Hierarchical video summarization for medical data

Xingquan Zhu[a*] , Jianping Fan[b**] , Ahmed K. Elmagarmid[a*] , Walid G. Aref[a*]

[a]Dept. of computer science, Purdue University, W. Lafayette, USA

[b]Dept. of computer science, University of North Carolina at Charlotte, USA

ABSTRACT

To provide users with an overview of medical video content at various levels of abstraction which can be used for more efficient database browsing and access, a hierarchical video summarization strategy has been developed and is presented in this paper. To generate an overview, the key frames of a video are preprocessed to extract special frames (black frames, slides, clip art, sketch drawings) and special regions (faces, skin or blood-red areas). A shot grouping method is then applied to merge the spatially or temporally related shots into groups. The visual features and knowledge from the video shots are integrated to assign the groups into predefined semantic categories. Based on the video groups and their semantic categories, video summaries for different levels are constructed by group merging, hierarchical group clustering and semantic category selection. Based on this strategy, a user can select the layer of the summary to access. The higher the layer, the more concise the video summary; the lower the layer, the greater the detail contained in the summary.

Keywords: Hierarchical video summarization, video processing, video classification.

1. INTRODUCTION

Video is increasingly the medium of choice for a variety of communication channels, resulting primarily from increasing levels of networked multimedia content. One way to keep our heads above the video waters is to provide summaries of video content in a more tractable format. Many tools have been developed for browsing videos by key frames [1][14] or by video objects [2][15]. A curve simplification video summarization strategy is introduced in [3], which maps each video frame into a vector in high dimensional feature space, and then segments the feature curve into units. A video summary is extracted based on relationships between the units. In video *Managa* [4], a pictorial video summary is presented with key frames of various sizes, where the importance of the key frame determines its size.

Since it is hard to find a general method for automatically extracting video semantic content, some abstraction strategies have been applied to specific kinds of video sources, such as home videos [5], stereoscopic videos [2], online presentation videos [6], etc. These strategies use knowledge among the video to analyze its content structure. Similar with video summary strategies, some approaches summarize a video by skimming [7]. Video skimming may be useful for some purposes, since compared to still pictorial images, a skimmed video stream is more attractive for users. However, the amount of time required for viewing a skim suggests that skimmed video is not appropriate for a quick overview. Hence, neither pictorial abstract nor skimming has the greater value, and both are supported by the strategy presented in this paper.

In general, the aforementioned methods work with nearly the same strategy: grouping videos, selecting important units, acquiring users' specification of summary length, assembling. However, there are two problems with this strategy. First, important unit selection is a semantic-related topic. Different users have different value judgments, and it would be relatively difficult to determine how much more important one unit is than another. Second, the length of the users' specification for the summary is not always reasonable in unfolding the video content, especially if the user is unfamiliar with videos in the database.

Contact information: *{zhuxq,ake,aref}@cs.purdue.edu; **jfan@uncc.edu

Storage and Retrieval for Media Databases 2002, Minerva M. Yeung, Chung-Sheng Li,
Rainer W. Lienhart, Editors, Proceedings of SPIE Vol. 4676 (2002)

Ideally, a video summary should briefly and concisely present the content of the input video source. It should be shorter than the original, focus on the content, and give the viewer an appropriate overview of the whole. However, the problem is that what's *appropriate* varies form viewer to viewer, depending on the viewer's familiarity with the source and genre, and with the viewer's particular goal in watching the summary. Hence, a hierarchical video summary strategy would be most helpful in assisting the viewer to determine what is *appropriate*. In [12], a key frame based hierarchical video summary strategy is presented. Key frames are organized in a hierarchical manner from coarse to fine temporal resolution using a pairwise clustering process to merge neighboring key frames with significant similarity into one cluster. Instead of letting a user accept the generated video summary passively, *movieDNA* [16] supplies the user with a hierarchical visualized video feature map called *DNA*. The hierarchical video summary is interactive: by rolling the mouse over the *DNA*, users can brush through the video, pulling up detailed meta-information on each segment. However, both strategies fail to address the correlation among the video units (key frame, shot, scene) in temporal series, since similar units may be shown in the video several times.

In this paper, a hierarchical video summary strategy for medical video is presented. The strategy integrates visual information and common knowledge about the video, such that the semantic content and visual feature based correlation among units are both analyzed.

