Supervised And Unsupervised Classification Post-Processing for Visual Video Summaries

Gianluigi Ciocca and Raimondo Schettini

Abstract — *Representation of the video content using a set of key frames is one of the most common techniques for video summarization. Summaries composed of key frames allow users to grasp the overall content of a video, and access specific sequences. The post-processing algorithm presented in this paper makes it possible to create visual video summaries that are exhaustive, but not redundant. In three steps the method removes meaningless key frames, groups the key frames into clusters to allow multi-level summary presentation, and determines the default summary level. The algorithm utilizes both supervised and unsupervised classification strategies to perform these tasks. It requires no previous knowledge about the video contents, nor is any assumption made about the input data .*

Index Terms — Visual video summaries post-processing, supervised and unsupervised classification, multilevel visual summaries.

I. INTRODUCTION

The growing interest of consumers in the acquisition of and access to visual information has created a demand for new technologies to represent, model, index, and retrieve multimedia data [1]. Very large databases of images and videos require efficient algorithms that enable fast browsing and access to the information pursued [2]. In the case of videos, in particular, much of the visual data offered is simply redundant, and we must find a way to retain only the information strictly needed for functional access, browsing and querying. Representation of video contents using a set of frames (key frames) is one of the most common techniques for video summarization. Summaries composed of key frames can resume the video contents in a rapid and compact way: users can grasp the overall contents more quickly from key frames than by watching a set of video sequences. Besides providing video browsing capability and content description, key frames act as video "bookmarks" that designate the interesting events captured, supplying direct access to video subsequences.

Videos such as motion pictures, are strongly structured and organized: shots, scenes and acts are tied together by a story line that defines the relations between the different components of the video. News programs, TV shows, and documentaries also have a distinctive structure of shots and scenes. In this kind of video, the extraction of the key frames can be driven by the detection of the video shots (see [3], [4], and [5] among others). Even in videos such as home videos, that apparently lack a clearly defined structure, a certain framework can be found, and it can be exploited to detect meaningful video segments and extract significant frames to use in video summarization [6]. For example, the visual summary shown in Fig. 1 has been created using the approach described in [5]. The video was processed by a shot detection algorithm to identify the boundaries of each shot. The visual contents of the frames were analyzed for different clues will indicating the presence of editing effects such as cuts, fades and dissolves. After each shot was detected, the frames composing it were further analyzed, and the key frames automatically extracted at salient points where the pictorial elements of the frames changed in a significant way, allowing the selection, for each shot, of a variable number of key frames.

From the user's point of view, and regardless of the key frame extraction strategy adopted, not all the images extracted are important or necessary to convey the visual contents of the video. Several different factors may render the visual summaries less useful or less attractive. For example, they may contain uninteresting or meaningless key frames (in Fig. 1, those corresponding to the color bars), overexposed or underexposed key frames, or close-ups with very few details. We consider all these as meaningless for key frames since they provide so little information. However, to our knowledge no previous attempts have been made to systematically remove from visual summaries frames that are meaningless in terms of the properties displayed with the exception of almost unicolor images (e.g. [7]).

Video summaries may also present a succession of very similar key frames. This problem may originate from different causes. For example, similar key frames may be repeatedly extracted if, while shooting the scene, an object moves very close in front of the camera, shadowing the whole scene. This may be interpreted as an abrupt change in the video sequence, and the shot detection algorithm may identify boundaries where they do not exist. Another cause is the presence of interruptions within a sequence. This is the case, for example, when a single, long sequence (e.g. an interview), is edited to reduce its duration. Although the sequence is semantically contiguous, the interruptions may be considered shot boundaries by the shot detection algorithm and the key frame extraction algorithm itself may cause the extraction of an over-large number of key frames due to the imprecise evaluation of the variations in the frame sequences.

G. Ciocca is with the Department of IT, Systems and Communications (DISCo) at the University of Milano-Bicocca, Milan, Italy (email: ciocca@disco.unimib.it).

R. Schettini is with the Department of IT, Systems and Communications (DISCo) at the University of Milano-Bicocca, Milan, Italy (email: schettini@disco.unimib.it).

Manuscript received March 13, 2006

Fig. 1. Example of visual summary. The original video ("DVD0019a") is a MPEG-4 sequence 30 minutes and 30 seconds long, composed of 45,753 frames (384x288 pixels at 25fps). The summary is composed of 55 key frames and was created with the algorithm described in [5]. The video is part of the AESS video archive (see below) and is an excerpt from a documentary on the "Summer White Carnival" at Cegni, showing a feast in progress, and a participant preparing for the traditional Carnival masquerade.

Examples of these situations can be seen at frame 25773, frame 28737, and the sequence of close-ups starting at frame 35890 of Fig. 1.

