{21}	Y_{22}	Y_{23}
Y_{30}	Y_{31}	Y_{23}	Y_{33}

Figure 41: Grey scale image

In YUV each cell needs a byte.

- <u>YUV 4:2:2:</u>

In this format each four bytes is two pixels (because each value needs a byte). Each four bytes is two luminance values, one U value and one V value. This YUV format appears when sub sampling 4:2:2 is applied to the image. As a result, the U and V components have half the horizontal resolution of the Y component.

Y_{00}	U_{00}	V_{00}	Y_{01}	U_{01}	V_{01}	Y_{02}	U_{02}	V_{02}	Y_{03}	U_{03}	V_{03}
Y_{10}	U_{10}	V_{10}	Y_{11}	U_{11}	V_{11}	Y_{12}	U_{12}	V_{12}	Y_{13}	U_{13}	V_{13}
Y_{20}	U_{20}	V_{20}	Y_{21}	U_{21}	V_{21}	Y_{22}	U_{22}	V_{22}	Y_{23}	U_{23}	V_{23}
Y_{30}	U_{30}	V_{30}	Y_{31}	U_{31}	V_{31}	Y_{23}	U_{23}	V_{23}	Y_{33}	U_{33}	V_{33}

⇩ Applying 4:2:2 sub sampling to the chrominances components

Y_{00}	U_{00}	Y_{01}	V_{00}	Y_{02}	U_{01}	Y_{03}	V_{01}
Y_{10}	U_{10}	Y_{11}	V_{10}	Y_{12}	U_{11}	Y_{13}	V_{11}
Y_{20}	U_{20}	Y_{21}	V_{20}	Y_{22}	U_{21}	Y_{23}	V_{21}
Y_{30}	U_{30}	Y_{31}	V_{30}	Y_{23}	U_{31}	Y_{33}	V_{31}

Figure 42: Generated image applying 4:2:2 sub sampling

- YUV 4:1:1:

When this method is applied to an image, twelve bytes are going to be necessary to represent eight pixels. In these twelve bytes are included four luminance values, and two pairs of U and V values. The U and V components have one fourth the horizontal resolution of the Y component.

Y_{00}	U_{00}	V_{00}	Y_{01}	U_{01}	V_{01}	Y_{02}	U_{02}	V_{02}	Y_{03}	U_{03}	V_{03}
Y_{10}	U_{10}	V_{10}	Y_{11}	U_{11}	V_{11}	Y_{12}	U_{12}	V_{12}	Y_{13}	U_{13}	V_{13}
Y_{20}	U_{20}	V_{20}	Y_{21}	U_{21}	V_{21}	Y_{22}	U_{22}	V_{22}	Y_{23}	U_{23}	V_{23}
Y_{30}	U_{30}	V_{30}	Y_{31}	U_{31}	V_{31}	Y_{23}	U_{23}	V_{23}	Y_{33}	U_{33}	V_{33}

Applying 4:1:1 sub sampling to the chrominances components

U_{00}	Y_{00}	V_{00}	Y_{01}	U_{01}	Y_{02}	V_{01}	Y_{03}	Y_{04}	Y_{05}	Y_{06}	Y_{07}
U_{10}	Y_{10}	V_{10}	Y_{11}	U_{11}	Y_{12}	V_{11}	Y_{13}	Y_{14}	Y_{15}	Y_{16}	Y_{17}
U_{20}	Y_{20}	V_{20}	Y_{21}	U_{21}	Y_{22}	V_{21}	Y_{23}	Y_{24}	Y_{25}	Y_{26}	Y_{27}
U_{30}	Y_{30}	V_{30}	Y_{31}	U_{31}	Y_{23}	V_{31}	Y_{33}	Y_{34}	Y_{35}	Y_{36}	Y_{37}

Figure 43: Generated image applying 4:1:1 sub sampling

- YUV 4:2:0:

This format is quite different to the previous ones, because it is a planar format (the others are packed formats). In the planar formats each component (luminance and the two chrominances) are separated into three sub-images or planes. Firstly, the luminance plane. Followed in memory for the U plane (there is one value U for four luminance values). The next plane in memory is the V plane.

In case that the Y plane has pad bytes after each row, then the U and V planes have half as many pad bytes after their rows. As a result, two chrominance rows (including padding) are exactly as long as one luminance row (including padding, too).

