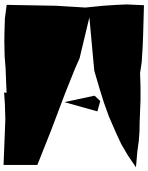


Universitat Autònoma de Barcelona

Multiple feature temporal models for the semantic characterization of video contents

A dissertation submitted by **Juan M. Sánchez Secades** at Universitat Autònoma de Barcelona to fulfil the degree of **Doctor en Informática**.

Bellaterra, September 3, 2003

Director: **Dr. Xavier Binefa Valls**
Universitat Autònoma de Barcelona
Dept. Informàtica

Universitat Autònoma de Barcelona

This document was typeset by the author using LaTeX 2_ε.

The research described in this book was carried out at the Universitat Autònoma de Barcelona.

ISBN 84–932156–7–8

Printed by Ediciones Gráficas Rey, S.L.

A los que me lo demuestran día tras día.

Acknowledgements

I want to express my gratitude to many people for their help, advice and support during the years I have been working on this thesis. Their combination has provided me a perfect "working environment".

First, I want to give special thanks to my advisor, Xavier Binefa, for his guidance and unconditional support. He introduced me in the world of Computer Vision and Multimedia research, and transmitted me his ideas and his knowledge. Thanks to him I made my first publication, I attended conferences all over the world, and I met people from everywhere. I will always be grateful for his personal support and understanding from the very first moment.

Thanks to John Kender, for welcoming me at his lab and his home, and providing his advice and guidance, which had great influence on my work.

Thanks to the members of the VISTA group at IRISA, specially to Patrick Bouthemy for welcoming me at his lab, and to Ronan Fablet. They provided me with a perfect starting point for my work.

Thanks to the members of the Computer Vision Center who supplied various help over the past years.

Thanks to my colleague, officemate and friend Xavier Orriols. During this years we have shared concerns, work and lots of fun. And much more to come.

Thanks to my other fellow students who also shared with me much more than work along these years, specially Lluís Barceló, Ramon Felip and David Guillamet.

Special thanks to my family and my dearest friends for bearing my good and bad moments and mood changes. I couldn't have done anything without you.

Gracias a mis padres, mi hermano y a los y las que comparten conmigo los momentos especiales de mi vida.

Abstract

The high-level structure of a video can be obtained once we have knowledge about the domain plus a representation of the contents that provides semantic information. In this context, intermediate-level semantic representations are defined in terms of low-level features and the information they convey about the contents of the video. Intermediate-level representations allow us to obtain semantically meaningful clusterings of shots, which are then used together with high-level domain-specific knowledge in order to obtain the structure of the video. Intermediate-level representations are usually domain-dependent as well. The descriptors involved in the representation are specifically tailored for the application, taking into account the requirements of the domain and the knowledge we have about it. This thesis proposes an intermediate-level representation of video contents that allows us to obtain semantically meaningful clusterings of shots. This representation does not depend on the domain, but still provides enough information to obtain the high-level structure of the video by combining the contributions of different low-level image features to the intermediate-level semantics.

Intermediate-level semantics are implicitly supplied by low-level features, given that a specific semantic concept generates some particular combination of feature values. The problem is to bridge the gap between observed low-level features and their corresponding hidden intermediate-level semantic concepts. Computer vision and image processing techniques are used to establish relationships between them. Other disciplines such as filmmaking and semiotics also provide important clues to discover how low-level features are used to create semantic concepts. A proper descriptor of low-level features can provide a representation of their corresponding semantic contents. Particularly, color summarized as a histogram is used to represent the appearance of objects. When this object is the background, color provides information about location. In the same way, the semantics conveyed by a description of motion have been analyzed in this thesis. A summary of motion features as a temporal cooccurrence matrix provides information about camera operation and the type of shot in terms of relative distance of the camera to the subject matter.

The main contribution of this thesis is a representation of visual contents in video based on summarizing the dynamic behavior of low-level features as temporal processes described by Markov chains (MC). The states of the MC are given by the values of an observed low-level feature. Unlike keyframe-based representations of shots, in-

formation from all the frames is considered in the MC modeling. Natural similarity measures such as likelihood ratios and Kullback-Leibler divergence are used to compare MC's, and thus the contents of the shots they are representing. In this framework, multiple image features can be combined in the same representation by coupling their corresponding MC's. Different ways of coupling MC's are presented, particularly the one called Coupled Markov Chains (CMC). A method to find the optimal coupling structure in terms of minimal cost and minimal loss of information is detailed in this dissertation. The loss of information is directly related to the loss of accuracy of the coupled structure to represent video contents. During the same process of computing shot representations, the boundaries between shots are detected using the same modeling of contents and similarity measures.

When color and motion features are combined, the CMC representation provides an intermediate-level semantic descriptor that implicitly contains information about objects (their identities, sizes and motion patterns), camera operation, location, type of shot, temporal relationships between elements of the scene and global activity understood as the amount of action. More complex semantic concepts emerge from the combination of these intermediate-level descriptors, such as a "talking head" that combines a close-up with the skin color of a face. Adding the location component in the News domain, talking heads can be further classified into "anchors" (located in the studio) and "correspondents" (located outdoors). These and many other semantically meaningful categories are discovered when shots represented using the CMC model are clustered in an unsupervised way. Well-defined concepts are given by compact clusters, which can be determined by a measure of their density. High level domain knowledge can then be defined by simple rules on these salient concepts, which will establish boundaries in the semantic structure of the video. The CMC modeling of video shots unifies the first steps of the video analysis process providing an intermediate-level semantically meaningful representation of contents without prior shot boundary detection.

Resumen

La estructura de alto nivel del vídeo se puede obtener a partir de conocimiento sobre el dominio más una representación de los contenidos que proporcione información semántica. En este contexto, las representaciones de la semántica de nivel medio vienen dadas en términos de características de bajo nivel y de la información que expresan acerca de los contenidos del vídeo. Las representaciones de nivel medio permiten obtener de forma automática agrupamientos semánticamente significativos de los *shots*, que son posteriormente utilizados conjuntamente con conocimientos de alto nivel específicos del dominio para obtener la estructura del vídeo. En general, las representaciones de nivel medio también dependen del dominio. Los descriptores que forman parte de la representación están específicamente diseñados para una aplicación concreta, teniendo en cuenta los requisitos del dominio y el conocimiento que tenemos del mismo. En esta tesis se propone una representación de nivel medio de los contenidos videográficos que permite obtener agrupamientos de *shots* que son semánticamente significativos. Esta representación no depende del dominio, y sin embargo aporta la información necesaria para obtener la estructura de alto nivel del vídeo, gracias a la combinación de las contribuciones de diferentes características de bajo nivel de las imágenes a la semántica de nivel medio.