2. SYSTEM ARCHITECTURE

Fig. 1 presents the architecture of our hierarchical video summarization strategy. First, all key frames are classified into two categories: special frames and non-special frames. The skin, blood-red regions and faces in non-special frames are also detected for semantic classification. Then, a shot grouping method is applied to merge those temporally or spatially related shots into groups. Based on the group information and common knowledge about the medical video, a semantic classification strategy is implemented to assign groups into predefined semantic categories, such as "Dialog", "Presentation", etc. Since a video scene may be separated into several groups, a visual feature-based group merging method is also used to merge neighboring groups from the same semantic category into new groups. Finally, a group clustering method is applied to all semantic categories to erase those redundant groups and acquire a hierarchical structure. As a result, the video summary for different layer would be constructed by combining semantic categories and their structure information.

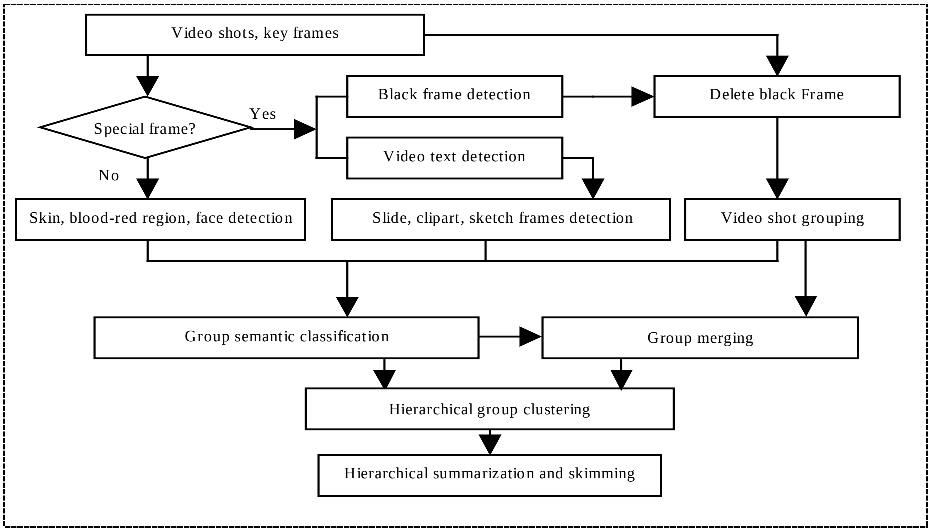
Video shots, key frames
Special frame?
Yes
No
Black frame detection
Video text detection
Delete black Frame
Skin, blood-red region, face detection
Slide, clipart, sketch frames detection
Video shot grouping
Group semantic classification
Group merging
Hierarchical group clustering
Hierarchical summarization and skimming

Figure 1. System architecture.

We do not use closed-captioned or speech recognition information. Closed-captioned text is not always available in medical videos, and audio signals in medical video are generally narrative descriptions of the disease, surgery, diagnosis, etc. instead of voices of persons from the video scene.

3. KEY FRAME PREPROCESSING

For the sake of simplicity, the first frame in each shot is selected as the key frame. To analyze video content efficiently, all key frames in the video are preprocessed to extract some useful information.

3.1 Special frame detection

Since medical videos are generally used for educational purposes and not for amusement, one relatively distinct feature is the existence of some manually created frames, such as black frames, slides, clip art, sketch drawings, etc. within certain video shots. We define these shots as *special shots*, and identify the key frames in these shots as *special frames*. Since these manually created frames may provide us with some semantic information about the video, special key frame detection is our first step in video processing.

According to our observations and statistics, special frames in medical videos usually have the following common features:

1. Since all special shots use the slides, clip art, etc. to introduce topics or present general information about the video content, the camera motion in the special shot would be still, i.e., the frame difference in the shot will be very low.
2. Because all special frames are man-made images (at least partly), the color information contained in special frames is relatively simple compared to the other natural images.
3. In general, the number of special shots in the medical video is far less than the number of non-special shots.
4. Since special frames are man-made images and are mainly used to present textual information or illustrations, most have a special background with relatively simple color and texture when compared to other natural images. Hence, they have a very low similarity with the other non-special frames. Also, since a medical video usually covers a single topic, the non-special frames will hopefully have a relatively larger average similarity with each other. Thus, the average similarity between a given key frame and all other key frames indicates the possibility that the given key frame is also a special frame.

