PERGAMON

Pattern Recognition 000 (2001) 000–000

PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY

www.elsevier.com/locate/patcog

Indexing for reuse of TV news shots

M. Bertini *, A. Del Bimbo, P. Pala

Dipartimento Sistemi E Informatica, Universita' di Firenze, Via S Marta 3, 50139 Firenze, Italy

Received 16 November 2000; accepted 16 November 2000

Abstract

Broadcasters are demonstrating interest in building digital archives of their assets for reuse of archive materials for TV programs or on-line availability. This requires tools for video indexing and retrieval by content. Effective indexing by content of videos is based on the association of high-level information associated with visual data. In this paper a system is presented that enables content-based indexing and browsing of news reports; the annotation of the video stream is fully automated and is based both on visual features extracted from video shots and on textual descriptors extracted from captions and audio tracks. © 2001 Published by Elsevier Science Ltd on behalf of Pattern Recognition Society.

Keywords: Multimedia databases; Video content analysis; Content-based video retrieval; Video shots classification

1. Introduction

Broadcasters are demonstrating interest in building large digital archives of their assets for reuse of archive materials for TV programs or on-line availability to other companies and the general public. To satisfy this request there is need of systems that are able to provide efficient management of visual data in terms of storage, transmission, retrieval and browsing. Solutions to storage and transmission issues involve analysis and processing of data streams regardless of their content. Differently, effective retrieval and browsing of images and videos is based on the extraction of content level information associated with visual data and on a compact representation of retrieved shots.

While effective content-based retrieval of images is accomplished by supporting content representation through low-level image features, the same does not apply to content-based retrieval of videos, except for very limited application contexts. Instead, effective retrieval of videos must be based on high-level content descriptors.

Specific knowledge of the application content ease the extraction of high-level descriptors [1].

Recently, news videos have received great attention by the research community. This is motivated by the interest of broadcasters in building digital archives of their assets for reuse of archive materials. On the one hand, reuse of archive materials is identified as one key method of improving production quality by bringing added depth, and historical context, to recent events. On the other hand, the use of stock footage allows to produce faster the news services. An example of the first case is the reuse of shots that show the scene of a crime: they can be reused later to provide the historical context. An example of the second case is the reuse of "generic" shots, e.g. shots that show an airport may be used in a news service about an airport strike. Anyway, it is not possible to reuse all the shots of a news video: the information contained in the speech of an anchorman or in the text and the graphs of a computer graphics shot became obsolete after a short time and can be easily and inexpensively replaced by new shots. An effective reuse of archive materials is possible if the shot description is rich enough to allow retrieval by content and the content has been classified: a thorough description of the contents allows to search the shots that fit into a request of the video producer, while shot classification allows to skip those that cannot

* Corresponding author. Tel.: +39-055-4796540; fax: +39-055-4796363.

E-mail addresses: bertini@dsi.unifi.it (M. Bertini), delbimbo@dsi.unifi.it (A. Del Bimbo), pala@dsi.unifi.it (P. Pala).

PII: S0031-3203(01)00061-9

be reused. News have a rather definite structure and do not offer a wide variety either of edit effects, which are mainly cuts, or of shooting condition (e.g. illumination). The definite structure of news is suitable for content analysis and has been exploited for automatic classification of news sequences in Refs. [2–6]. In all of these systems a two stage scene classification scheme is employed. First, the video stream is parsed and video shots are extracted. Each shot is then classified according to content classes such as *newscaster*, *report*, *computer graphics*, *weather forecast*. The general approach to this type of classification relies on the definition of one or more image templates for each content class. To classify a generic shot, a *key frame* is extracted and matched against the image template of every content classes. Other works [6,7] deal with the problem of video indexing using information sources like the text of the captions and the audio track.

This is due to the fact that news videos images have an ancillary function with respect to words and video content is strongly related to textual and audio information which is contained in the audio track.

1.1. Previous work

A method for shot classification based on the syntax and the structure of news videos has been proposed in Ref. [2]. Shot classification is based on the similarity match of frames against a pre-determined set of prototype anchorman images. However, as noted in Refs. [3,5] the validity of this approach is limited by the difficulty to find a representative set of prototype anchorman images. These should account for different cases including news editions, change of dresses, modifications of studio layout. Furthermore, the method proposed is computation intensive since it requires the calculation of similarity between each frame and prototypical image of the anchorman.

In order to diminish dependency from the set of sample frames in Refs. [3,5] has been proposed to use a different approach to the definition of the anchorman frame model. In this model each anchorman frame is considered a composition of distinctive regions, like the shape of the anchorman, the caption of the reporter's name, the graphics that sometimes appear in the top third of the frame. A model of the anchorman frame is built, which accounts for the spatial distribution of basic elements and is independent of the anchorman's sex, apparel and appearance. To determine whether a shot contains an anchorman all the frames are compared with the model; if they match, they are classified as "anchorman", thus building a set of model images for each video. Only the frames of the shots that satisfy the similarity criteria according to the spatial model are then compared with the model-image set, using a new similarity measure. One of the limits of this method is that if the style of the news changes the database must be updated.

A different approach, based on frame statistics, is presented in Ref. [8]. The system uses hidden Markov models to classify frames of news videos. The classification process takes into account several clues, including feature vectors based on difference images, average frame color and audio signal. Parameters of the hidden Markov model are determined in a training stage using a ground-truth database of news videos.