La semántica de nivel medio se encuentra implícita en las características de bajo nivel, dado que un concepto semántico concreto genera una combinación específica de valores de las mismas. El problema consiste en "tender un puente sobre el vacío" entre las características de bajo nivel que se observan y sus correspondientes conceptos semánticos de nivel medio ocultos. Para establecer relaciones entre estos dos niveles, se utilizan técnicas de visión por computador y procesamiento de imágenes. Otras disciplinas como la cinematografía y la semiótica también proporcionan pistas importantes para determinar como se usan las características de bajo nivel para crear conceptos semánticos. Una descripción adecuada de las características de bajo nivel puede proporcionar una representación de sus correspondientes contenidos semánticos. Más en concreto, el color resumido en un histograma se utiliza para representar la apariencia de los objetos. Cuando el objeto es el fondo de la escena, su color aporta información sobre la localización. De la misma manera, en esta tesis se analiza la semántica que transmite una descripción del movimiento. Las características de movimiento resumidas en una matriz de coocurrencias temporales proporcionan información sobre las operaciones de la cámara y el tipo de toma (primer plano, etc.) en función de la distancia relativa entre la cámara y los objetos filmados.

La principal contribución de esta tesis es una representación de los contenidos visuales del vídeo basada en el resumen del comportamiento dinámico de las características de bajo nivel como procesos temporals descritos por cadenas de Markov. Los estados de la cadena de Markov vienen dados por los valores observados de una característica de bajo nivel. A diferencia de las representaciones de los *shots* basadas en *keyframes*, el modelo de cadena de Markov considera información de todos los *frames* del *shot* en la misma representación. Las medidas de similitud naturales en un marco probabilístico, como la divergencia de Kullback-Leibler, pueden ser utilizadas para comparar cadenas de Markov y, por tanto, el contenido de los *shots* que representan. En la misma representación se pueden combinar múltiples características de las imágenes mediante el acoplamiento de sus correspondientes cadenas. Esta tesis presenta diferentes formas de acoplar cadenas de Markov, y en particular la llamada Cadenas Acopladas de Markov (*Coupled Markov Chains*, CMC). También se detalla un método para encontrar la estructura de acoplamiento óptima en términos de coste mínimo y mínima pérdida de información, ya que esta merma se relaciona directamente con la pérdida de precisión de la estructura acoplada para representar contenidos de vídeo. Durante el proceso de cálculo de las representaciones de los *shots* se detectan las fronteras entre éstos usando el mismo modelo y medidas de similitud.

Cuando las características de color y movimiento se combinan, la representación en cadenas acopladas de Markov proporciona un descriptor semántico de nivel medio que contiene información implícita sobre objetos (sus identidades, tamaños y patrones de movimiento), movimiento de cámara, localización, tipo de toma, relaciones temporales entre los elementos que componen la escena y actividad global, entendida como la cantidad de acción. Conceptos semánticos más complejos emergen de la unión de estos descriptores de nivel medio, tales como "cabeza parlante", que surge de la combinación de un primer plano con el color de la piel de la cara. Añadiendo el componente de localización en el dominio de Noticiarios, las cabezas parlantes se pueden subclasificar en "presentadores" (localizados en estudio) y "corresponsales" (localizados en exteriores). Estas y otras categorías semánticamente significativas aparecen cuando los *shots* representados usando el modelo CMC se agrupan de forma no supervisada. Los conceptos mejor definidos se corresponden con grupos compactos, que pueden ser detectados usando una medida de densidad. Conocimiento de alto nivel sobre el dominio se puede definir mediante simples reglas basadas en estos conceptos, que establecen fronteras en la estructura semàntica del vídeo. El modelado de contenidos de vídeo por cadenas acopladas de Markov unifica los primeros pasos del proceso de análisis semántico de vídeo y proporciona una representación de nivel medio semánticamente significativa sin necesidad de detectar previamente las fronteras entre *shots*.

Resum

L'estructura d'alt nivell del vídeo es pot obtenir a partir de coneixement sobre el domini més una representació dels continguts que proporcioni informació semàntica. En aquest context, les representacions de la semàntica de nivell mig venen donades en termes de característiques de baix nivell i de la informació que expressen sobre els continguts del vídeo. Les representacions de nivell mig permeten obtenir de manera automàtica agrupaments semànticament significatius dels *shots*, que s'utilitzen més tard conjuntament amb coneixements d'alt nivell expecífics del domini per obtenir l'estructura del vídeo. En general, les representacions de nivell mig també depenen del domini. Els descriptors que formen part de la representació són dissenyats específicament per a una aplicació en concret, considerant els requeriments del domini i el coneixement que tenim d'aquest. En aquesta tesi es proposa una representació de nivell mig dels continguts vídeogràfics que permet obtenir agrupaments semànticament significatius dels *shots* que el composen. Aquesta representació, tot i que no depèn del domini, aporta la informació necessària per obtenir l'estructura d'alt nivell del vídeo, gràcies a la combinació de les contribucions de diferents característiques de baix nivell de les imatges a la semàntica de nivell mig.

La semàntica de nivell mig es troba implícita en les característiques de baix nivell, ja que un concepte semàntic concret genera una combinació específica de valors d'aquestes. El problema consisteix a "estendre un pont sobre el buit" entre les característiques de baix nivell que s'observen i els seus corresponents conceptes semàntics de nivell mig ocults. Per establir relacions entre aquests dos nivells, es fan servir tècniques de visió per computador i processament d'imatges. Altres disciplines com la cinematografia i la semiòtica també donen indicacions per determinar com s'utilitzen les característiques de baix nivell en la creaci de conceptes semàntics. Una descripció adequada de les característiques de baix nivell pot proporcionar una representació dels seus corresponents continguts semàntics. Concretament, el color resumit com un histograma s'empra per representar l'aparença dels objectes. Quan l'objecte és el fons de l'escena, el seu color aporta informació sobre la localització. Així mateix, la semàntica que es deriva d'una descripció del moviment és analitzada en aquesta tesi. Les característiques d'aquest resumides en una matriu de coocurrències temporals proporcionen informació sobre el moviment de la càmera i el tipus de presa (primer pla, etc.) en funció de la distància relativa entre la càmera i els objectes filmats.