Y_{00}	U_{00}	V_{00}	Y_{01}	U_{01}	V_{01}	Y_{02}	U_{02}	V_{02}	Y_{03}	U_{03}	V_{03}
Y_{10}	U_{10}	V_{10}	Y_{11}	U_{11}	V_{11}	Y_{12}	U_{12}	V_{12}	Y_{13}	U_{13}	V_{13}
Y_{20}	U_{20}	V_{20}	Y_{21}	U_{21}	V_{21}	Y_{22}	U_{22}	V_{22}	Y_{23}	U_{23}	V_{23}
Y_{30}	U_{30}	V_{30}	Y_{31}	U_{31}	V_{31}	Y_{23}	U_{23}	V_{23}	Y_{33}	U_{33}	V_{33}

Y_{00}	Y_{01}	Y_{02}	Y_{03}
Y_{10}	Y_{11}	Y_{12}	Y_{13}
Y_{20}	Y_{21}	Y_{22}	Y_{23}
Y_{30}	Y_{31}	Y_{23}	Y_{33}

U_{00}	U_{01}
U_{10}	U_{11}

V_{00}	V_{01}
V_{10}	V_{11}

Figure 44: Generated image applying 4:2:0 sub sampling

- <u>YUV 4:1:0:</u>

As the previous one, this format is also a planar format. The only difference with the previous format is the sub sampling model that has been used. In YUV 4:1:0 the sub sampling model is 4:1:0, what means that each pair of chrominances (one U and one V) belongs to 16 pixels (a four-by-four image).

If the Y plane has pad bytes after each row, then the U and V planes have ¼ as many pad bytes after their rows. In other words, four chrominace rows (including padding) are exactly as long as one Y row (including padding).

Y_{00}	U_{00}	V_{00}	Y_{01}	U_{01}	V_{01}	Y_{02}	U_{02}	V_{02}	Y_{03}	U_{03}	V_{03}
Y_{10}	U_{10}	V_{10}	Y_{11}	U_{11}	V_{11}	Y_{12}	U_{12}	V_{12}	Y_{13}	U_{13}	V_{13}
Y_{20}	U_{20}	V_{20}	Y_{21}	U_{21}	V_{21}	Y_{22}	U_{22}	V_{22}	Y_{23}	U_{23}	V_{23}
Y_{30}	U_{30}	V_{30}	Y_{31}	U_{31}	V_{31}	Y_{23}	U_{23}	V_{23}	Y_{33}	U_{33}	V_{33}

Y_{00}	Y_{01}	Y_{02}	Y_{03}
Y_{10}	Y_{11}	Y_{12}	Y_{13}
Y_{20}	Y_{21}	Y_{22}	Y_{23}
Y_{30}	Y_{31}	Y_{23}	Y_{33}

U_{00}

V_{00}

Figure 45: Generated image applying 4:1:0 sub sampling

There are lots of different formats more, but these are the most representative ones. Moreover, YUV compression can be also called YCC, YcbCr, YpbPr or YIQ depending on the used colour considerations.

YUV-RGB conversion:

There are many different formulas to make the conversion between YUV and RGB. The existing difference among them is the coefficients in the formulas. The ITU-R 601 Standard specifies the correct coefficients.

The formulas that define the translation between RGB and YUV are:

$Y = E_Y = 0.299 \cdot E_R + 0.587 \cdot E_G + 0.114 \cdot E_B$

It must be taken into account that the E_R, E_G, E_B coefficients are the result of normalizing the primary coefficients (now, they vary between 0 and 1V) and previously the correction of the primary coefficients with the gamma function.

With these coefficients we will calculate the colour differences that later we will use to calculate the chrominance values.

$(E_R - E_Y) = 0.701 \cdot E_R – 0.587 \cdot E_G – 0.114 \cdot E_B$

$(E_B - E_Y) = - 0.299 \cdot E_R – 0.587 \cdot E_G + 0.886 \cdot E_B$

As it has been said before the coefficients of the luminance and primary colours are normalized, but it does not happen with the colour differences. These vary between [-0.701, +0.701] and [-0.886, +0.886] respectively.

The next step consists on normalizing the colour differences between -0.5 and 0.5V. That can be done by using the next formulas:

$C_r = 0.713 \cdot (E_R – Y)$

$C_b = 0.564 \cdot (E_B – Y)$

Once we have calculated C_b and C_r, uniform quantification is applied to the luminance and chrominance coefficients. However, the extremes will be kept to padding information.