The problem of text extraction has been investigated by several researchers. A method for the extraction of captions and scene texts (e.g. street names or shop names in the scene) from movies has been presented in Ref. [9]. Techniques for the extraction and OCR of caption text for the news video indexing have been examined in Ref. [7]. The first problem that must be solved for effective text extraction is to determine which frames contain captions and the position of the text in the frame. The method presented in Ref. [7] is based on the search of rectangular regions, composed by elements with sharp borders, appearing in sequences of frames; it is also based on the assumption that the captions have a high contrast on the background.

For the purpose of video content annotation, speech transcriptions has been used in the CMU Informedia project as extremely important source of information [10–12].

News-on-demand is an application within the Informedia digital video library project [6] that indexes news from TV and radio sources and allows the user to retrieve news by content. The system creates a time-aligned transcript from speech recognition and captions. The video data is segmented into news stories using the presence of silence and captions as "paragraph" boundaries, while scene breaks and keyframes are identified using algorithms based on color histograms. The CMU Sphinx-II speech recognition system is used both for the speech transcription and for the user interface of the content-based retrieval system. There is no shot classification and the speech recognition system uses the whole audio track, obtaining variable error rates that depend on the audio source [11].

Two prototypes for the construction of personalized TV news programs have been presented in Ref. [13]. The first prototype allows category-based retrieval using manual annotation provided by the news producer. The second prototype indexes the shot content using teletext data that are provided for deaf people by a French TV channel. The indexing of news videos uses the video parsing system presented in Ref. [14]. This system detects cuts computing the difference of color histograms of consecutive frames. Shots containing anchorman are identified by combining shot similarity, person detection and the "high variance factor" which accounts for the "regular spot presence" of the anchorman shots.

1.2. The news indexing and annotation system

In this paper a system for content-based indexing and annotation of news videos is presented. Videos are segmented to identify video shots. On the basis of the first frame of each shot, a statistical analysis is performed to detect which shots recur throughout the video. The shots are thus classified as newscaster shots, and the others are classified as report shots. The content of a generic report shot is described through the use of both visual and textual information and is further classified as computer graphics (*non-realistic*) or *realistic*, in order to improve the reuse of realistic shots, as needed by the broadcasters. Textual information is automatically extracted from textual captions included in the video and from speech associated with the video. Differently from Ref. [6] only anchorman shots are used for speech recognition. A retrieval engine allows the user to search by content and browse through video shots.

This paper is organized as follows: in Section 2, the video segmentation technique used to identify video shots is presented. In Section 3 the shot classification system is presented, and a comparison is carried out with respect to other techniques. In Section 4, video content description is expounded with reference to the extraction of textual information from OCR and speech recognition. Finally, in Section 5 retrieval and browsing examples are provided.

2. Video segmentation

In order to perform segmentation of news videos two problems must be dealt with: (i) avoiding incorrect identification of shot changes due to rapid motion or sudden lighting change in the scene (false positives), (ii) identification of sharp shot transitions (cuts) as well as gradual (dissolves, matte). Ref. [15] reports a thorough comparison of video segmentation algorithms. In the following, we concentrate on cuts since they are, by far, the most commonly employed edit effect in news videos. Furthermore, for the purpose of content-based indexing, it is not important to classify the edit effect, but to detect changes of visual content. Table 1 shows the number of sharp and gradual edit effects used in 4 h of news videos of the three most important Italian broadcasters.

The identification of gradual as well as sharp transitions can be performed through a cut detection algorithm, provided that the video is suitably sub-sampled in time. In fact, gradual transitions become sharp if the video is sub-sampled in the time variable since the difference between consecutive frames increases. The cut detection algorithm is developed following two distinct steps:

Table 1
Shot boundary statistics for news videos

Shots	Cuts	Diss.+ Wipe	Matte+
1797	1702 (94.7%)	95 (5.3%)	

Preliminary cut detection: Rapid motion in the scene and sudden change in lighting produce a low correlation between contiguous frames especially in case a high temporal sub-sampling rate is adopted. To avoid false cut detection, a metric has been studied which proves highly insensitive to such variations, while being reliable in detecting "true" cuts [16]. Each frame is partitioned into nine sub-frames. Each of these is represented by considering its color histogram in the HSI color space. Actually, to improve independence with respect to lighting conditions, the histogram takes into account only hue H and saturation S properties. The HSI color space has been chosen, since as reported in Ref. [17], it is a good compromise between missed detection and computational costs.

Edit effect detection is performed considering the volume of the difference of sub-frame histograms in two consecutive frames. Cuts correspond to zero crossings of the difference of the average values of the difference of the volumes. This method allows edit effect identification also when the frame color statistic remains the same but the position of the color spots is different.

To keep false positive detection low, results of the first pass are refined using a method based on video structure and shot similarity.

Cut detection refinement: The algorithm described above features a high false positive detection rate in some critical situations, such as: (i) color instabilities due to noise in the digitalization process, (ii) insertion of graphics or other changes of large zones in images, (iii) news shots recorded in critical situations, or news shots featuring sudden lighting changes. Typically, lighting changes are due either to long sequences of flashes like in press conferences, or to sudden camera movements (like *panning* and *zoom*) and free hand takes, like in reports on demonstrations or war actions.