La principal contribució d'aquesta tesi és una representació dels continguts visu-

als del vídeo basada en el resum del comportament dinàmic de les característiques de baix nivell com a processos temporals descrits per cadenes de Markov. Els estats de la cadena de Markov estan determinats pels valors que s'observen en una característica de baix nivell. A diferència de les representacions dels *shots* basades en *keyframes*, el model de cadena de Markov considera informació de tots els *frames* del *shot* dins de la mateixa representació. Les mesures de similitud naturals dins d'un marc probabilístic, com la divergència de Kullback-Leibler, poden ser utilitzades per comparar cadenes de Markov i, per tant, el contingut dels *shots* que representen. En la mateixa representació, es poden combinar múltiples característiques de les imatges mitjançant l'acoblament de les seves corresponents cadenes. Aquesta tesi presenta diferents formes d'acoblar cadenes de Markov, i en particular les anomenades Cadenes Acoblades de Markov (*Coupled Markov Chains*, CMC). També es detalla un mètode per trobar l'estructura d'acoblament òptima en termes de cost mínim i mínima pèrdua d'informació, puix aquesta minva incideix en la pèrdua de precisió de l'estructura acoblada per representar continguts de vídeo. Durant el procés de càlcul de les representacions dels *shots* es detecten també les fronteres entre aquests fent servir el mateix model i mesures de similitud.

Quan es combinen les característiques de color i moviment, la representació en cadenes acoblades de Markov proporciona un descriptor semàntic de nivell mig que conté informació implícita sobre objectes (les seves identitats, mides i patrons de moviment), moviment de càmera, localització, tipus de presa, relacions temporals entre els elements que composen l'escena i activitat global, entesa com a quantitat d'acció. Conceptes semàntics més complexes emergeixen de la unió d'aquests descriptors de nivell mig, com ara "cap parlant", que sorgeix de la combinació d'un primer pla amb el color de la pell de la cara. Afegint el component de localització al domini de Noticiaris, els caps parlants es poden subclassificar en "presentadors" (localitzats dins l'estudi) i "corresponsals" (localitzats en exteriors). Aquestes i altres categories semànticament significatives apareixen quan els *shots* representats fent servir el model CMC s'agrupen de forma no supervisada. Els conceptes més ben definits es corresponen amb grups compactes, que poden ser detectats mitjançant una mesura de densitat. Mitjançant simples regles sobre aquests conceptes –que estableixen fronteres en l'estructura semàntica del vídeo– es pot definir coneixement d'alt nivell sobre el domini. El modelatge de continguts de vídeo per cadenes acoblades de Markov unifica els primers passos del procés d'anàlisi semàntic de vídeo i proporciona una representació de nivell mig semànticament significativa sense que calgui detectar prèviament les fronteres entre *shots*.

Contents

List of Tables

List of Figures

Chapter 1

The semantic structure of video

The high-level structure of a video can be obtained once we have knowledge about the domain plus a representation of the contents that provides semantic information. In this context, intermediate-level semantic representations are defined in terms of low-level features and the information they convey about the contents of the video. Intermediate-level representations allow us to obtain semantically meaningful clusterings of shots, which are then used together with high-level domain-specific knowledge in order to obtain the structure of the video. Intermediate-level representations are usually domain-dependent as well. The descriptors involved in the representation are specifically tailored for the application, taking into account the requirements of the domain and the knowledge we have about it. This thesis proposes an intermediate-level representation of video contents that allows us to obtain semantically meaningful clusterings of shots. This representation does not depend on the domain, but still provides enough information to obtain the high-level structure of the video by combining the contributions of different low-level image features to the intermediate-level semantics.

1.1 Introduction

Video contents are naturally structured during the production process. This structure is hierarchical and has at least two levels. In the lowest level, we find camera shots, which are the basic units that are concatenated during edition. In the next level, shot aggregates form "story units". Depending on the domain, story units get different names: scenes in the case of feature films, news items for newscasts, plays for sports, and so on. In all cases, the shots are logically grouped in terms of contents semantics. Some video domains show a better defined structure than others. News videos are a good example of highly structured audio-visual data, while movie trailers have very little structure. Home videos have been considered unstructured by some authors [39]. However, we can still find this basic two-level

structure, as shots are grouped with the purpose of recording an event or a place [46]. Some domains may show structures with higher levels of aggregation, like sections in newscasts or game periods in sports. The automatic identification of the structure of video contents is a basic task in order to organize them, so that they can be easily located, retrieved, browsed, and summarized.

The MPEG-7 standard provides tools to describe multimedia contents. The main basic goal of MPEG-7 is to make multimedia contents searchable. MPEG-7 is intended to be generic, not application-dependant. The tools provided allow us to describe contents using different levels of abstraction, from low-level video and audio descriptors to high-level semantics and structure. Most low-level video and audio feature descriptors like color, texture, motion activity, timbre or pitch can be automatically obtained using well-known image and audio analysis techniques. However, semantic concepts and the structure of contents usually have to be defined manually.

Many recent works on multimedia content analysis have been devoted to the automatic identification of the high-level semantic structure of video. This structure depends on the domain and, in many cases, on the application. For example, focusing on the structure of news videos, we can observe that:

- Every piece of news begins with an anchor shot. Typically, two or three different anchors may appear in the same news program.
- The most important news items are summarized at the beginning of the newscast, reusing part of the video footage that will illustrate them later in the program.
- Different TV stations use the same video footage provided by news agencies in order to illustrate the news.

These facts are found in the typical structure of news videos nowadays, which is depicted in fig. 1.1. Many applications can be faced from the semantic standpoint using this structure. For example, when a user queries a news archive, he is usually looking for a particular piece of news, and not for the complete news program. Once he finds it, he may want to watch the history of that topic, that is, how it evolved along time. He may also want to compare how this topic is treated by different stations. The automatic extraction of the structure of news videos allows us to organize a news archive in order to easily provide support for the previously mentioned functionalities, and also the automatic creation of an optimal Table-of-Contents (ToC) of each news video for quick preview and browsing of its contents.

1.2 Extracting semantics from video

We have seen that knowledge about the domain has to be defined and properly represented in order to automatically obtain the structure of a video. In our example, we have implicitly defined a set of rules that are common to all news videos. For

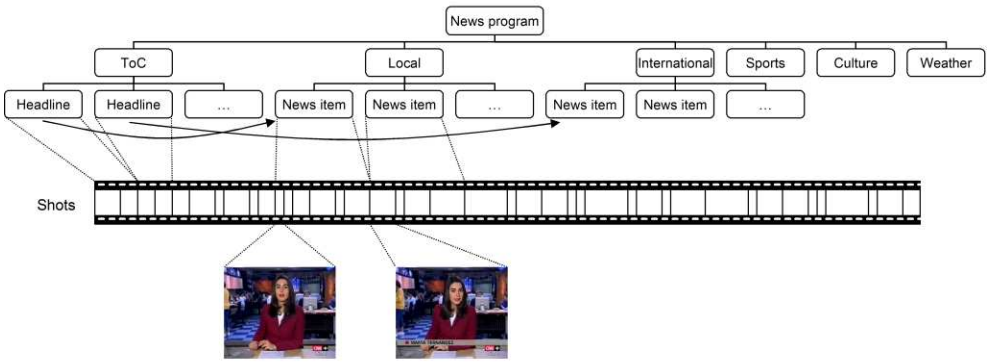


Figure 1.1: Typical semantic structure of a news video. News items begin with an anchor shot. Headlines from the table of contents (ToC) are linked to their associated news items.

instance, "*if an anchor shot is found, then a news item begins*". Domain knowledge can be represented in different ways like rules, finite state machines (FSM) or hidden Markov models (HMM). In any case, this knowledge is expressed in terms of intermediate-level semantic concepts like "anchor shot". Some of these concepts can be obtained using pattern recognition and computer vision techniques on low-level features that can be directly computed from the images. Different low-level features convey different semantics about the contents. Chapter 2 of this thesis will review the semantics that can be extracted from basic color, texture and orientation descriptors, and will analyze the kind of semantics that is carried by motion information.