The management of potentially hundreds of key frames is another aspect that must be taken into account. Managing visual summaries with hundreds of key frames is a rather difficult and inefficient task for the human user, who is more interested in the underlying plots of a video sequence then in knowing all the details. Consequently the number of key frames must be limited while preserving the overall contents and structure to allow users a practical way of managing visual summaries. However, since different users may be interested in different levels of detail, it is also necessary to provide a hierarchical structure containing different views of video contents, from a coarse view showing the overall plot to finer views containing an increasingly number of details. Users can then quickly browse through a video sequence, rapidly get an overview of the contents, and navigate to different levels of detail to locate the segments of interest. Although there are already only 55 key frames in the summary in Fig. 1, under certain viewing conditions (for example on devices with small monitors) it might be useful to further reduce the number of key frames.

The problem of redundant key frames and that of multilevel summarization can be dealt with by exploiting clustering algorithms to group similar key frames. These algorithms are widely exploited in the field of video analysis, particularly for video segmentation and summarization. In [8], an agglomerative clustering algorithm is used to group similar shots in clusters based on temporal constraints and cophenetic dissimilarity criteria. The average signature computed on the corresponding key frames describes each shot. In [9], sport videos are segmented into shots and summarized by selecting five key frames at equal intervals within each shot. The scenes in the sequence are identified by merging the shots on the basis of a scene-likeness measure computed on the HSV color histogram produced for the shot's key frames. Fisher's discriminant analysis is used to determine the correct number of clusters. A similar approach is found in [10] where the shots extracted from sport videos are clustered into dominant scene. A scene is defined dominant is appears repeatedly over an extended portion of the video footage. The differences between a set of shots and a prototypical shot are used to group similar shots together, assigning those that are similar to the a group, and the others to a second group. All the shots in the first group are held to belong to the dominant scene in the video sequence. By selecting another prototypical shots within the second group, shots belonging to the second most dominant scene can be identified, and so on. In [11] the authors are interested in determining the number of distinct shots within a video sequence. This knowledge is then used to set the threshold for the segmentation of the sequence. The frames, described in terms of texture features are grouped by the c-means clustering algorithm. A temporal validity index is defined so that only consecutive frames are merged. To determine the optimal number of clusters (and thus the number of shots), the authors rely on the Davies Bouldin index. In [12] the shot clustering proposed is hierarchical. A compatibility measure based on a fuzzy membership function is used to assess the degree of dissimilarity between shots belonging to a predefined temporal window. The compatibility measure represents the correlation among shots in the video sequence. Time constraints are used to split clusters that exceed a duration threshold. The clustering process terminates when no changes can be made in the clusters found. In [13] a methodology for discovering the cluster structure in home videos is presented. Statistical models of visual similarity, duration and temporal adjacency are exploited. The hierarchical clustering process is formulated as a sequential binary Bayesian classification process. The features used in the process comprise color histograms, color ratios, edge density, edge directions and

temporal adjacency. The authors of [14], present a multilevel hierarchical video summarization strategy. The pictorial key frames are used to visualize the video contents ,and a hierarchical strategy is introduced to construct multilevel summaries. Key frames (level 1 summary) are grouped into video super groups (level 2 summary) on the basis of their visual similarities. Temporal adjacency is used to create video groups, and a temporal interlaced video correlation is then used to identify the video scenes, which are finally clustered together to eliminate visual redundancies. These operations create higher (level 2+) summaries. In [15], shots are clustered on the basis of their visual similarity and temporal closeness. The hierarchical clustering process continues until all the distances between the clusters are greater then a given threshold. A scene transition graph (STG) is derived from the clustering phase to partition the video into story units.

II. PROPOSED POST-PROCESSING ALGORITHM

Fig. 2 shows the three steps of the post-processing pipeline. The first removes meaningless key frames, using supervised classification performed by a neural network on the basis of pictorial features derived directly from the frames, together with others derived from the processing of the frames by a visual attention model algorithm. The second step provides for the grouping of the key frames into clusters to allow multilevel summary using both low level and high level features. The third step identifies the default summary level that is shown to the users: starting from this set of key frames, the users can then browse the video content at different level of detail.

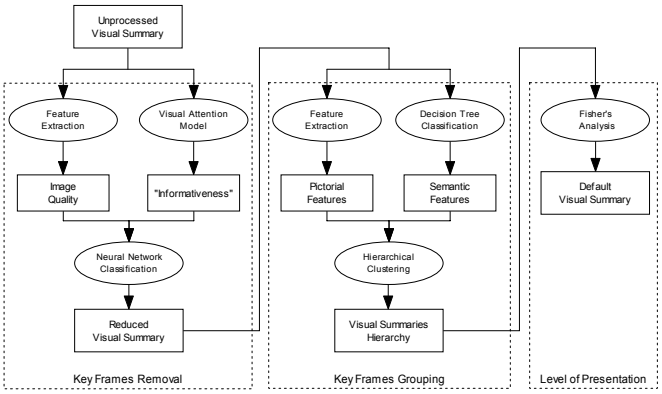

Fig. 2. Pipeline of the proposed post-processing algorithm.