Based on all these features, three factors are extracted for special frame detection.

3.1.1 Region based frame difference (***RBFD***)

As mentioned in Section 3.1, the frame difference among special shots may be used for shot classification. However, for some special shots which contain slides, the text or figures in the slide may change slightly as time advances, since the speaker would use those changes to develop his topic step by step. As shown in Fig.2, slides *a*, *c*, and *d* belong to the same shot. Note that as time advances, the text among slides changes slightly, but the title and background region are unchanged. Hence, a region-based method is proposed to detect changes in the title or background region in each shot.

$$ED_i = \frac{\sum\limits_{j=1}^{N_i-1} \sum\limits_{w=0,h=0,P_{0,w,h}\in E}^{h=H-1,w=W-1} |(P_{i,j,w,h} - P_{i,0,w,h})|}{N_i \sum\limits_{w=0,h=0,P_0,w,h\in E}^{h=H-1,w=W-1}} \tag{1}$$

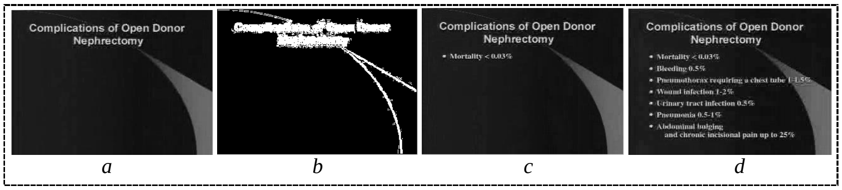
Complications of Open Donor Nephrectomy
Complications of Open Donor Nephrectomy
Complications of Open Donor Nephrectomy
• Mortality < 0.03%
Complications of Open Donor Nephrectomy
• Mortality < 0.03%
• Bleeding 0.5%
• Pneumothorax requiring a chest tube 1-1.5%
• Wound infection 1-2%
• Urinary tract infection 0.5%
• Pneumonia 0.5-1%
• Abdominal bulging and chronic incisional pain up to 25%
a
b
c
d

Figure 2. Region based frame difference. Images *a*, *c*, and *d* are slides in the same shot, image *b* is the segmented region areas of image *a*. As time advances, the contents of the slides in the shot change slightly.

For any shot $S_{i,}$, suppose there are N_i frames $F_{i,j}$ ($j=0,..N_i-1$) contained in the shot. The edge operator is used first to detect edge pixels in the first frame $F_{i,0.}$ Then, any pixel ($P_{i,0,w,h}$ w$\in[0,W]$, $h\in[0,H]$ in $F_{i,0}$ *(where* W,H are the width and height of the image), would be classified into two categories: edge pixel (E) versus non-edge pixel($\overline{E}$). The region based frame difference (ED_i) in shot S_i is defined by Eq.(1). The value of ED_i indicates the possibility that the current shot S_i belongs to the special shot category, where the lower the value, the higher the probability. However, since any exactly still shot will result in a low value in ED_i, this strategy can only be used to select special shot candidates: any shot S_i with ED_i less than threshold T_{Edge} is taken as a special shot candidate.

3.1.2 Average frame similarity (*AFS*)

Suppose there are M shots ($S_{i,}$ $i=0,\ldots,M-1$) in the current video, The average frame similarity would be defined as the average similarity between each key frame and all other key frames, as given by Eq.(2).

$$AFS_i = \frac{\sum_{j=0,j\neq i}^{M-1} sim(S_i, S_j)}{M-1} \tag{2}$$

Since special frames are man-made images, they usually contain special backgrounds and relatively simple color or texture. In addition, the number of special shot in the video is far less than the number of other kinds of shots. Hence, the average similarity of a special key frame would be relatively lower than that of the non-special frames.