There are intermediate-level semantic concepts, whose definition is very subtle. This is the case of emotions and other kind of unspoken messages conveyed in video. Research on communication theory and semiotics can be of great help in order to obtain intermediate-level representations that take into account this kind of concepts. Analyzing the relationships established by semiotics between low-level image features and unspoken messages conveyed by TV commercials, it can be shown that there exists a relationship between the 4 classes of emotions (relax, happiness, pleasantness and action) that can be conveyed by a commercial, and the visual features used in its production. Different mentions to these relationships will be done along this dissertation. Other works on this topic have been carried out by the group of Del Bimbo at University of Florence [13, 14, 11].

Mosaic images can also be seen as an intermediate-level representation of video shots, where the descriptors are not explicitly given. Mosaics contain information about location [3], objects, their trajectories and sizes [4], object and global camera-

induced motion, and even spatial and temporal descriptions of the actions in the scene (implicit in object trajectories).

Video shots, which are considered the basic semantically meaningful unit of video contents, are thus clustered in terms of their intermediate-level semantics, so that domain knowledge can be applied in order to obtain the high-level structure. This entire framework can be summarized in one simple equation:

Video structure = Domain knowledge + Intermediate-level semantics

The main problem is that not only domain knowledge is domain-specific, but also intermediate-level semantics. For instance, in our example of a news program, we have defined the concept of "anchor shot" as the fundamental view that indicates the boundaries of higher-level structures. The usual approach would be to develop an algorithm for detecting "anchor shots" in a video sequence. An anchor is a concept that only appears in newscasts, and maybe in other similar domains. Therefore, intermediate-level descriptors are also specifically tailored to the domain. This thesis provides a generic tool to characterize intermediate-level semantic concepts, so that specific algorithms for the detection of, for instance, "anchor shots" do not have to be developed.

1.3 Prior work on video structure analysis

A considerable research effort has been recently devoted to the analysis of video sequences in order to obtain their high-level semantic structure [16, 85, 2, 25, 33, 66, 77, 81, 41, 45, 46, 47]. The usual approach follows the scheme shown in fig. 1.2. First, the video sequence is partitioned into shots, which are the basic semantically meaningful units. Each shot is represented by one or several representative images or keyframes, depending on the complexity of its contents. Feature vectors are extracted from the keyframes, so that a similarity measure can be defined in order to allow the clustering of shots based on their keyframe representation. This clustering is intended to provide semantically meaningful groupings of the shots, which sometimes can be attached to semantic labels [55, 75, 86, 25, 80]. Semantics can be expressed implicitly by the feature vector, or explicitly by the definition of a set of appropriate intermediate-level semantic descriptors. A semantic clustering, together with properly defined domain-specific knowledge, leads to the high-level semantic structure of the video. A review of prior work on the different parts of this process follows.

1.3.1 Shot boundary detection

Most of the algorithms found in the literature follow the same paradigm [42]. They obtain a certain feature of each frame, and then a distance between the features of adjacent frames is computed. When this distance exceeds a certain pre-defined threshold, a shot boundary is detected. This approach has been used either on compressed

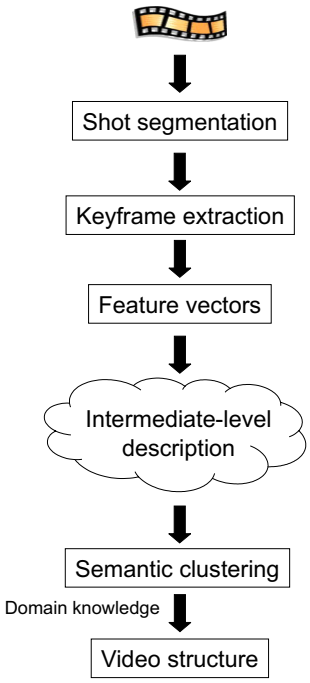


Figure 1.2: Block diagram of the usual approach to video structure analysis.

and uncompressed video. Cabedo et al. reported that the feature that provides better results is the global intensity or color histogram, and using the cosine distance [72].

These algorithms have two main problems. First, the selection of a pre-defined threshold is extremely difficult. A fixed threshold depends on the domain of the contents (sports, news, commercials, ...). O'Toole et al. proposed in [57] a semi-automatic selection of the threshold depending on the domain of video, which must be known a priori. However, even within the same domain, threshold selection requires a trade-off between recall and precision that depends on the target application. Usually, a small number of false detections are harmless, while a missed boundary can be dramatic. However, a threshold tailored to avoid missing true boundaries can report an overwhelming number of false detections.

Second, the frame-to-frame approach works well with abrupt transitions, also known as cuts. However, it is not appropriate for gradual transitions [24]. Particularly, the variation of a global intensity or color histogram between adjacent frames in a gradual transition is very subtle and difficult to detect. Zabih et al. developed in [82] an approach to gradual transitions detection based on the analysis of intensity edges. This method also has limitations due to the edge detection process. Boreczky and Rowe argue in [7] that a combination of features might produce better results than each of them individually.

1.3.2 The CECA algorithm for shot boundary detection

Edges and color

Shot segmentation algorithms in uncompressed images are mainly based on a single image feature. Combining different features during the analysis we can take advantage of their strengths, and mutually conceal their weaknesses. An algorithm that combines information from color and edges is presented in this section.