A. Removal of Meaningless Key Frames

To remove meaningless key frames, we to extract a set of features that describe them in terms of quality and information supplied. To assess image quality, we use some of the features considered in judging the quality of images taken by digital cameras: the percentage of dark and bright pixels, to identifying overexposed and underexposed images; the dynamic range of the image, to single out flat-looking images; and the color balance [16]. To assess the amount of information the key frames convey (that is, their "informativeness"), we use a visual attention model proposed by Corchs et. al [17] to locate the Regions of Interest (ROIs) on a saliency map. This visual attention model can detect those portions of the input image where highly informative contents are located, and suppress the remaining parts. It produces a map of activities of the original image: high values in the resulting map indicate areas of high neural activity, i.e. areas where we expect visual attention to be focused, and thus significant information to be located. Fig. 3 shows the processing steps followed to obtain the ROIs of a given image.

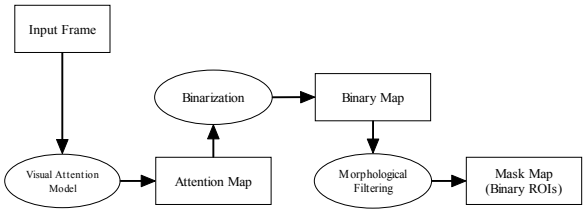

Fig. 3. ROIs extraction pipeline.

The first step is the computation of the neuronal activity map by the visual attention models. The second is the binarization of the attention map in order to retain only those regions of greater neural activity. The values of the attention map are analyzed, and the statistics regarding their distribution are used to determine a threshold. Values above the computed threshold are considered to belong to the region of interest. The regions not selected are assumed to concern non-relevant and noisy data. We choose to use as threshold the average of the activity values. Only those pixels for which the corresponding activity is above the average are retained. The result of the binarization process is a preliminary binary mask representing the possible location of the ROIs. Since the binary mask may be composed of isolated pixels or fragmented regions, the third step in the processing pipeline is binary morphological filtering, applied to remove noisy pixels, and obtain regions with smooth borders and uniform areas. The morphological operators used are the opening operator followed by the closing operator. The opening operator allows us to remove the smaller regions in the image, while the closing operator allows us to close most of the gaps within the regions. Different levels of filtering can be achieved by varying the size and shape of the structuring element. We have employed a square structuring element 3x3 pixels in size. The result of the third step is a mask image containing the locations of all the filtered ROIs. Fig. 4 shows some examples of maps of activities and the ROIs extracted from uninformative images, while Fig. 5 gives some examples of maps of activities and ROIs extracted from informative images.

Intuitively, images with a large volume of ROIs are more informative then images with a small volume of ROIs. The location of these ROIs is also helpful in discriminating between informative and uninformative images. If the ROIs are located near the center of the image, the object of attention is clearly in the foreground. If, instead, the ROIs are located near the borders of the image, the object of interest is in the background, and not clearly identifiable (see Fig. 6).

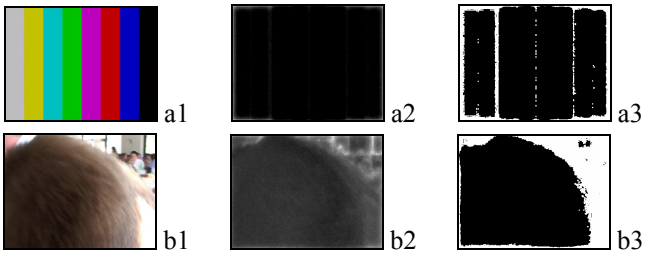

Fig. 4. Examples of uninformative frames (left) with their corresponding maps of activities (center), and the regions-of-interest extracted (right).

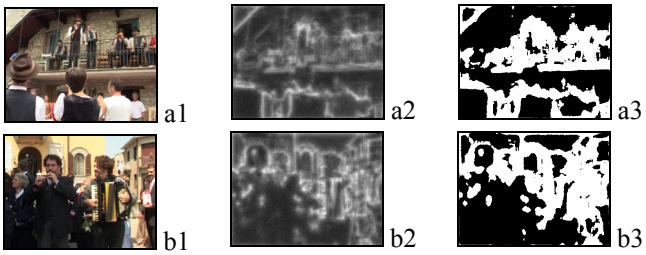

Fig. 5. Examples of informative frames (left) with their corresponding maps of activities (centre), and regions-of-interest extracted (right).

Consequently two features extracted as criteria of "Informativeness" are the percentage of pixels belonging to ROIs within a central region of the image, and their dispersion with respect to the center of the image. The central region has been set at about 65% of the whole image.

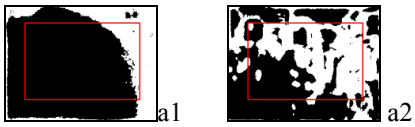

Fig. 6. The central region of the images is used to compute the "Informativeness" feature.

In deciding whether or not to reject a key frame, candidate frames are classified in two groups: key frames to reject, and key frames to retain. Due to the small number of features involved (eight in all here), the classification of the key frames is performed using a neural network classifier [18].

For our purposes, the neural network was composed of one input node for each feature value (i.e. eight input nodes) plus a node with a bias term; a number of hidden layers, each with the same number of nodes as the input (we experimented with one, two, and three hidden layers), and a single final node which gave an output value in the range of [0,..,1]. The output could be interpreted as the probability that the processed key frame should be discarded.