The luminance coefficients will be able to take 220 different values. While the chrominance values will have 225 different levels (the null value will be 128). (The following formulas take into account that to quantify we are working with eight bits, they cannot be used if we want to quantify with a different number of bits.)

$Y' = 219 \cdot E_Y + 16$

$C_R' = 224\,(C_r) + 128 = 160 \cdot (E_R - Y) + 128$

$C_B' = 224\,(C_b) + 128 = 126 \cdot (E_{-B} - Y) + 128$

E FFMPEG

FFMPEG is mainly a fast video and audio converter. Furthermore, it is a computer program that can record, convert and stream digital audio and video in numerous formats.

The multimedia compression formats accepted in FFMEP are:

Video compression	**ISO/IEC**	**ITU-T**	**Others**
	MPEG-1 MPEG-2 MPEG-4	H.261 H.263 H.264	WMV 7 VC1 RealVideo 1.0 RealVideo 2.0
Audio compression	**ISO/IEC**	**Others**	
	MPEG-1 Layer III (MP3) MPEG-1 Layer II MPEG-1 Layer I AAC	AC3 ATRAC3 RealAudio WMA	
Image compression	**ISO/IEC/ITU-T**	**Others**	
	JPEG PNG	GIF TIFF	

Table 11: Most important multimedia compression formats accepted in FFMPEG

FFmpeg is also a command line tool that is composed of a collection of open source libraries. Moreover it allows the following actions:

- To generate videos from an array of image files.
- To convert digital audio and video between various formats.
- Streaming real time video from a TV card.

Finally, Ffmpeg can also convert from any sample rate to any other, and resize video.

PRESUPUESTO

1)	**Ejecución Material**	
	• Compra de ordenador personal (Software incluido)	2.000 €
	• Alquiler de impresora láser durante 6 meses	50 €
	• Material de oficina	150 €
	• Total de ejecución material	2.200 €
2)	**Gastos generales**	
	• 16 % sobre Ejecución Material	352 €
3)	**Beneficio Industrial**	
	• 6 % sobre Ejecución Material	132 €
4)	**Honorarios Proyecto**	
	• 32 semanas/20h semanales → 640 horas a 15 € / hora	9600 €
5)	**Material fungible**	
	• Gastos de impresión	60 €
	• Encuadernación	200 €
6)	**Subtotal del presupuesto**	
	• Subtotal Presupuesto	12060 €
7)	**I.V.A. aplicable**	
	• 16% Subtotal Presupuesto	1929.6 €
8)	**Total presupuesto**	
	• Total Presupuesto	13989,6 €

Madrid, Junio de 2008

El Ingeniero Jefe de Proyecto

Fdo.: Virginia Fernández Arguedas
Ingeniero Superior de Telecomunicación

PLIEGO DE CONDICIONES

Este documento contiene las condiciones legales que guiarán la realización, en este proyecto llamado “Summarization of Surveillance Videos based on visual activity”. En lo que sigue, se supondrá que el proyecto ha sido encargado por una empresa cliente a una empresa consultora con la finalidad de realizar dicho sistema. Dicha empresa ha debido desarrollar una línea de investigación con objeto de elaborar el proyecto. Esta línea de investigación, junto con el posterior desarrollo de los programas está amparada por las condiciones particulares del siguiente pliego.

Supuesto que la utilización industrial de los métodos recogidos en el presente proyecto ha sido decidida por parte de la empresa cliente o de otras, la obra a realizar se regulará por las siguientes:

Condiciones generales

1. La modalidad de contratación será el concurso. La adjudicación se hará, por tanto, a la proposición más favorable sin atender exclusivamente al valor económico, dependiendo de las mayores garantías ofrecidas. La empresa que somete el proyecto a concurso se reserva el derecho a declararlo desierto.

2. El montaje y mecanización completa de los equipos que intervengan será realizado totalmente por la empresa licitadora.

3. En la oferta, se hará constar el precio total por el que se compromete a realizar la obra y el tanto por ciento de baja que supone este precio en relación con un importe límite si este se hubiera fijado.

4. La obra se realizará bajo la dirección técnica de un Ingeniero Superior de Telecomunicación, auxiliado por el número de Ingenieros Técnicos y Programadores que se estime preciso para el desarrollo de la misma.