The colors around the edges of an image are a very important source of data for visual recognition. Analyzing image similarities in this way lets us capture the content of the scene, while certain variability during time is accepted. Regions around edges have interesting characteristics, as they can be seen as two different sub-regions which belong to different scene elements or to different parts of the same element. Moreover, these sub-regions have uniform colors, so that the analysis of their color content can be interpreted with respect to the elements of the scene. Imagine an object moving over an irregularly colored and textured background. In this situation, the color of the sub-region that belongs to the background may change from one frame to the next, but the one belonging to the object will remain unchanged. Therefore, the criterion used in the Combined Edges and Color Analysis (CECA) algorithm in order to determine the continuity of a region around an edge is:

Criterion 1 *Given a specific region defined by the surroundings of an edge, if the content of at least one of its sub-regions has not significantly changed from frame i to frame $i+1$, then the probability of having continuity in the scene increases.*

This criterion is checked for every region by building a color histogram of the pixels around the edge. Using a queue based algorithm, the pixels of the sub-regions at a distance less than or equal to d from their boundary are determined, so that the color histogram is built using the same number of pixels from both sub-regions (fig. 1.3). Suppose that we have the regions around edge j in frames i and $i+1$, which we call R_j^i and R_j^{i+1}, where the color in only one of its sub-regions has changed. Their associated color histograms h_j^i and h_j^{i+1} are bimodal and their intersection is the color corresponding to the unchanged sub-region. Given that both sub-regions contribute to the histograms with the same number of pixels, this intersection comprises 50% of the volume of a full histogram (which is normalized to 1). Thus, criterion 1 becomes true for regions R_j^i and R_j^{i+1} when:

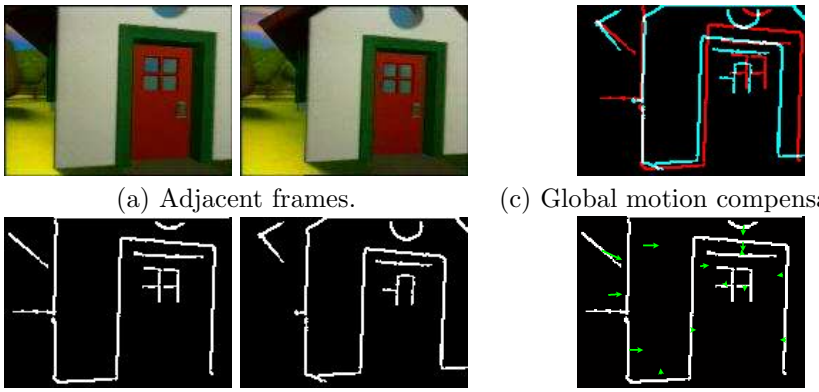
$$\sum_n \min(h_j^i(n), h_j^{i+1}(n)) \geq 0.5 \tag{1.1}$$

Finding matching edges between two images

In order to be able to apply criterion 1, we must find a correspondence between the edges in frames i and $i+1$. Edges may have moved and changed of shape due to

(a) (b) (c) (d)

Figure 1.3: (a) An image, (b) its edges, (c) the pixels used to build a color histogram around a particular edge, and (d) a 2-D projection of the bimodal color histogram obtained.



(a) Adjacent frames. (c) Global motion compensation.

(b) Edges from (a). (d) Local motion approach.

Figure 1.4: Global motion compensation is not enough when there is a camera zoom. There is a zoom out effect between the images in (a). When global motion is compensated, their edges (b) do not intersect correctly (c). Local motion compensation (d) works well, but must be followed by a second test. In our case, color continuity is checked. The arrows show the motion estimated for each edge segment.

camera operation. Global motion compensation could be applied like in [82], so that intersecting edges from the two images would be assigned to each other. However, correcting global translations may not suffice when there are multiple motions in the scene, or when there is a camera zoom, like in fig. 1.4(c). For this reason, our approach consists of breaking all edges into smaller segments and performing local edge motion estimation. In the example shown in fig. 1.4(d), all edges are moving towards the center of the image due to the zoom-out effect.

Every edge segment is located in the following frame using a correlation based search within a neighborhood of its edge image. Given a particular edge from frame i, if there is not a corresponding edge in frame $i+1$, then we consider a disappearing edge. That is, this region of the scene has changed, so that the probability of being a shot boundary increases. This consideration leads us to criterion 2, which must be applied prior to criterion 1.

Criterion 2 *Given an edge, if a corresponding one can not be found within a neighborhood in the next frame, then the probability of continuity of the scene decreases.*

Given a region R_j^i, if it does not fit criterion 2, then it is added to a set called P^i. Otherwise, criterion 1 is checked and the region is added to a second set called Q^i if it fails (see fig. 1.5).

Detecting and classifying scene breaks

These are the steps to be followed in order to detect changed regions in a frame with respect to the next one:

– Initial sets of changed regions: $P^i = \emptyset$, $Q^i = \emptyset$.
– Edge detection: threshold on the color gradient image (average of the gradient of the RGB channels).
– Edge rejection: remove small edges.
– Edge partition: divide large edges into smaller pieces.
– Region definition: define regions R_j^i, $\forall j$.
– FOR every j DO
  IF a feasible R_j^{i+1} is located THEN
    IF $\sum_n min(h_j^i(n), h_j^{i+1}(n)) < 0.5$ THEN
      $Q^i = Q^i \cup \{R_j^i\}$
    ENDIF
  ELSE
    $P^i = P^i \cup \{R_j^i\}$
  ENDIF
– ENDFOR

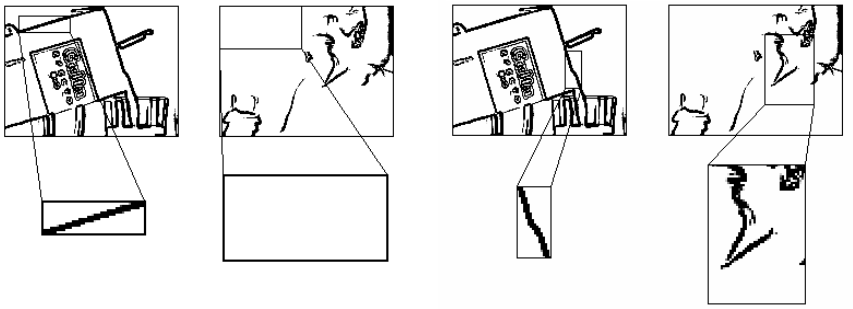
In order to obtain a global measure of the scene variation in consecutive frames a ratio of the changed regions with respect to the total regions is computed as:

$$V_1(i) = \frac{\sum_{j=0}^{L-1} |P_j^i| + \sum_{j=0}^{M-1} |Q_j^i|}{\sum_{j=0}^{N-1} |R_j^i|} \tag{1.2}$$

where $|x|$ denotes the number of pixels in region x. The contribution of each region is weighed up by the number of pixels it contains, so that variations in large regions are more significant than in small ones.

Sharp and gradual transitions are detected and classified using equation (1.2). Cuts are characterized by high values of $V_1(i)$ with the contribution of both the Q_j^i's and the P_j^i's, while they are mainly due to the P_j^i's in dissolves because new edges appear far from the locations of old edges, as observed in [82], and color variation

(a) Frames before and after a cut.



(b) Detection of a P-type region. (c) Detection of a Q-type region.