To avoid data overfitting, the ground truth set (composed by 440 elements) was partitioned into a training set used for learning, a validation set used to decide the termination of learning, and a test set used to evaluate the classification performance. We created the ground truth by asking several users to select those key frames they would like to remove on the basis of the quality of the frames, and the information conveyed. The final class of each key frame was then assigned by a majority vote. Experimentally, the best results were obtained with a neural network containing two hidden layers. Table I shows the classification results of this network on the training set and on the evaluation set. The results are in percentages, and refer to the neural network with two hidden layers.

TABLE I
CONFUSION MATRICES OF THE FRAMES CLASSIFICATION

	Training Set		Evaluation Set	
	Class 0	Class 1	Class 0	Class 1
Class 0	98.00	02.00	94.98	06.02
Class 1	37.50	62.50	26.00	74.00

The moderately low precision in assigning key frames to Class 1 is mainly due to the difficult and often very subjective characterization of the class of the rejected key frames. Another cause of misclassification of rejected key frames is the relatively small amount of ground truth data available. The moderately low precision in assigning key frames to Class 1 is mainly due to the difficult and often very subjective characterization of the class of the rejected key frames. Another cause of misclassification of rejected key frames is the relatively small amount of ground truth data available.

Fig. 7. The summary of Fig. 1 after key frame classification and the removal stage. Nine key frames have been removed, resulting in a summary composed of 46 key frames.

In any case, a large misclassification error in the rejected key frames class is obviously preferable to a large a high misclassification error in the class of retained key frames since dropping informative key frames may compromise the visual summary.

Fig. 7 shows the summary of Fig. 1 after the key frames removal stage performed using the trained neural network with two hidden layers. As it can be seen, nine key frames have been removed: among the others, the two color bars (frames 217, and 45262), some strong close ups (frames 35890, 36002, 36069, and 38962), and a frame mainly showing the shadow of a person leaving the room (31719).

Because each key frame is used as an access point to the video contents and is as such associated with a video subsequence, to ensure that all the video contents remain accessible after removal of the key frames, the subsequences targeted by the key frames that have been removed are assigned to the nearest key frames left in the summary. When there is a choice, priority is given to the key frame belonging to the same shot.

B. Key Frames Grouping

The key frame grouping algorithm that we propose is conceptually different from the scene clustering task where key similar key frames that are spread along the summary are deemed to belong to the same scene, and grouped together. We are mainly interested in merging similar key frames while preserving the temporal ordering of the remaining key frames regardless of the shot boundaries. Key frames belonging to the same shot may be merged in different clusters if they are sufficiently dissimilar in appearance and content. Instead of using only simple pictorial features to describe the key frame content (as most of the scene clustering algorithms proposed in the literature do), we use a two level description approach. The key frames are described by low level (pictorial) and high level (semantic) features.

The low level features used are the color histogram, the wavelet statistics and the edge's direction histogram for a total of 156 values. The color histogram used is composed of 64 bins determined by sampling groups of meaningful colors in the HSV color space. The edge direction histogram is composed of 72 bins corresponding to intervals of 2.5 degrees. Two Sobel filters are applied to obtain the gradient of the horizontal and the vertical edges of the luminance frame image. These values are used to compute the gradient of each pixel: those pixels that exhibit a gradient over a predefined threshold are taken to compute the gradient angle, and then the histogram. Multiresolution wavelet analysis can provide information about the overall texture of the image at different levels of detail. At each step in multiresolution wavelet analysis four sub-images (or sub-bands) are obtained with the application of a low-pass filter (L) and high-pass filter (H) in the four possible combinations of LL, LH, HL and HH. We apply multiresolution wavelet analysis on the luminance frame image, using three-step Daubechies multiresolution wavelet expansion to produce ten sub-bands. Two energy features, the mean and the variance, are computed on each sub-band, resulting in a 20-value descriptor. These complementary visual features (colors, textures and edges) capture the contents in an efficient and compact way, and have already been successfully used in the extraction of key frames [5].

Given this set of pictorial features, the pictorial signature (*Ps*) of a key frame *KF* can be defined as:

$$Ps(KF) = < H, W, D > \tag{1}$$

where H represents the color histogram, W the wavelet statistics and D the edge direction histogram. To compare two pictorial signatures the following measure is used:

$$\begin{aligned} Diff_{HWD}(Ps_1, Ps_2) = &\left[d_H(H_1, H_2) \cdot d_W(W_1, W_2)\right] + \\ &\left[d_W(W_1, W_2) \cdot d_D(D_1, D_2)\right] + \left[d_D(D_1, D_2) \cdot d_H(H_1, H_2)\right] \end{aligned} \tag{2}$$

where d_H, is the histogram intersection distance [19]:

$$d_H(H_1, H_2) = 1 - \sum_{j=0}^{63} \min(H_1(j), H_2(j)) \tag{3}$$

and d_D and d_W are the Euclidean distance:

$$\begin{aligned} d_D(D_1, D_2) &= \sqrt{\sum_{j=0}^{71} (D_1(j) - D_2(j))^2} \\ d_W(W_1, W_2) &= \sqrt{\sum_{j=0}^{19} (W_1(j) - W_2(j))^2} \end{aligned} \tag{4}$$