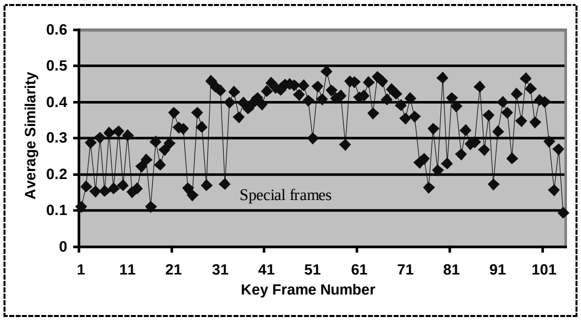

Figure 3. Average frame similarity in the video

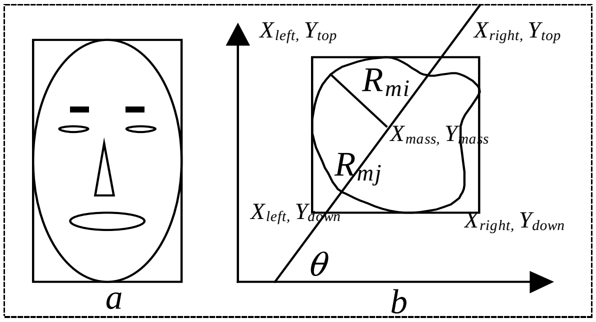

Figure 4. The relationship constraints for face detection

Figure 2 presents an experimental result of the average similarity between each key frame and all other key frames, where the x-axis identifies the key frame number and the y-axis indicates the average visual feature similarity (using a 256-bin color histogram in *HSV* color space). There are 105 key frames (shots) in the video; 16 of them are special frames as defined above. It's obvious that the special frame and non-special frame have a relatively distinct distribution in average similarity. A single threshold T_{AvgSim} (T_{AvgSim} is equal to 0.2 in Fig. 2) can be used for classification. The *AFS* can be used to find special shots that contain the motion or changes in motion, and these are extracted as special shot candidates as well.

3.1.3 Dominant color percentage (*DCP*)

There is no doubt that the methods described above (for *RBFD* and *AFS*) may run the risk of selecting some non-special shots as special shot candidates. In this section, dominant colors will be used to refine the special shot candidates. Suppose $p_{i,c}$ ($i=0,\ldots,M-1;\ c=0,1\ldots,L-1;\ L$ is the number of the dominant color) is the percentage (ranked in descending order) of the dominant color c in frame i,. Then, the sum of the first 4 dominant colors' percentages ($DCP_i = \sum_{c=0}^{3} p_{i,c}$) indicate the color complexity of each image, where the lower the DCP_i , the more complicated the color contained in the image. In this manner, the special frames would be assigned a larger DCP_i value. However, due to the diversity of the natural images, there is no evidence to show that any key frame with simple color information must belong to the special key frame category. Our experimental results show that almost all special key frames are assigned a higher value (larger than 0.8) for DCP_i. There are also some non-special frames with large values in DCP_i, but they generally also have a larger value for ED_i or ASF_i. Hence, this indicates that the *DCP* can be used to refine special frame candidates efficiently.

3.1.4 Special frame detection strategy

Based on the three factors above, our algorithm for special frame detection is described as follows:

Input: Video shots, S={*shot0*, …, *shot N*}; **Output:** The special shot and special frame.
Procedure:

1. Input all video shots; use the Prewitt edge detector to segment the first frame of each shot.
2. For each shot *i*, calculate its edge based frame difference ED_i using Eq.(1). If this value is lower than threshold T_{Edge} (T_{Edge}=10 in our system), the shot is taken as a special shot candidate.
3. Return to step 2 to process all other shots in the video so that all shots in the video are classified into two categories: special shot candidates versus non-special shot candidates.
4. For all non-special shot candidates, calculate their AFS_i using Eq.(2). Video shots with value smaller than the threshold T_{AvgSim} (T_{AvgSim} =0.2 in our system) are also taken as special shot candidates.
5. For all special shot candidates, a 36-bin histogram in *HSV* color space (H=18, S=2) is calculated, and their DCP_i is computed. A candidate shot with *DCP* value larger than threshold $T_{DomainColor}$ ($T_{DomainColor}$ =0.8 in our system) would be identified as a special shot, and its key frame is identified as special frame.

3.2 Special frame classification

Currently, all detected special frames are classified into three categories: black frames, slides and clip art, and sketch drawings. In order to detect black frames, each special frame is separated into several blocks; if the average intensity of a block is lower than a threshold, it is identified as a black block. A frame with all blocks identified as black blocks will be identified as a black frame.

Since slides are often used for presentations, they generally contain some text. Hence, our algorithm to detect slides is relatively easy compared to other Gaussian model based methods [8]: Any special frame containing text will be identified as a slide. For this, a relatively simple videotext detection method is used [9]. Since we check only whether there is text in a given frame, the precision and recall of the strategy are acceptable. According to our experimental results, with about 188 special frames (containing 84 slides), the precision and recall were 91% and 93% respectively.