Figure 1.5: Region R_j^i may become a P_j^i or a Q_j^i in the presence of a cut between frame i and frame $i+1$ (a). If no R_j^{i+1} that correlates well with R_j^i is found in a neighbourhood (b), it becomes a P-type region. If a feasible R_j^{i+1} is found (b), its color content is checked and it becomes a Q-type region if it does not match.

between consecutive frames is low. Figure 1.5 shows these contributions in a common cut. In a fade-out, i.e. a gradual transition of the scene into black, every R_j^i turns out to be a P_j^i because all the edges disappear, but a fade-in, which is the opposite transition, can not be detected using $V_1(i)$. Since the edges gradually appear, we can first define the regions R_j^{i+1} instead and compute an equivalent measure $V_2(i)$ with respect to the regions in frame i.

We can take advantage of the need to compute $V_2(i)$ in order to make the detection more robust, using the sum of both measures $V(i) = V_1(i) + V_2(i)$ instead of only one of them. Table 1.1 summarizes the different contributions of P and Q-type regions in the computation of V_1 and V_2.

Results and discussion

The CECA algorithm was tested on a video sequence of commercials from a Spanish TV broadcast. The sequence was 11,800 frames long and contains different kinds of shot transitions, which are summarized in the ideal detection column in table 1.2. The sequence also contains plenty of synthetic images, camera operation and multiple object motions, as commonly found in commercials. Detection results are given in table 1.2 in terms of precision and recall.

We have compared the results obtained by the CECA with algorithms that only

Transition effect	Contribution to V_1		Contribution to V_2	
	P-type	Q-type	P-type	Q-type
Cut	+	+	+	+
Dissolve	–	–	–	–
Fade-out	+	0	0	0
Fade-in	0	0	+	0

Table 1.1: Contribution of P- and Q-type regions to the computation of V_1 and V_2. + and – stand for high and low contributions and 0 for no contribution at all.

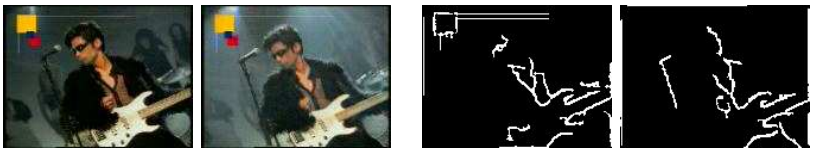
	Ideal	CECA	HI ($th = 0.25$)	HI ($th = 0.3$)	Edges
Cuts detected	246	246	210	202	234
Fade-in's detected	12	10	4	4	8
Fade-out's detected	9	7	3	3	5
Dissolves detected	18	9	0	0	1
False positives	0	45	67	48	203
Precision	1	0.86	0.76	0.81	0.55
Recall	1	0.96	0.77	0.74	0.88
Recall (only cuts)	1	1	0.85	0.82	0.95

Table 1.2: Comparative results of different scene break detection algorithms on a 11,800 frames long video sequence.

rely on one of these visual cues, either edges or color. As a color-based algorithm, we have implemented the widely used frame-to-frame color histogram difference with respect to their intersection. On the other hand, we have tested an algorithm based on the work by Zabih et al. in [82]. Our particular implementation uses the same edge detection strategy that was used in the CECA algorithm, and then compensates global motion by finding the maximum correlation position of edge images. The number of intersecting edge pixels is computed after applying a dilation to the motion compensated image, so that small local variations are allowed.

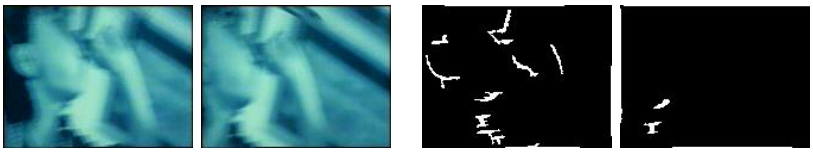
The performance of all the algorithms compared in our tests is very poor when applied to gradual transitions, especially for the simple color histogram detector. Moreover, the number of these transitions is relatively low with respect to the number of cuts in the test sequence. For these reasons, we have also considered a recall measure that only takes into account sharp transitions, and not gradual ones.

First of all, results show a great cut detection accuracy of the CECA algorithm, and a significantly better detection of gradual transitions than with single cue approaches. Fades are easier to detect than dissolves because image features completely appear or disappear. However, frame-to-frame approaches to shot boundary detection are not appropriate to deal with gradual transitions due to their extremely smooth image variations. A larger number of frames should be considered like in the twin threshold mechanism by Zhang et al. [84]. Even so, gradual transition detection results obtained using CECA are significantly good.



(a) Adjacent frames. (b) Edges of images from (a).

Figure 1.6: (a) Color is affected by dramatic luminance changes. (b) Edges may be affected as well.



(a) Adjacent frames. (b) Edges of images from (a).

Figure 1.7: Motion blur makes edge detection difficult. The CECA algorithm is affected by appearing and disappearing edges.

On the other hand, the number of false positives is kept within reasonable values, i.e. we will not be overwhelmed by a huge number of redundant key-frames. We have noticed in our tests that false positives given by the CECA algorithm are always due to one of these facts:

- Dramatic luminance changes: Camera flashes, explosions, and so on, not only cause sudden changes in image colors, but edge detection may also be affected, as shown in fig. 1.6. Single feature approaches are also affected by them.
- Fast and sudden motion: Large objects may appear and disappear from the scene. Motion blur can also make edge detection difficult, like in fig. 1.7.

These are the main sources of false positives using the color histogram detector as well. However, this algorithm has low recall. When a lower threshold is used in order to obtain a better recall rate, then precision gets worse, as shown in table 1.2. The enhanced analysis performed by the CECA algorithm lets us detect shot boundaries that go unnoticed using only color, like in fig. 1.8, without being less precise.

On the other hand, the purely edge based algorithm reports a quite good recall rate, but precision is low. When false positive detections are thoroughly analyzed, we notice that the algorithm is affected by camera operation, other than simple translations, and by multiple motions in different directions. Figures 1.4(c) and 1.9 show that global motion compensation is not suitable in these situations. The local approach used in the CECA algorithm can handle them properly. A local motion approach is prone to find matching edges even in significantly different images because the global structure of the scene is not considered. Therefore, a test that confirms them is needed, which in this case is based on the color around them.

(a) Frames before and after a cut. (b) Edges of images from (a).

Figure 1.8: The color histograms of images in (a) are very similar, so the cut between them is not detected using a color based algorithm. However, their edges (b) are significantly different.

(a) Adjacent frames. (b) Global motion compensation.

Figure 1.9: (a) The face is moving right and the background is going left. (b) Global motion compensation can not fit both of them.