The three distances are mapped into the range of [0,1] before being combined to form the final difference measure d_{HWD} (to simplify the notation we have indicated them as d_H, d_W, and d_D only) :

$$d_{HWD} = (d_H \cdot d_W) + (d_W \cdot d_D) + (d_D \cdot d_H) \tag{5}$$

The high level features are obtained by applying the classification strategies described in [20], where the images are classified as indoor, outdoor, or close-up images. The classification is based on the use of ensembles of decision trees, often called decision forests. The trees of the forests are constructed according to CART methodology [21]. The features used are related to color (moments of inertia of the color channels in the HSV color space, and skin color distribution), texture and edge (statistics on wavelets decomposition and on edge and texture distributions), and composition of the image (in terms of fragmentation and symmetry). To fully exploit the fact that trees allow a powerful use of high dimensionality and conditional information, all the features are taken together, letting the training process perform complexity reduction, and redundancy detection. An ambiguity rejection option is also included, for a more accurate classification. The final decision

to classify an image in one of the three classes, or reject it, is made by a majority vote among the decision results of the trees in the forest. Examples of key frame classification results are presented in Fig. 8. Fig. 9 shows examples of misclassified key frames. For a more detailed description of the classification process, see [20]. Instead of assigning an image to a single class we have used the probabilities that a given image belongs to each class as a semantic histogram signature. The semantic signature (*Ss*) of a key frame *KF* is then defined as:

$$Ss(KF) = < I, O, C > \tag{6}$$

where *I*, *O* and *C* are the percentage of the frame belonging to the Indoor, Outdoor and Close-up classes respectively. Two semantic signatures are compared using the Euclidean distance:

$$Diff_{IOC}(Ss_1, Ss_2) = \sqrt{(I_1 - I_2)^2 + (O_1 - O_2)^2 + (C_1 - C_2)^2} \tag{7}$$

The overall difference (*Diff*) between two key frames is computed by a linear combination of the pictorial signature difference and the semantic signature difference (to simplify the notation, we omit the arguments of the differences):

$$Diff(KF_1, KF_2) = \alpha(Diff_{HWD}) + (1-\alpha) Diff_{IOC} \tag{8}$$

The factor α is used to weight the contribution of one signature over the other. In our experiments the weight was set at 0.5, equally weighting the two signatures. The use of features at different levels of abstraction allows us to reduce the error made in comparing two key frames. If both the pictorial content and the semantic content are similar the two key frames probably belong to the same shooting sequence; they are consequently merged together.

To group similar key frames a hierarchical clustering based on the complete link strategy has been adopted adding few constraints. We let $\{KF_1, \ldots, KF_N\}$ be the *N* key frames that compose the original visual summary, G_i the i-th cluster, and g_i its representative element. Initially each key frame KF_i belongs to a cluster G_i and is the representative element (key frame) of that cluster:

$$G_i = \{KF_i\} \text{ and } g_i = KF_i \tag{9}$$

Representative key frames are used to visually represent the cluster in the hierarchical visual summary. The clusters are arranged according to the chronology of the representative key frames. The hierarchical clustering is performed by computing the distance between clusters within a temporal window *t* and using the complete link strategy.

$$Dist(G_i, G_{i-t}) = \max_{KF_n \in G_i, KF_m \in G_{i-t}} \{Diff(KF_n, KF_m)\} \tag{10}$$

Diff( ) is the key frame difference measure as defined above and is computed on the representative key frames. The temporal window *t* has been set at was set at 1, meaning that only pairs of consecutive clusters are evaluated. This allows the merging of successive similar clusters (key frames) first. If the temporal window is set at an arbitrary value *T*, the clustering is equivalent to that of scene clustering algorithms with a scene duration model of *T* lengths. Once the cluster distances have been calculated, the two clusters that are closest are merged to form a new cluster. To decide which of the two should be deleted from the list of clusters, we apply a rule based on the "Informativeness" features described above and, specifically, on the percentage of ROIs: the cluster having the representative key frame with the highest percentage of ROIs is retained, while the other is deleted and all its elements added to the winner cluster. The rule ensures that of two representative key frames only the one with the highest informative content is retained at each iteration. The representative key frame of the winning cluster is not changed. After each clustering iteration, the set of representative key frames is reduced by only one element. Since the set of representative key frames forms a summary, with *N* key frames, the clustering algorithm produces a hierarchy of *N-1* summaries (levels). The original summary corresponds to the 0 level summary. Browsing from the higher levels to the lower levels, it is possible to view the summary from a rough summary (few key frames) to a finer summary (more key frames).

Fig. 8. Examples of correctly classified key frames. The numbers below each image refer to the probability that the image belongs to the indoor (I), outdoor (O) or close up (C) classes.

Fig. 9. Examples of misclassified key frames. The numbers below each image refer to the probability that the image belongs to the indoor (I), outdoor (O) or close up (C) classes.