After all the black frames and slides are detected, the remaining special frames are specified as clip art or sketch drawing frames. Currently, no other refining strategy is used to distinguish between clip art or sketch drawing frames, since shots containing either will always belong to the same semantic category defined in Section 5.

3.3 Face and special color region detection

Since many medical videos are related to humans, the face, skin or blood-red regions will give us some hint in acquiring the semantic content. A color based region detection method is used to detect special regions (skin regions and blood-red regions) and the human face. The skin color map is generated automatically from a set of normalized face training images. The generated skin color distribution model is used to detect the homogeneous color regions, where homogeneous color regions that have higher skin color likeness should be taken as the skin color regions. The same strategy is used to detect the blood-red regions.

Since face regions should be included in the collection of skin color regions, the identified skin color regions are first taken as candidates for human faces. However, the mere presence of a skin-color region is not enough to hypothesize a face; human faces are then verified by a set of facial filters [10]. The design of the facial filters is based on shapes of human faces and facial features, such as the elliptical shape of facial region and the overall spatial relationship constraints among these facial features, as shown in Fig.4. Since the aspect ratio and area ratio of a human face should be distributed in a narrow range, the aspect ratio and area ratio filters are first performed. The Hausdorff distance is then used as a measure for comparing the similarity between the shape of the face candidate and the boundary of its corresponding ellipse model. Face candidates whose similarity measures are less than a pre-specified threshold are rejected. Since those small faces or skin regions would have less importance in addressing the content of the video, any skin region with size less than 5% of the image size is deleted from further processing. This restriction would not, however, be applied in detecting the blood-red regions. A set of face detection results is shown in Fig.5.

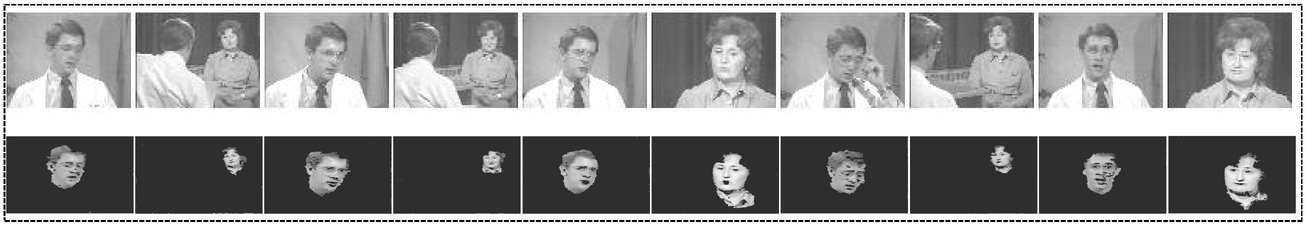

Figure 5. Face detection results. The top row contains the original images in a dialog scene; the bottom row contains the face detection results for the scene.