5. Aparte del Ingeniero Director, el contratista tendrá derecho a contratar al resto del personal, pudiendo ceder esta prerrogativa a favor del Ingeniero Director, quien no estará obligado a aceptarla.

6. El contratista tiene derecho a sacar copias a su costa de los planos, pliego de condiciones y presupuestos. El Ingeniero autor del proyecto autorizará con su firma las copias solicitadas por el contratista después de confrontarlas.

7. Se abonará al contratista la obra que realmente ejecute con sujeción al proyecto que sirvió de base para la contratación, a las modificaciones autorizadas por la superioridad o a las órdenes que con arreglo a sus facultades le hayan comunicado por escrito al Ingeniero Director de obras siempre que dicha obra se haya ajustado a los preceptos de los pliegos de condiciones, con arreglo a los cuales, se harán las modificaciones y la valoración de las diversas unidades sin que el importe total pueda exceder de los presupuestos aprobados. Por consiguiente, el número de unidades que se consignan en el proyecto o en el presupuesto, no podrá servirle de fundamento para entablar reclamaciones de ninguna clase, salvo en los casos de rescisión.

8. Tanto en las certificaciones de obras como en la liquidación final, se abonarán los trabajos realizados por el contratista a los precios de ejecución material que figuran en el presupuesto para cada unidad de la obra.

9. Si excepcionalmente se hubiera ejecutado algún trabajo que no se ajustase a las condiciones de la contrata pero que sin embargo es admisible a juicio del Ingeniero Director de obras, se dará conocimiento a la Dirección, proponiendo a la vez la rebaja de precios que el Ingeniero estime justa y si la Dirección resolviera aceptar la obra, quedará el contratista obligado a conformarse con la rebaja acordada.

10. Cuando se juzgue necesario emplear materiales o ejecutar obras que no figuren en el presupuesto de la contrata, se evaluará su importe a los precios asignados a otras obras o materiales análogos si los hubiere y cuando no, se discutirán entre el Ingeniero Director y el contratista, sometiéndolos a la aprobación de la Dirección. Los nuevos precios convenidos por uno u otro procedimiento, se sujetarán siempre al establecido en el punto anterior.

11. Cuando el contratista, con autorización del Ingeniero Director de obras, emplee materiales de calidad más elevada o de mayores dimensiones de lo estipulado en el proyecto, o sustituya una clase de fabricación por otra que tenga asignado mayor precio o ejecute con mayores dimensiones cualquier otra parte de las obras, o en general, introduzca en ellas cualquier modificación que sea beneficiosa a juicio del Ingeniero Director de obras, no tendrá derecho sin embargo, sino a lo que le correspondería si hubiera realizado la obra con estricta sujeción a lo proyectado y contratado.

12. Las cantidades calculadas para obras accesorias, aunque figuren por partida alzada en el presupuesto final (general), no serán abonadas sino a los precios de la contrata, según las condiciones de la misma y los proyectos particulares que para ellas se formen, o en su defecto, por lo que resulte de su medición final.

13. El contratista queda obligado a abonar al Ingeniero autor del proyecto y director de obras así como a los Ingenieros Técnicos, el importe de sus respectivos honorarios facultativos por formación del proyecto, dirección técnica y administración en su caso, con arreglo a las tarifas y honorarios vigentes.

14. Concluida la ejecución de la obra, será reconocida por el Ingeniero Director que a tal efecto designe la empresa.

15. La garantía definitiva será del 4% del presupuesto y la provisional del 2%.

16. La forma de pago será por certificaciones mensuales de la obra ejecutada, de acuerdo con los precios del presupuesto, deducida la baja si la hubiera.

17. La fecha de comienzo de las obras será a partir de los 15 días naturales del replanteo oficial de las mismas y la definitiva, al año de haber ejecutado la provisional, procediéndose si no existe reclamación alguna, a la reclamación de la fianza.

18. Si el contratista al efectuar el replanteo, observase algún error en el proyecto, deberá comunicarlo en el plazo de quince días al Ingeniero Director de obras, pues transcurrido ese plazo será responsable de la exactitud del proyecto.

19. El contratista está obligado a designar una persona responsable que se entenderá con el Ingeniero Director de obras, o con el delegado que éste designe, para todo relacionado con ella. Al ser el Ingeniero Director de obras el que interpreta el proyecto, el contratista deberá consultarle cualquier duda que surja en su realización.