To reduce errors due to multiple and rapid variation of visual contents of the shot, the knowledge of the specific structure of news videos has been considered. In fact, unlike other types of videos, such as commercials and movies, where the editing can reach frantic levels, in news videos the duration of the shots is long enough to let the audience "understand" the subject. Thus, there is always a minimum temporal distance τ_L between two consecutive cuts. This rule is adopted to disregard all those cuts that are less distant than τ_L seconds from the preceding cut (inter-cut time difference constraint). Furthermore, since cuts identify a change of the video content the key-frames of shots for two consecutive cuts cannot be too similar. This rule is used to disregard all those cuts whose similarity with the preceding cut exceeds a threshold τ_S (inter-cut frame similarity constraint).

Table 2
Statistics of all the videos

Shots	Detected shots	False detections	Missed detections
731	765	43 (5.9%)	9 (1.2%)

The performance of the proposed technique has been evaluated with reference to a test database composed of 12 videos from 6 Italian TV channels: RAI 1, 2 e 3, Mediaset Canale 5 and Cecchi Gori TeleMonteCarlo 1, for a total time of 2 h and 42 min.

Table 2 includes the number of video shots, cuts, gradual edit effects, falsely detected edit effects and missed detections. With respect to cut detection based exclusively on color histogram, the use of cut detection refinement results in a 37% improvement in false detection (from 69 to 43).

3. Shot classification

The main goals of shot classification are the classification of reusable and not reusable shots, and the indexing of the video. For each video shot, the first frame is used as the key-frame. Video shots are classified into two main classes: anchorman and news reports. Sub-classifications of the anchorman shots (like "weather forecasts") are obtained considering the speech content as explained in Section 4. Shot classification is a two step process: the first step classifies anchorman and report shots, using a statistical approach and motion features of the anchorman shots, without requiring any model. Then news report shots are processed in order to detect those that contain computer graphics.

Classification of anchorman and computer graphic shots is important since they cannot be reused. Fig. 1 shows an example of reusable shots: the anchorman introduces a report about accidents in the home, then after some *realistic* shots that show typical house works there is a computer graphic shot that will show some statistics. While the realistic shots are reuseable in an another report that deals about house works, the anchorman and the graphics are not, and will be replaced by another anchorman and by newer computer graphics.

3.1. Classification of anchorman shots

3.1.1. Classification based on statistical features

Shots of the anchorman are repeated at intervals of variable length throughout the video. The first step for the classification of these shots stems from this assumption and is based on the computation, for each video shot S_k, of its *shot lifetime* $\mathscr{L}(S_k)$. The shot lifetime measures the shortest temporal interval that includes all the occurrences of shots with similar visual content, within the video. Given a generic shot S_k its lifetime is computed by considering the set

$$T_k = \{t_i | \mathscr{S}(S_k, S_i) < \tau_s\},$$

where $\mathscr{S}(S_k, S_i)$ is a similarity measure applied to key-frames of shots S_k and S_i, τ_s a similarity threshold and t_i is the value of the time variable corresponding to the occurrence of the key-frame of shot S_i. The lifetime of shot S_k is defined as $\mathscr{L}(S_k) = \max(T_k) - \min(T_k)$.

Shot classification is based on fitting values of $\mathscr{L}(S_k)$ for all the video shots in a bimodal distribution. This is used to identify a threshold value τ_l that is used to classify shots into service and anchorman categories. Particularly, all the shots S_k so that $\mathscr{L}(S_k) > \tau_l$ are classified as anchorman shots, where τ_l is determined according

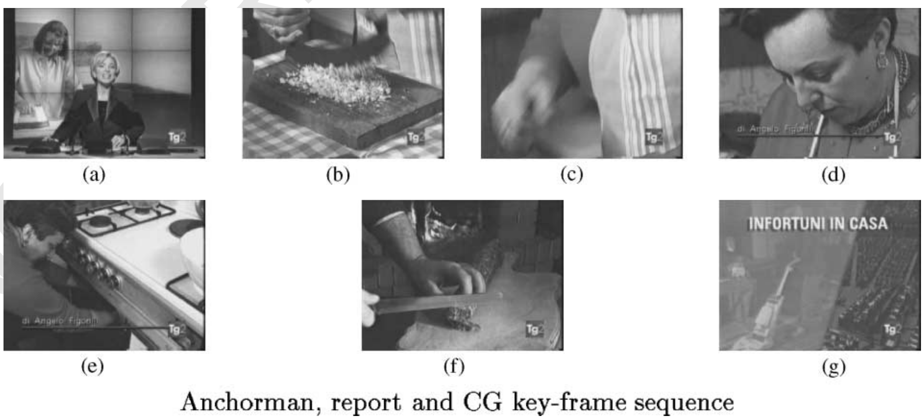


Fig. 1. Example of reusable (b, c, d, e, f,) and not reusable(a, g) shots.