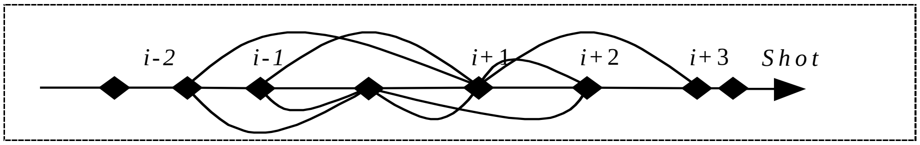

Figure 6. Shot grouping strategy

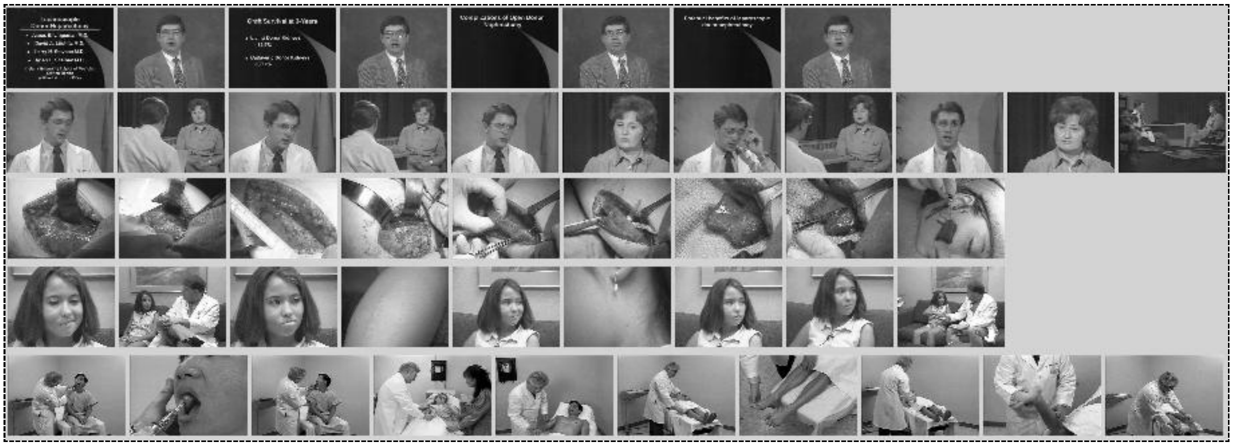

Figure 7. Shot grouping results, with groups from the top row to the bottom row identifying (in order): presentation, dialog, surgery, diagnosis and diagnosis.

4. SHOT GROUPING

In order to partition a video into semantically richer entities, shots must be grouped together based on content. The ultimate goal is to automatically determine a shot cluster which a human would judge as a "scene". Compared to other strategies that emphasize grouping all shots in the scene into one unit [11], our method emphasizes an analysis of the detailed information to determine the correlation among shots in the scene. For this, a given shot is compared with the shots that precede and succeed it (no more than 2 shots) to determine the correlation between them, as shown in Fig.6. Suppose $Sim(S_i,S_j)$ is the similarity between shot i and j. Our video shot grouping procedure is stated below:

$$CL_i = max\{ Sim(S_i,S_{i-1}), Sim(S_i,S_{i-2})\}; \qquad CR_i = max\{ Sim(S_i,S_{i+1}), Sim(S_i,S_{i+2})\} \tag{3}$$

$$CL_{i+1} = max\{ Sim(S_{i+1},S_{i-1}), Sim(S_{i+1},S_{i-2})\}; \quad CR_{i+1} = max\{ Sim(S_{i+1},S_{i+2}), Sim(S_{i+1},S_{i+3})\} \tag{4}$$

$$R(i)=(CR_i+CR_{i+1})/(CL_i+CL_{i+1}). \tag{5}$$

For any shot S_i, if $R(i)>TH_1$ or $CR_i<TH_2$, $LR_i<TH_2$, we claim a new group starts at S_i, otherwise, S_i is absorbed in the current group (the thresholds TH_1, TH_2 are set as TH_1=1.5, TH_2=0.5). We use a 256-bin dimensional *HSV* color

histogram in our algorithm to evaluate the similarity between shots. Suppose $H_{i,j}$, $j \in [0,255]$ is the normalized color histogram of the key frame *i*, then the similarity between shot *i*, *j* is defined by Eq. (6).

$$Sim\ (S_i, S_j) = \sum_{k=0}^{255} \min(H_{i,k}, H_{j,k}) \qquad (6)$$

Using this strategy, two kinds of shots would be absorbed into a given group:

- Shots related in temporal series, such as a dialog, presentation, where similar shots are shown recursively.
- Shots similar in visual perception, where all shots in the group are similar in visual features.

Figure 7 presents the experimental results of our video grouping strategy. Since all shots in one scene are semantically related, parts of the shots would share the same background or the same dominant color. Hence, this operation will help in merging the shots of the scene into one or several groups. However, due to the inherent insufficiency of low-level features in addressing the semantic content, this operation will not entirely unfold the semantic structure of the video. Thus, the next step in our process will emphasize the semantic analysis and discovery of correlations among groups. Our goal is to use this analysis to help us acquire an efficient hierarchical video summary.