1.3.3 Keyframe selection

A keyframe is a representative image from a shot. The problem is how to select the image that best represents shot contents. Sometimes, more than one image may be needed to represent all the contents in a single shot, depending on its complexity. The standard and most straightforward approach to keyframe selection is to choose images at fixed positions in the shot, like the first, the last, or a temporal sampling of a certain number of frames. However, considering that editors, authors and artists use camera operations to communicate some specific intentions, this standard keyframe selection approach may not represent properly the semantic information of the shot.

Several works can be found in the literature that perform a content-based selection of keyframes. Different visual criteria have been defined in order to select keyframes. Wolf uses motion analysis in [79]. He proposes stillness as a selection criterion, arguing that the camera stops on a new position or the characters hold gestures to emphasize their importance. Considering the same observation, Gresle and Huang propose in [32] a selection criterion based on the local minima of an activity indicator. Zhang et al. propose in [83] the use of multiple visual criteria to extract keyframes:

- Shot criterion: The first frame of a shot is always selected as a keyframe.
- Color criterion: The current frame is selected as a keyframe if it has significantly changed with respect to the last keyframe selected.
- Motion criterion: If there is a zoom in the shot, the first and last frames are selected. If there is a pan, frames with less than 30% overlap are selected.

Zhuang et al. present a clustering-based approach in [87]. The N frames of a shot are clustered into M clusters, and the frames closer to their centroids are selected as keyframes. Girgensohn et al. propose a similar approach in [29], but they also apply temporal constraints to the selection of keyframes from the clusters.

The clustering-based approach reduces redundancies and obtains more compact representations. Other ways of reducing redundancy have been proposed. Gong et al. use Singular Value Decomposition (SVD) in [30] for the summarization of shots and optimal selection of keyframes. Orriols et al. present in [56] a generative model approach to extract a reduced number of image-like data structures, which are semantically meaningful and have the ability of representing the dynamic evolution of the sequence.

1.3.4 AudiCom: An example of keyframe-based representation

AudiCom is an example of application that uses a keyframe-based representation of video contents to recognize TV commercials. The automatic recognition of commercials in TV broadcasts is an important task that can be faced from the pattern recognition standpoint. Possible applications of this technology are commercials isolation and removal, copy detection, and, like in the case of AudiCom, logging of broadcast times and durations in order to allow advertisers to check their correctness.

In AudiCom, we deal with the problem of matching video segments under some specific constraints imposed by the domain of TV commercials. Spot producers make an extensive use of synthetic production effects and other techniques due to the large amount of information they want to convey to the viewer within a very strict time constraint, usually from 10 to 40 seconds. In this sense, information must be understood from the semiotic point of view instead of from the information theory definition. This kind of information is embedded in the audio and visual streams, but must be inferred by the viewer by considering previously established semantic codes. These characteristics make automatic tasks like scene break detection to be especially difficult. Besides, there are two main sources of variability that affect commercials recognition:

- The length of commercials is reduced after a short time of being aired. Shorter spots act as reminders of their longer versions, while broadcast cost is reduced. Shorter versions are built from the original set of shots by removing and/or trimming some of them.
- Color intensity variations caused by the acquisition of video imagery from different sources, i.e. broadcasts from different TV stations or digitization using different video devices. These variations distort the appearance of images and complicate the matching process.

A commercial is represented in AudiCom by a set of keyframes. Video segment matching is thus turned into a problem of matching static images. The first image

of every shot is chosen as its keyframe. This keyframe selection strategy has the following motivations:

- When advertisers make shortened versions of their commercials by trimming shots, they usually remove frames from their end, thus relying on the memory of the viewer.
- Taking the first frame is costless, and using more complex criteria does not guarantee a better performance of the system.
- The recognition rates obtained by the system provide an empirical validation of this criterion.

Principal component analysis (PCA) is then used as a dimensionality reduction technique in order to obtain a compact feature vector for each keyframe. Matching of keyframes is then performed in the low-dimensional feature space using minimum Euclidean distance. Heuristics are introduced in the database lookup process in order to consider the sequentiality of video segments. If we already know which commercial is currently being aired because any of its shots has been identified, then the next shot will probably belong to the same commercial. If it does not, then the commercial has probably finished, and a new one may be starting. We must also consider that shots can be removed in shorter versions of the commercials, even the first one of the sequence. Therefore, the search sequence is as follows:

1. Shots from the current commercial.
2. First shot of the rest of commercials.
3. All non-first shots of all commercials.

The main drawback of appearance-based representations, like the one obtained by PCA, is their sensitivity to slight changes of view and color intensity variations. If the first frame of each shot is taken as its keyframe, changes of view might only be caused by a lack of precision in shot boundary detections. Fortunately, most of the algorithms are precise in this sense, as they are based on computing frame to frame difference measures. On the other hand, color intensity variations are a very meaningful source of variability in commercials. A color normalization step must be applied to keyframes prior to obtaining their low-dimensional representation. Several normalization algorithms have been developed. The grayworld approach has been shown to be suitable for pure recognition purposes. It is based on the assumption that the average color in all images is an ideal or canonical gray, following the diagonal model of color correction [20]. The scale factors for each RGB color channel are $R_g/\bar{R}$, $G_g/\bar{G}$ and $B_g/\bar{B}$, where (R_g, G_g, B_g) is the canonical gray RGB value and $(\bar{R}, \bar{G}, \bar{B})$ is the average image color.

The keyframes-based representation and the video segment matching strategy implemented in AudiCom have some disadvantages. First, the presence of keyframes

from monochrome shots may lead to confusion. Monochrome shots are used in many contexts, not only in commercials. Considering knowledge about the domain, these keyframes can be removed from the representation, as they will always be found between two other keyframes of the same commercial. On the other hand, the performance of the video segment recognizer relies on the accuracy of the shot boundary detector. Due to their particular complexity, we can easily come across commercials that can not be correctly represented by their keyframes because shot boundaries are not properly detected. On the other hand, shot boundary detection techniques with better recall usually have poor precision. In this case, oversegmentation produces an overwhelming number of redundant keyframes that will slow down and reduce the performance of the recognition process. For these reasons, an accurate shot boundary detection algorithm, like the previously discussed CECA, is needed.

The keyframe-based approach also disregards the temporal information contained in video sequences, which can provide very significant information about their contents. Some works try to overcome this fault by defining some kind of activity-related descriptor [66, 76] or simply by accumulating static image feature descriptors through several consecutive images of the video [60]. A different approach is based on representing a shot by its temporal fingerprint, which is a concatenation of one or several features of each image of the sequence. In this way, a string of image features is obtained, so that string edit distances can be applied [1]. Lienhart et al. presented in [44] a method for recognizing TV commercials based on fingerprints. Other work on automatic recognition of TV commercials was also carried out by Hampapur et al. in [34]. In this case, they compare different features and distance measures for matching video sequences.