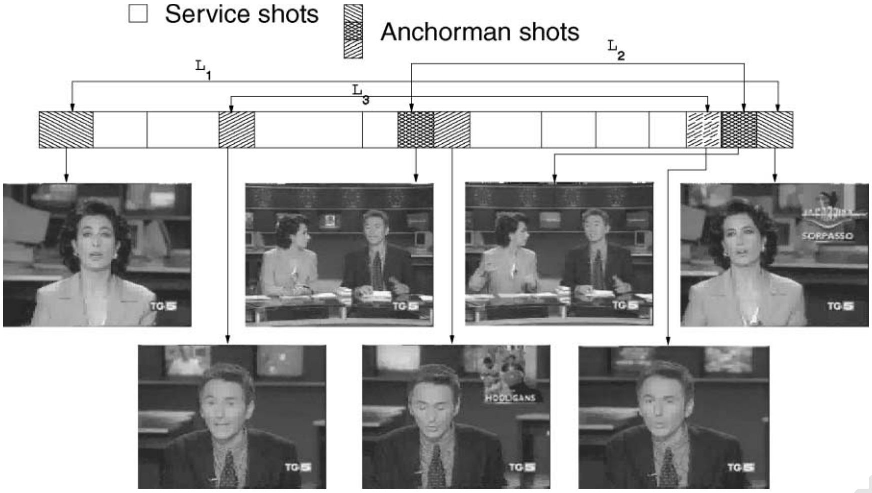


Fig. 2. Lifetime of anchorman shots.

to the statistics of the test database, and set to 4.5 s. Remaining shots are classified as news service shots.

This classification method does not rely on any pre-defined model of the anchorman shots; rather it is based on the time structure of news videos. Fig. 2 shows lifetimes for three different types of anchorman shots identified in a news video.

3.1.2. Classification based on motion features

Shot classification based on statistical features can sometimes lead to the erroneous classification of some news service shots as anchorman shots. This occurs mainly in correspondence to interviews and reports. In fact, in

Interviews: The camera alternatively takes shots of the interviewer and the interviewed people. Erroneous classification of interview shots have been discussed in Ref. [8].

In reports: A reporter describes the content shown in some shots; at the end of every shot (or series of shots) there is the shot of a new reporter describing the next series; this structure replicates the whole structure of news video. An example is shown in Fig. 3 where the recurrence of shots (a), (c) and (g) leads to erroneous classification of these shots as anchorman shots.

To avoid these errors, the preliminary classification based on statistical feature is refined considering motion features of the anchorman shots. Classification refinement stems from the assumption that in an anchorman shot, both the camera and the anchorman are almost motionless. In contrast, for both interview and news service shots, background objects and camera movements—persons and vehicles, free-hand shots, camera panning and zooming—cause relevant motion components throughout the shot.

Classification refinement is performed by computing an index of the *quantity of motion* $\mathcal{Q}_S$, for each possible anchorman shot. The algorithm for the analysis of this index takes into account the frame to frame difference between the shot key-frame f_1 and subsequent frames f_i in the shot according to

$$\mathcal{Q}_S = \sum_{f_i \in \mathcal{S}} D_i$$

with

$$D_i = \sum_{xy} d_{RGB}(f_1(x,y), f_i(x,y)), \tag{1}$$

$$d_{RGB}(f_1(x,y), f_i(x,y)) = \begin{cases} 0 & \text{if } \|f_1(x,y) - f_i(x,y)\| < \tau_{RGB}, \\ 1 & \text{if } \|f_1(x,y) - f_i(x,y)\| \geqslant \tau_{RGB}. \end{cases} \tag{2}$$

To enhance sensitivity to motion the shot is sub-sampled in time, and the frames are compared to the key-frame f_1. Only those shots whose $\mathcal{Q}_S$ does not exceed a threshold τ_Q are definitely classified as anchorman shots. By using this classification refinement, false anchorman shots shown in Fig. 3 are eliminated. In fact, shots (a), (c) and (g) feature a relevant motion component on account of camera zooming and panning and movement of people and objects in the background.

3.2. Classification of computer graphics shots

Shots classified as containing news report are processed in order to detect whether they contain computer graphics. Fig. 4 shows an example of computer graphics shot. Usually, those type of shots show information about money change rates, economic indexes and other

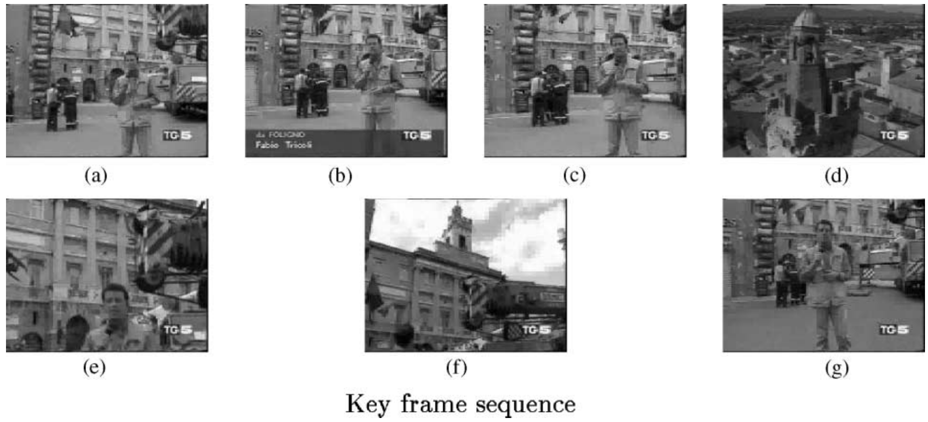


Fig. 3. Example of “false positive” class.