C. Default Summary Level

From the visual summary we have derived a hierarchical structure that represents the visual summary at different levels of detail (i.e. with different numbers of clusters/key frames). This allows the user to inspect the contents of the video by navigating throughout the levels. The problem of the multilevel summary is to decide which summary to present to the user as the optimal or default summary level. When the number of levels is low (i.e. the number of key frames composing the original visual summary is low), the user can easily browse all the levels until he eventually finds the one with the information he is searching for. When the original summary is composed of hundreds of key frames, browsing through all the levels is cumbersome and time consuming. It is necessary to define a strategy to select, if it exists, the summary that is least redundant in terms of pictorial information (i.e. few similar key frames). It also affords an optimal starting point for browsing the remaining summary levels.

The idea underlying the strategy for detecting the default summary level is that the frame differences used in merging the clusters in the previous section be considered merging costs, that is, the costs of reducing the summary by one key frame. In the initial phase of the clustering process, the costs will be relatively small, that is, the merged clusters will have small merging costs because the representative key frames are similar, and their merging will not significantly reduce the summary information contents. As clustering continues, the merging costs will increase, meaning that the representative key frames will be increasingly dissimilar, and their merging will result in a summary with fewer information contents. We select the level at which the merging costs rise significantly as the level of the default summary. Starting from there and moving toward lower levels we obtain summaries with more key frames and redundant information. Moving toward higher levels, we have summaries with few key frames and more compact contents. To select the clustering level we have used an approach based on the peer-group filtering (PGF) scheme proposed by Deng et. al [22][23]. The authors employed it to filter image noise and to quantize color space for the purpose of image enhancement. In [10] the PGF scheme has been employed to detect dominant scenes in sports videos. The objective of the PGF is to group a set of data into two classes by minimizing intra-group differences while maximizing their inter-group separation, using Fisher's discriminant analysis [24].

We apply the PGF scheme to the sequence of merging costs determined in the previous clustering process. We let $\{c_0, c_1, \ldots, c_{N-1}\}$ be the costs associated to the summary levels, that is, c_0 is the cost of obtaining the summary at level 0 (assumed to be equal to zero), c_1 the cost of obtaining the summary at level 1, and so on. Since the costs refer to the merging of pairs of frames with a minimum difference, they are naturally ranked in ascending order. We define the set of partitions (P_i) of the cost values into two groups as:

$$P_i = \{\{c_0,\ldots,c_i\},\{c_{i+1},\ldots,c_{N-1}\}\}, \quad i=0,..,N\text{-}2 \tag{11}$$

For each partition P_i the Fisher's index is computed as:

$$F_i = \frac{\left|\mu_i^1 - \mu_i^2\right|^2}{\varpi_i^1 + \varpi_i^2} \quad i=0,\ldots,N\text{-}2 \tag{12}$$

where

$$\mu_i^1 = \frac{1}{i+1}\sum_{j=0}^{i} c_j \qquad \mu_i^2 = \frac{1}{N-i-2}\sum_{j=i+1}^{N-2} c_j \tag{13}$$

and

$$\varpi_i^1 = \left(\sum_{j=0}^{i}\left(c_j - \mu_i^1\right)^2\right)^2 \qquad \varpi_i^2 = \left(\sum_{j=i+1}^{N-2}\left(c_j - \mu_i^2\right)^2\right)^2 \tag{14}$$

The default summary level is determined by the maximum of Fisher's index values:

$$D = arg\max_i \{F_i\} \tag{15}$$

The level corresponding to the maximum value D, indicates the partition (in terms of Fisher's analysis) where the costs assigned to each group are homogeneous, and the costs of the second group are visibly higher than those of the first. Reducing the summary by one key frame at a time from this point on, means merging pairs of dissimilar key frames (greater differences), with a consequent loss of information.

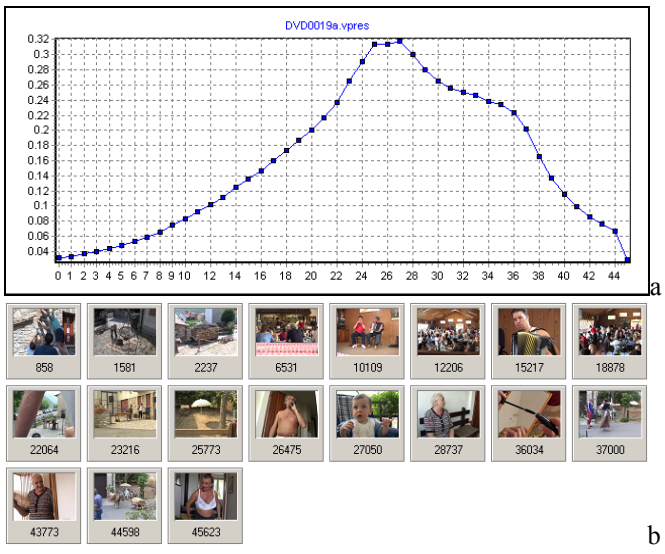

Fig. 10. Result of the selection of the default summary level for the summary in Fig. 14. a) Fisher's analysis of the merging costs. b) the level 27 summary.