20. Durante la realización de la obra, se girarán visitas de inspección por personal facultativo de la empresa cliente, para hacer las comprobaciones que se crean oportunas. Es obligación del contratista, la conservación de la obra ya ejecutada hasta la recepción de la misma, por lo que el deterioro parcial o total de ella, aunque sea por agentes atmosféricos u otras causas, deberá ser reparado o reconstruido por su cuenta.

21. El contratista, deberá realizar la obra en el plazo mencionado a partir de la fecha del contrato, incurriendo en multa, por retraso de la ejecución siempre que éste no sea debido a causas de fuerza mayor. A la terminación de la obra, se hará una recepción provisional previo reconocimiento y examen por la dirección técnica, el depositario de efectos, el interventor y el jefe de servicio o un representante, estampando su conformidad el contratista.

22. Hecha la recepción provisional, se certificará al contratista el resto de la obra, reservándose la administración el importe de los gastos de conservación de la misma hasta su recepción definitiva y la fianza durante el tiempo señalado como plazo de garantía. La recepción definitiva se hará en las mismas condiciones que la provisional, extendiéndose el acta correspondiente. El Director Técnico propondrá a la Junta Económica la devolución de la fianza al contratista de acuerdo con las condiciones económicas legales establecidas.

23. Las tarifas para la determinación de honorarios, reguladas por orden de la Presidencia del Gobierno el 19 de Octubre de 1961, se aplicarán sobre el denominado en la actualidad “Presupuesto de Ejecución de Contrata” y anteriormente llamado ”Presupuesto de Ejecución Material” que hoy designa otro concepto.

Condiciones particulares

La empresa consultora, que ha desarrollado el presente proyecto, lo entregará a la empresa cliente bajo las condiciones generales ya formuladas, debiendo añadirse las siguientes condiciones particulares:

1. La propiedad intelectual de los procesos descritos y analizados en el presente trabajo, pertenece por entero a la empresa consultora representada por el Ingeniero Director del Proyecto.

2. La empresa consultora se reserva el derecho a la utilización total o parcial de los resultados de la investigación realizada para desarrollar el siguiente proyecto, bien para su publicación o bien para su uso en trabajos o proyectos posteriores, para la misma empresa cliente o para otra.

3. Cualquier tipo de reproducción aparte de las reseñadas en las condiciones generales, bien sea para uso particular de la empresa cliente, o para cualquier otra aplicación, contará con autorización expresa y por escrito del Ingeniero Director del Proyecto, que actuará en representación de la empresa consultora.

4. En la autorización se ha de hacer constar la aplicación a que se destinan sus reproducciones así como su cantidad.

5. En todas las reproducciones se indicará su procedencia, explicitando el nombre del proyecto, nombre del Ingeniero Director y de la empresa consultora.

6. Si el proyecto pasa la etapa de desarrollo, cualquier modificación que se realice sobre él, deberá ser notificada al Ingeniero Director del Proyecto y a criterio de éste, la empresa consultora decidirá aceptar o no la modificación propuesta.

7. Si la modificación se acepta, la empresa consultora se hará responsable al mismo nivel que el proyecto inicial del que resulta el añadirla.

8. Si la modificación no es aceptada, por el contrario, la empresa consultora declinará toda responsabilidad que se derive de la aplicación o influencia de la misma.

9. Si la empresa cliente decide desarrollar industrialmente uno o varios productos en los que resulte parcial o totalmente aplicable el estudio de este proyecto, deberá comunicarlo a la empresa consultora.

10. La empresa consultora no se responsabiliza de los efectos laterales que se puedan producir en el momento en que se utilice la herramienta objeto del presente proyecto para la realización de otras aplicaciones.

11. La empresa consultora tendrá prioridad respecto a otras en la elaboración de los proyectos auxiliares que fuese necesario desarrollar para dicha aplicación industrial, siempre que no haga explícita renuncia a este hecho. En este caso, deberá autorizar expresamente los proyectos presentados por otros.

12. El Ingeniero Director del presente proyecto, será el responsable de la dirección de la aplicación industrial siempre que la empresa consultora lo estime oportuno. En caso contrario, la persona designada deberá contar con la autorización del mismo, quien delegará en él las responsabilidades que ostente.