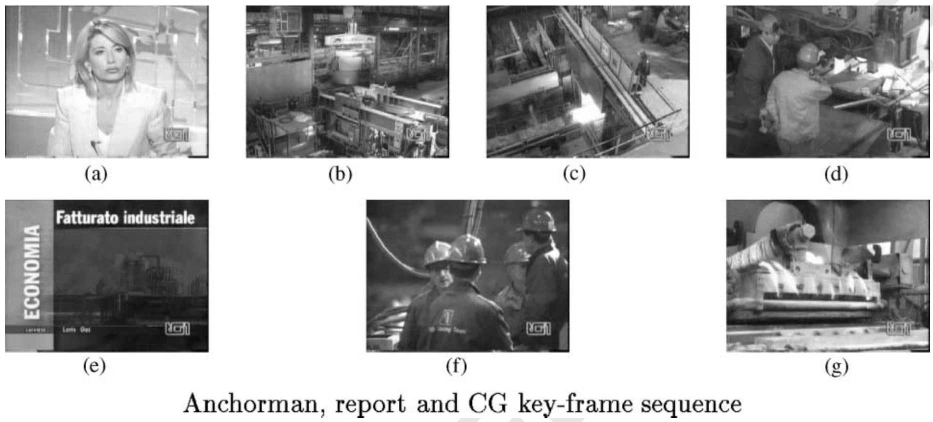


Fig. 4. Example of reusable (report) and not reusable (anchorman and CG) shots.

graphs. They are not reusable due to the fact that the information they convey is subject to fast changes, and can be inexpensively replaced. The shot represented by key-frame (e) in Figs. 4 and 5 shows the sales of February 2000 compared to those of February 1999, and has little reuse value. Unlike it, the shots that show workers in a factory can be reused in other reports. Unlike the anchorman shots it is possible to use neither the structure of the video nor the layout, as a hint to detect the computer graphic shots.

The features used to classify those shots are based on statistical parameters and motion features. The first step calculates an index of the *quantity of motion* $\mathcal{Q}_{CG}$ dividing each shot into sub-shots and taking into account the frame to frame difference between the sub-shot key-frame f_i and $f_{(i+1)}$ according to the previous equation.

To reduce possible misclassification of still images the preliminary classification is refined analyzing the color histogram in the HSI color space. The histogram takes into account only the H and S components, and calculates two indexes: N_{bin} is the number of histogram bins whose value is higher than a τ_{bin} percentage of frame pixels; N_{pix} is the percentage of pixels represented by a selection of the biggest bins of the histogram. N_{bin} and N_{pix} are calculated for each key-frame of the sub-shots and are summed; if one of these values exceed a threshold they are discarded. N_{bin} and N_{pix} take into account the fact that computer graphics shots present a more “compact” color histogram than realistic shots, with a low contrast background that allows higher quality legibility of text and graphics. Table 4 reports the performance of the computer graphics shots classification.

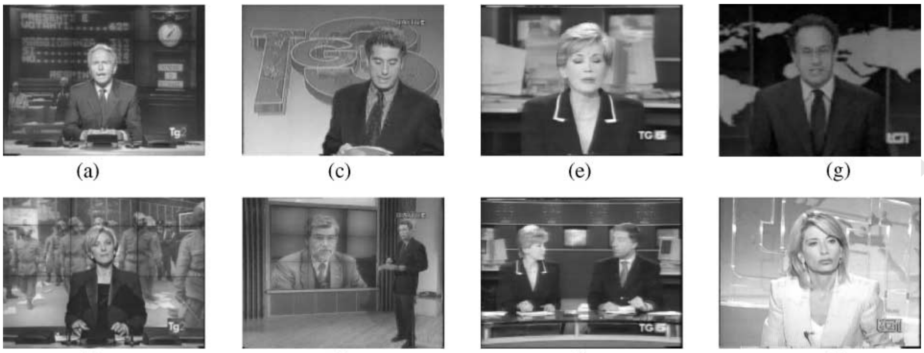
The algorithm takes into account the feasability of the presence of small motion in the CG shot, due to moving text and graphics; An example is shown in Fig. 5.

3.3. Performance evaluation

The shot classification algorithms have been tested on a test database of news video. To verify the robustness of the classification process the database includes news videos of different broadcasters, featuring different styles

(a) (b) (c)

Fig. 5. Example of computer graphics text.



(a) (c) (e) (g)

(b) (d) (f) (h)

Fig. 6. Example of anchorman shots' styles.

and layouts, both for the anchorman shots and for the computer graphic shots.

A short analysis of the most recurrent styles for anchorman shots is reported below:

- The anchorman shots are taken using a fixed position camera while there's an image in the background that shows the subject of the incoming service (Fig. 6(a) and (b)). This style is adopted for all the editions of RAI TG2.
- The anchorman can be either standing or seated, with camera movements and edit effects. The background is usually static (photos or logos). An example is shown in Fig. 6(c) and (d). For example, this style is used in the evening edition of RAI TG3.
- Two anchorman alternate each other. This is shown in Fig. 6(e) and (f). Background is almost fixed or the movement is in small regions. This is used in the evening edition of Mediaset TG5 and in some CNN editions.
- There is a more or less uniform background, some camera movements, limited number of anchorman shots, for example front view and 3/4 view (Fig. 6(g) and (h)).