1.3.5 Intermediate-level descriptions

Intermediate-level semantics can be either implicit or explicit. Low-level features implicitly contain semantic information about contents. A proper representation of low-level features can semantically characterize contents. For example, a simple color histogram can be used for object recognition [69, 27, 10]. Therefore, the representation of color as a histogram contains semantic information about the identity of the object that generates that particular distribution of colors.

When used individually, some image features (like color, texture, shape, motion or even audio in video sequences) may provide a better representation of contents than others, so that one can be more appropriate for a particular application than another. There is a believe that a combination of image features provides a better representation than the features independently. Several works have presented different ways to use multiple features [49, 59, 52]. Combining features is not straightforward. For this reason, some authors [55, 86] have used multiple features, but at different stages of a hierarchical clustering, one at a time. Other authors have used different features specifically for different classifications [75]. For example, Vailaya et al. use in [73] color for an "indoor vs. outdoor" classifier and edges for "city vs. landscape". The definition of a multiple-feature similarity measure is complex and usually requires user

intervention in order to specify the relative importance of each feature in the measure [52].

Intermediate-level descriptors that are fed to the higher-level domain knowledge reasoning system can also be explicitly defined. Qian et al. state in [65] that, in general, an explicit intermediate-level description may consist of descriptors that refer to:

Objects Whether a particular object appears in the scene or not. For instance, a person, a face, an animal, a tree, the sky/clouds, grass, buildings, and so on. Also, the size of the objects and their motion pattern, if any.

Space Spatial relationships between objects or elements in the scene (next to, on top of, ...).

Time Temporal description of the actions that happen in the scene (beginning, while, ...).

Motion Description of global camera-induced motion (zoom, pan, ...) and object motion and their trajectories.

Two more categories should be included above, due to their relevance in terms of semantics and video structure analysis:

Shot type Description of how the scene was shot (wide-angle, close-up, ...).

Affective descriptors For example, emotions, expressions, unspoken messages conveyed in video, the impact of graphics or the sense of activity.

In some cases, ways to automatically obtain these descriptors are described [65, 46]. Computer vision and pattern recognition play a main role providing algorithms for face detection, object recognition, and other semantically relevant tasks. In other cases, the intermediate-level is so complex that video shots must be annotated by hand [45].

1.3.6 Domain knowledge representation

Domain knowledge can be specified as rules about the way a video sequence is generated [2, 12, 85]. These rules are derived from film-making and from experimental observations. In many cases, the rules are based on fundamental views that indicate the boundaries of higher-level structures. This is the case of news videos, where the fundamental view "anchor shot" indicates the boundaries of the higher-level structures "news items".

Other tools that have been used for representing domain knowledge are temporal models of behavior, like finite state machines (FSM) [33, 43] or hidden Markov models (HMM) [16, 43, 45]. In the case of HMM, its learning capabilities can be used to

automatize the acquisition of knowledge up to a certain level. Probabilistic reasoning and belief networks have been also used for knowledge representation by Paek et al. in [58]. Knowledge discovery and data mining techniques were used in this context by Benitez et al. in [5] to automatically obtain high-level knowledge from image sets.

1.4 An intermediate-level representation of video contents

In this thesis, an intermediate-level representation based on Markov chains is presented. Markov chains provide a simple representation of a temporal process, and thus can be used to represent the temporal behavior of a certain pixel-based image feature. The representation is not keyframe-based, but is an aggregate of the information contained in all the frames of a shot. This framework allows us to integrate information from multiple image features into the same model of shot contents, and provides natural similarity measures. Multiple features are straightforwardly combined by considering the Cartesian product of feature state spaces. When multiple features are combined in this way, the size of the representation of the model grows exponentially. For this reason, a method to obtain a lower-cost representation with the minimum loss of information is presented. The method also proves that the amount of information (in the information theory sense) contained by the model is directly related to its accuracy to represent video contents. This method leads us to the coupled Markov chains (CMC) model, which is a simplification of the straight Cartesian-product model that establishes certain independencies between random variables in the chains. Also, the method can automatically determine the presence of irrelevant features that can be removed from the representation, and other independencies between random variables that may exist. In this way, the model can be further simplified with a minimum loss of representation accuracy.

The CMC representation of shot contents is a form of intermediate-level description. For instance, when color and motion features are considered, the CMC representation contains object descriptions, object sizes, global and object motions, temporal relationships, the type of shot, and information about location given by the background. Besides, the quality of the image features computed is not critical for the accuracy of the representation of contents, as it is based on accumulating and averaging feature values. Shot boundary detection is implicit in the generation of shot contents representations, and can also be used as a stand-alone shot segmentation technique. The contributions of the CMC modeling to the representation of visual contents in video can be summarized as follows:

- Shot contents are seen as a temporal process, thus a keyframe-based representation is avoided.
- The partition of the video into shots is automatically obtained during the generation of the CMC shot representations.
- Natural similarity measures are provided by the probabilistic framework.

- Multiple image features can be combined in the same representation, with reasonable cost in time and space and a minimum loss of representation accuracy with respect to the optimal combination.
- The representation captures the intermediate-level information that can be extracted from the image features used.
- Given that the CMC modeling provides an intermediate-level description of shot contents, an intermediate-level semantic clustering can be obtained using the similarity measures available.

The representation given by a CMC model can be used for object detection in video, intermediate-level semantic video retrieval and, particularly, to obtain an intermediate-level semantic clustering of video shots. This clustering provides a large amount of information that can be used together with domain-specific knowledge about the structure of videos in order to automatically obtain ToC's and semantic indexes. We have defined very simple rules about the structure of news videos in order to automatically obtain their structure in terms of news items, the initial summary, and the links between them. The clustering obtained allows us to automatically label the cluster of anchors, so that the structure is obtained fully automatically.

1.5 Organization of the thesis

The rest of this dissertation is organized as follows:

Chapter 2 deals with the semantics conveyed by low-level features. The semantics that can be attached to color, and how it has been used for video structure analysis will be reviewed. Then, a deep analysis of the semantics that can be inferred from motion information is presented. Also, the combination of multiple features is discussed.

Chapter 3 presents an intermediate-level representation of contents based on Markov chains. This representation will be extended to jointly account for information from multiple low-level features. This extension leads to an increase in its computational cost. A method for model structure learning will be shown, so that a reduced cost model can be obtained with minimum loss of information and representation accuracy. This accuracy will be tested in a content-based video retrieval environment. Altogether will lead us to CMC as a multiple feature temporal model of visual contents in video.

Chapter 4 shows the application of the coupled Markov chains representation of visual contents in video to obtain an intermediate-level semantically meaningful clustering of video shots that will be used to automatically extract the high-level structure in the domain of news videos. Other applications like object detection in video and video retrieval will be considered.

Chapter 5 concludes this dissertation with a summary of the main contributions of this work and insights on issues that remain open to future research.