The default clustering level for the summary in Fig. 7 is shown in Fig. 10. Image a, is the plot of the Fisher's indexes with a maximum in the 27th position. Image b shows the summary corresponding to the 27th clustering level. If the maximum value is very close to the first or last clustering level, one of the two groups will be composed of very few elements and the Fisher analysis has not been able to identify a clear separation between the costs. We may conclude that, for that summary, the default level does not differ substantially from that of the original summary at level 0 which can then be used as the default level.

D. Computational Complexity

In order to estimate the computational complexity of the proposed pipeline, we have considered the number of operations required to perform each processing step. The evaluation is based on the actual implementation and does not take into account any specific algorithm optimization. Table II shows the computational complexity of the post-processing steps. We have indicated with P the number of pixels in the frames; N the number of bins in a generic histogram; K the number of key frames in the summary, and G the number of Gabor filters used to compute the ROIs. As expected, most of the computation load can be found in the first two processing steps where image features must be computed. In particular, the key frame removal step requires the computation of the regions of interest and this is the most time-consuming elaboration within the whole processing pipeline.

TABLE II
COMPUTATIONAL COMPLEXITY OF THE THREE PROCESSING STEPS

Processing Step	Complexity
Key Frame Removal	$O(20G^2P)$
Key Frame Grouping	$O(117P + 30N^2)$
Default Summary	$O(4K^2)$

The neurodynamical visual attention model we used is based on evidence of neurophysiological and psychological findings. To reduce the computational load, other algorithms such as in [25] that compute the regions of interest in a different manner can be used instead. It should be pointed out that the post-processing pipeline can be applied on frames with lower resolution. In our implementation, the frames are sampled (maintaining the aspect ratio) at 128 pixels on the higher dimension.

III. RESULTS AND FINAL REMARKS

Since all the above processing steps (removal of meaningless key frames and classification) rely on semantic information which no objective quality measure can effectively incorporate, the post-processing algorithm was heuristically tested by domain experts on a set of videos belonging to the "Archivio di Etnografia e Storia Sociale - AESS" [26]. These experts manage video footages on a daily basis, manually extracting relevant information the videos to use for content cataloging and publication through distributed channels. The test set was composed of 14 non-professional videos, about 4 hours of footage. The experts evaluated the processed summary in terms of compactness, and information contents as well as the effectiveness of the multilevel summary. They judged the results positive, and our key frames extraction algorithm [5] and post-processing pipeline will be incorporated into the informative system of the AESS archive. The results of some of the AESS videos processed with the proposed algorithm can be found at http://www.ivl.disco.unimib.it/Activities.html. As the post-processing algorithm does not use previous knowledge about the video contents, nor is any assumption made about the input data, it can be used in different domains as a general purpose algorithm. Nevertheless, some improvements can be made. The key frame removal stage could be extended with more pictorial quality features (both low level and high level) in order to better cover the many factors that can cause a user to reject a frame (e.g. wrong skin tone, half faces, etc…). The key frame grouping stage could also be extended. We have introduced three generic classes for the classification of the key frames but more classes can be added in the decision trees to enlarge the semantic dictionary. The generic approaches used in the whole post processing pipeline, mean that it can easily be specialized to support domain specific applications by taking into account the appropriate pictorial and semantic properties.

REFERENCES

[1] N. Dimitrova, H. Zhang, B. Shahraray, M. Sezan, T. Huang and A. Zakhor, "Applications of video content analysis and retrieval", *IEEE MultiMedia* 9(3), 44-55, 2002.

[2] P. Agraim, H. Zhang, D. Petkovic, "Content-Based Representation and Retrieval of Visual Media: A State of the Art review", Multimedia Tools and Applications, vol. 3, pp. 179-202, 1996.

[3] A. Hanjalic, R.L. Lagendijk, J. Biemond, "A new Method for Key Frame Based Video Content Representation", *Image Databases and Multimedia Search*, World Scientific Singapore, 1998.

[4] S. Han, K. Yoon, I. Kweon, "A new Technique for Shot Detection and Key Frames Selection in Histogram Space", *Proc. 12th Workshop on Image Processing and Image Understanding*, pp. 475-479, 2000.

[5] G. Ciocca, R. Schettini, "Dynamic key-frame extraction for video summarization", Proc. SPIE , Vol. 5670, pp. 137-142, 2005.

[6] J.R. Kender, "On the Structure and Anlysis of Home Videos", *Proc. Asian Conference on Computer Vision*, Taipei, Taiwan, R.O.C., 2000.

[7] N. Dimitrova, T. McGee, H. Elenbaas, "Video Keyframe Extraction and Filtering: A Keyframe is not a Keyframe to Everyone", *Proc. of the sixth international conference on Information and knowledge management*, pp. 113-120, 1997.

[8] E. Veneau, R. Ronoard, P. Bouthemy, "From Video Shot Clustering to Sequence Segmentation", *IEEE International Conference on Pattern Recognition*, vol. 4, pp. 254-257, 2000.