The results on the test set used in Section 2 are reported in Table 3. The use of the motion feature reduces the number of false detection errors from 14 to 3.

Table 3
Results of the shots classification process

Anchorman shots	Detected anchorman shots	False detections	Missed detections
66	67	3 (4.5%)	2 (3%)

Table 4
Results of the CG shots classification process

Shots	CG shots	False detections	Missed detections
318	15	10	2

Missed detections occurred with type (d) shots when the background contains motion. False detection occurred in the presence of an interview which is similar to types (a) and (b). To improve false detection in (b) analysis was restricted to the central part of the frame, according to the broadcaster's style.

The test set used for the computer graphic classification is a sub-set of the one used for the anchorman classification (see Table 4).

Missed detections of computer graphics are due to fast action, like fast moving text, while the false detection occurred in the presence of still images, or shots that featured very little motion, with low contrast that lead to color histogram distributions similar to that non-realistic shots.

4. Video content representation

To support effective video retrieval by content, high-level information must be extracted from videos and used to perform shot sub-classification based on their content. Additional information to shot classification is extracted from text captions and anchorman speech.

4.1. Text recognition

In news videos, text captions are used to show several information about the shot being broadcasted, such as the site where the action takes place (in service shots) and the names of the people shown in the video (both in anchorman and service shots).

Extraction of text information from video captions has been performed by integrating a traditional OCR within our system. The OCR engine cannot be supplied with raw video frames: a pre-filtering phase is required. This phase includes two distinct steps: caption identification and text/background separation.

Caption identification: If a shot includes a caption, it is not guaranteed that the caption is present in the first frames of the shot. Sometimes the caption appears in the middle of the shot and disappears after the last frame of the shot. Identification of frames including a caption is based on the fact that captions are always used in combination with graphic elements that improve text readability. These graphic elements follow different styles and may include opaque backgrounds and colored lines (Fig. 7). Captions are always located in the lower part of the frame. Caption identification is based on the matching of a pre-defined model of the graphic elements with shapes extracted in the lower part of the frames. The model accounts for the presence of horizontal stripes/long lines either colored or opaque, according to the different broadcasters' styles.

Text/background separation: Text separation is complicated by the presence of captions featuring a poor text/background contrast (Fig. 7(c) and (d)). This sort of problem is dealt with by using a text/background separation method that exploits persistence of patterns over contiguous frames. This method is based on the assumption that for the entire display of a caption all the pixels corresponding to the text have more or less the same value. On the other hand, the value of the pixels in the background changes. Text/background separation is performed by highlighting the pixels the value of which is almost constant. Captions usually display over two or more consecutive shots. A critical instance of text/background separation occurs when a caption without an opaque background appears over a static scene (e.g. a photo or a painting) and is displayed only for the duration of a single shot. This is indeed a rare condition that we did not encounter in our test sequences.

Let us assume that $\{f_0, \ldots, f_k\}$ is a sequence of frames that has been identified as including a caption. A new sequence $\{\hat{f}_0, \ldots, \hat{f}_k\}$ is computed as follows:

$$\hat{f}_0(i,j) = 0,$$
$$\hat{f}_k(i,j) = \begin{cases} \min(255, \hat{f}_{k-1}(i,j) + \Delta) & \text{if } f_k(i,j) = f_{k-1}(i,j), \\ \max(0, \hat{f}_{k-1}(i,j) - \Delta) & \text{if } f_k(i,j) \neq f_{k-1}(i,j), \end{cases}$$

where $f_k(i,j)$ is the gray level value of pixel (i,j) in frame k and Δ a pre-defined incremental step. In this way, the sequence $\{\hat{f}_0, \ldots, \hat{f}_k\}$ is characterized by the text caption that gradually fades in Fig. 8. This method has proven to be robust even in those cases where the sequence of frames includes several captions that are separated by editing effects such as dissolves and cuts.

Finally, the sequence $\{\hat{f}_0, \ldots, \hat{f}_k\}$ is processed in order to extract some frames that are used to feed the OCR engine. For this purpose, the correlation $C(\hat{f}_{k-1}, \hat{f}_k)$ is computed for every pair of contiguous frames. Frames characterized by local maxima of the correlation function are passed to the OCR engine.

The graphic elements like the line and the TG2 logo are removed since they interfere with OCR processing.

The OCR engine: To increase separation between the single characters, and ease their segmentation, on the part of the OCR program, *thresholding* is applied to images extracted from the previous step.

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

Fig. 7. Different styles of captions.

Fig. 8. Character extraction.

Two OCR programs have been tested: (i) TextBridge OCR (Windows commercial OCR package), (ii) SOCR (open source OCR developed by the University of Waikato, New Zealand, http://www.socr.org): the most recent complete version is 0.1 and recognition rate changes according to the fonts employed. Results are shown in Table 5.

4.2. Speech recognition

During the anchorman shot, the content of the following news service is summarized. While in the news service the reporter provides detailed information on the topic.

In order to improve the speech recognition rate and extract only relevant information about news service content, the speech recognition engine is fed with the audio track of the anchorman shots only. In fact, as reported in Ref. [12], generally there is not an exact synchronization between speech and objects shown in the news service. Often the content of the shots does not correspond to the reporter's description, consequently the association of the audio track content with the corresponding shots may lead to erroneous results. Furthermore, the audio track of news services is typically disturbed by background noise and sometimes includes speech transmitted through low-quality telephone or satellite links.

Speech recognition engine: The speech recognition engine used is IBM ViaVoice 98 that features speaker independence and continuous speech processing.

The speech recognition engine is based on a hidden Markov model of the language and uses the following resources: (i) a language thesaurus that can be customized and enhanced, (ii) a customizable model of word usage, (iii) word pronunciation models. A database of audio tracks was used to train the words usage model. The database included the audio tracks of anchorman of several broadcasters. Sentences corresponding to speech were manually transcribed. Their content covered different topics such as sports news, politics, chronicles and gossip.

Table 5
OCR and speech recognition results

Speech recognition			OCR	
Anchor. shots		All audio track	TextBridge	SOCR
Not trained 57%	Trained 84%	Trained 52%	87%	~ 60%

The speech recognition rate was measured on a test database that did not include any of the audio tracks used for training. Results are shown in Table 5.

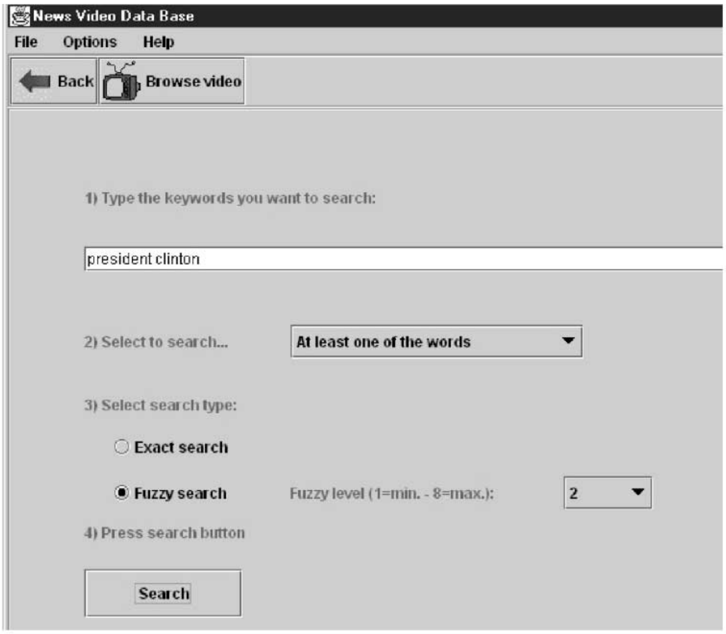
Words extracted by the speech recognition engine were filtered in order to wipe out all utility words (articles, pronouns, conjunctions and prepositions—this accounts approximately for 50% of the speech in latin languages). Remaining words are used to describe the content of the following news services.

5. Video retrieval

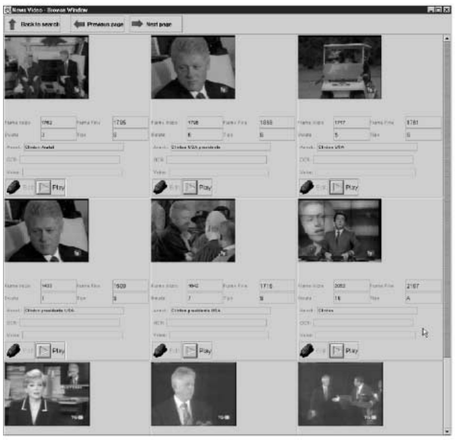
Techniques for video segmentation, shot classification and shot content description presented in the previous sections have been integrated into a system for content-based retrieval of TV news. At archiving time, news videos are automatically processed in order to extract content descriptors for each video shot. The content descriptor of a generic shot includes:

Shot type identifier: This can be either *anchorman*, *news service* or *computer graphics*.
TV broadcaster identifier.
Broadcast date and time.
Visual shot descriptor. This is the key-frame of the shot.
Textual shot descriptor. This is the set of words extracted from shot captions and from speech recognition of the previous anchorman shot. Manual annotation can be added.

At retrieval time, the system supports video querying and browsing. To reduce the effect of the errors of the OCR programs, the retrieval system uses the AGREP approximate text search that allows to find words that



(a)



(b)

(c)

Fig. 9. (a) Specification of a fuzzy search for “President Clinton”. (b) shows result page with shots that match both the keywords, (c) shows the last page with shots that match only the keyword “President”.

contain errors. Queries formulated according to TV broadcaster, date, time, content and any Boolean combination of these are supported. One or more words can be input by the user. These are matched against textual shot descriptors of database videos through the use of a thesaurus so as to support exact word and synonym matching. Matched shots are presented to the user for browsing. For each matched shot all the information stored in its content descriptor is shown.

In Fig. 9(a) a sample query by content is shown. The user enters a Boolean combination of the words ‘President’ or ‘Clinton’ to search for shots with similar content. Retrieval results are shown in Fig. 9(b). The query also retrieves shots classified with the Italian word “Presidente” (speech transcription and manual annotation), since the “fuzzy” search method is used. Retrieved shots are shown in decreasing order of match. The first shots match both query keywords and show news and anchorman shots related to “President Clinton”. The other shots retrieved match only the keyword “President” and show “President Milosevic” and “President Scalfaro”. The “Previous” and “Next” buttons on the top of the window allow the user to navigate through all the retrieved shots.