[9] W. Zhang, Q. Ye, L. Xing, Q. Huang, W. Gao, "Unsupervised sports video scene clustering and its applications to story units detection", *Proc. SPIE Visual Communications and Image Processing*, vol. 5960, pp. 446-455, 2005.

[10] H. Lu, Y.P. Tan, "Unsupervised clustering of dominant scenes in sports video", *Pattern Recognition Letters*, vol. 24, pp. 2651-2662, 2003.

[11] W. Ren, M. Sharma, S. Singh, "Automated Video Segmentation", *Proc. 3rd International Conference on Information, Communications & Signal Processing*, 2001.
[12] Y. Choi, S. J. Kim, S. Lee, "Hierarchical shot clustering for video summarization*", ICCS 2002, Lecture Notes on Computer Science*, vol. 2331, pp. 1100-1107, 2002.
[13] D. G. Perez, A. Loui, M.T. Sun, "Finding Structure in Home Videos by Probablistic Hierarchical Clustering", *IEEE Transaction on Circuits and Systems for Video Technology*, vol. 13, no. 6, pp. 539-548, 2003.
[14] X. Zhu, X. Wu, J. Fan, A. K. Elmagarmid, W. G. Aref, "Exploring video content structure for hierarchical summarization", *Multimedia Systems*, vol. 10, pp. 98-115, 2004.
[15] M. Yeung, B.L. Yeo, B. Liu, "Extracting Story Units from long Programs for video browsing and Navigation", *Proc. IEEE Conference on Multimedia Computing and Systems*, pp. 296-305, 1996.
[16] D. Vercauteren, *Quality assurance for digital camera images*, EPFL-IVRG Project Report, 2003.
[17] Corchs S., Deco G., "Large-scale neural model for visual attention: integration of experimental single-cell and fMRI data", *Cerebral Cortex*, vol. 12, pp. 339-348, 2002.
[18] D.E. Rumelhart, G.E. Hinton, R.J. Williams, "Learning Internal Representations by Error Propagation", *Parallel Distributed Processing*, MIT Press, Cambridge, MA, Volume 1, pp. 318-362, 1986.
[19] M. Swain, D. Ballard, "Color Indexing", *International Journal of Computer Vision*, vol. 7, no. 1, pp. 11-32, 1991.
[20] R. Schettini, C. Brambilla, C. Cusano, G. Ciocca, "Automatic classification of digital photographs based on decision forests*", Int. J. of Pattern Recognition and Artificial Intelligence,* vol. 18(5), pp. 819-846, 2004.
[21] L. Breiman, J. H. Friedman, R. A. Olshen and C. J. Stone, *Classification and Regression Trees*, (Wadsworth and Brooks/Cole), 1984.
[22] Y.N. Deng, C. Henny, M.S. Moore, B.S. Manjunath, "Peer group filtering and perceptual color quantization", *IEEE Int. Symposium on Circuits and Systems*, vol. 4, pp. 21-24.
[23] C. Kenney, Y.N. Deng, B.S. Manjunath, G. Hewer, "Peer Group image enhancement", *IEEE Trans. Image Processing*, 10(2), 326-334.
[24] R.A. Fisher, "The use of multiple measures in taxonomic problems", *Ann. Eugenics*, vol. 7, pp.179-188, 1936.
[25] Li-Qun Chen, Xing Xie, Xin Fan,Wei-Ying Ma, Hong-Jiang Zhang, He-Qin Zhou, A visual attention model for adapting images on small displays, Multimedia Systems vol. 9 no. 4, pp. 353-364, 2003.
[26] AESS Archive: http://aess.itc.cnr.it/index.htm.

Gianluigi Ciocca received his degree (Laurea) in Computer Science at the University of Milan in 1998, and since then he has been a fellow at the Institute of Multimedia Information Technologies and at the Imaging and Vision Laboratory in the ITC Institute of the Italian National Research Council, where his research has focused on the development of systems for the management of image and video databases and the development of new methodologies and algorithms for automatic indexing. He is currently a PhD student in computer science at the Department of Information Science, Systems Theory, and Communication (DISCo)of the University of Milano-Bicocca, working on video analysis and video abstraction.

Raimondo Schettini is an associate professor at DISCo, University of Milano Bicocca where he is in charge of the Imaging and Vision Lab. He has been associated with Italian National Research Council (CNR) since 1987. He has been team leader in several research projects and published more than 160 refereed papers on image processing, analysis and reproduction, and on image content-based indexing and retrieval. He is an associated editor of the Pattern Recognition Journal. He was a co-guest editor of three special issues about Internet Imaging (Journal of Electronic Imaging, 2002), Color Image Processing and Analysis (Pattern Recognition Letters, 2003), and Color for Image Indexing and Retrieval (Computer Vision and Image Understanding, 2004). He was General Co-Chairman of the 1st Workshop on Image and Video Content-based Retrieval (1998), of the First European Conference on Color in Graphics, Imaging and Vision (2002), and of the EI Internet Imaging Conferences (2000-2006).