Selection of a shot key-frame from the output interface allows display of the entire shot through a movie player application.

6. Conclusions

This paper presents a system for content-based indexing and retrieval of news videos. The system features content-based shot classification of anchorman and *non-realistic* shots, to allow the reuse of report shots. Extraction of high-level content descriptors through caption OCR and speech recognition. Shot classification is based on statistical and motion features of the news video structure, so as to provide independence from TV broadcaster style.

References

[1] A. Del Bimbo, C. Colombo, P. Pala, Semantics in visual information retrival, IEEE Multimedia 6 (3) (1999).
[2] D. Swanberg, C. Shu, R. Jain, Knowledge guided parsing in video databases, Spie 1908 (13) (1993) 13–24.
[3] S. Smoliar, H.J. Zhang, Y. Gong, Automatic parsing of news video, Proceedings of the IEEE Conference on Multimedia Computing and Systems, May 1994, pp. 45–54.
[4] T. Kanade, Y. Nakamura, Semantic analysis for video contents extraction spotting by association in news video, ACM Multimedia 97 (1997).
[5] H. Zhang, B. Furht, S.W. Smoliar, Video and Image Processing in Multimedia Systems, Kluwer Academic Publishers, Dordrecht, 1995.
[6] A.G. Hauptmann, M.J. Witbrock, Informedia: news-on-demand multimedia information acquisition and retrieval, Intell. Multimedia Informat. Retrieval (1997) 213–239.
[7] T. Sato, T. Kanade, E.K. Hughes, M.A. Smith, Video ocr for digital news archieve, IEEE International Workshop on Content-Based Access of Image and Video Databases CAIVD’ 98, 1998, pp. 52–60.
[8] S. Eickeler, S. Muller, Content-based video indexing of tv broadcast news using hidden Markov models, Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), March 1999, pp. 2997–3000.
[9] R. Lienhart, Indexing and retrieval of digital video sequences based on automatic text recognition, Fourth ACM International Multimedia Conference. 1996.
[10] A.G. Hauptmann, Speech recognition in the informedia digital video library: uses and limitations, ICTAI 95, 1995.
[11] A.G. Hauptmann, H.D. Wactlar, M.J. Witbrock, Informedia: News-on-demand experiments in speech recognition, ARPA Speech Recognition Workshop, 1996.
[12] M.J. Witbrock, A.G. Hauptmann, Speech recognition for a digital video library, JASIS, 1996.
[13] D. Luparello, B. Merialdo, K.T. Lee, J. Roudaire, Automatic construction of personalized tv news programs, Proceedings of ACM Multimedia 99, 1999, pp. 323–330.
[14] B. Merialdo, Automatic indexing of tv news, WIAMIS ’97, June 1997.
[15] J.S. Boreczky, L.A. Rowe, Comparison of video shot boundary detection techniques, Technical Report, Computer Science Division-EECS, University of Calfornia Berkeley.
[16] M. Caliani, C. Colombo, A. Del Bimbo, Commercial video retrieval by induced semantics, IEEE International Workshop on Content-Based Access of Image and Video Databases CAIVD ’98, 1998, pp. 72–80.
[17] U. Gargi, R. Kasturi, An evaluation of color histogram based methods in video indexing, First International Workshop on Image Database and Multi-media Search, 1996, pp. 75–82.

About the Author—MARCO BERTINI has a research grant and carries out his research activity at the Department of Systems and Informatics at the University of Florence, Italy. He received a MS in electronic engineering from the University of Florence in 1999. His main research interest is content-based indexing and retrieval of videos.

About the Author—ALBERTO DEL Bimbo graduated in 1978 from the University of Florence, Italy, where he is presently Full Professor of Computer Engineering. Presently he is the Director of the Master in Multimedia at the University of Florence and the Deputy Rector for Research and Innovation Transfer of the University of Florence.
His scientific interests and activities have addressed the subjects of Image Technology and Multimedia, with particular reference to object recognition and image sequence analysis, content-based retrieval for image and video databases and advanced man-machine interaction. Prof. Del Bimbo is the author of over 150 publications, that have appeared in the most distinguished international journals and conference proceedings and is the author of the monography “Visual Information Retrieval” edited by Morgan Kaufman in 1999. From 1996 to 2000, he has been the President (formerly Vice-President) of the Italian Chapter of IAPR, the International Association for Pattern Recognition. He obtained the IAPR fellowship in 2000. He presently serves as Associate Editor of IEEE Transactions on Multimedia, IEEE Transaction on Pattern Analysis and Machine Intelligence, Pattern Recognition, Pattern Analysis and Applications Journal, Journal of Visual Languages and Computing and Multimedia Tools and Applications Journal. Since 1999 he has been Member at Large of the IEEE Publications Board.

About the Author—PIETRO PALA is an assistant professor in the Department of Systems and Informatics at the University of Florecence, Italy. He received his MS in electronic engineering at the University of Florence in 1994. He received a Ph.D. in information science from the same university in 1998. His current research interests include pattern recognition, image and video retrieval by content, and related applications.