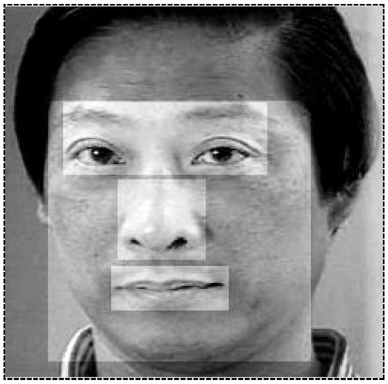

Figure 8. Regions for template matching strategy

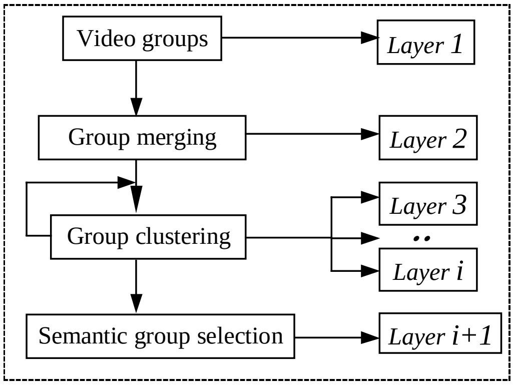

Figure 9. Hierarchical video summarization

5. SEMANTIC CLASSIFICATION

Currently, it would be difficult to develop a general semantic classification method that applied to all images or videos. However, since we have focused our research on medical videos, this restriction helps us in acquiring semantic information about the video. According to our statistics, we have defined four semantic categories: “Clinical Operation”, “Dialog”, “Presentation”, and “Unknown”. Any group obtained in Section 4 is assigned to one of these categories.

Since medical videos are often used for teaching, presentation or other educational purposes, the video content is usually recorded or edited in one of three ways:

- Using presentations by the doctor or experts to express the general topic information of the video, for example, to explain a new kind of surgery, present the definition of a disease, etc.
- Using clinical operations (such as the diagnosis, surgery, the picture of the organ, etc.) to present the detail of the disease, their symptoms, comparisons and steps of the surgery, etc.
- Using dialog between the doctor and patients to acquire other knowledge about the disease, such as how the patient feels, their lifestyle, precautionary methods, etc.

There may be other kinds of scenes which are not covered by these categories, however, according to our analysis of 4 hours of medical video data (where all videos are processed by using our video grouping methods first, and then by using human judgment to determine the semantic content of each group), nearly 90% of the video groups are covered by these semantic categories. If we can assign each video group into the corresponding semantic category, it would be a significant step toward determining the structure of the video. First, we state the definition of our semantic categories:

1. A “Presentation” is defined as a group of shots that contain at least one slide or one clip art/sketch drawing frame. At least one close face (human face with size larger than 10% of the total frame size) should be shown in the group.
2. A “Dialog” is a group of shots containing at least two different faces which are shown in different shots, where at least one person’s face is repeated one or more times.
3. The “Clinical Operation” includes three kinds of medical scenes (surgery, diagnosis, symptoms). Since it often shows detailed information related to the clinical topic, most of the shots in this category contain skin or blood-red regions. In this paper, the “Clinical operation” is defined as a video group containing at least one blood-red region or one close skin region (skin region with size larger than 20% of the total frame size). Also, if more than 80% of the shots contain skin or blood-red regions, the video group is identified as a clinical operation.
4. If the current group does not belong to any of the three semantic categories above, it is defined as “Unknown”. Fig.7 presents several video groups and their semantic category. It is relatively easy to find the visual differences among these semantic categories.

Based on these definitions, our semantic classification algorithm is described as below:

Input: Video group, G_i={$shot_{i,0}$, …, $shot_{i,k}$ }; **Output:** Category for the current video group.
Procedure:

```
1. Input the video shots and their key frame preprocessing results. If there are no remaining groups, stop.
2. Test whether the current group belongs to “Presentation”:
    a. If there is no slide, clip art, or sketch-drawing frame contained in the group, go to step 3.
    b. If there is no close face contained in the group, go to step 3.
    c. Assign the current group to the “Presentation” category; go to step 1 to process another video group.
3. Test whether the current video group belongs to “Dialog”:
    a. If there is either no face or less than two shots contain faces, go to step 4.
    b. If all faces among shots belong to one person, go to step 4.
    c. If at least one person’s face is shown two or more times, the current group is claimed as a “Dialog”, go to step 1 to process another video group, otherwise go to the next step.
4. Test whether the current video group belongs to “Clinical Operation”:
    a. If there are any close skin regions or blood-red regions detected, the current video group is assigned to “Clinical Operation”; go to step 1.
    b. For all shots in the group, if more than 80% of the key frame contains skin regions, the video group is assigned as a “Clinical Operation”, go to step 1 to process another group.
5. Assign the group to “Unknown” and go to step 1.
```

In order to detect the “Dialog”, all faces in the group are classified and recognized to verify whether they belong to one or to different persons. For the sake of simplicity, a template-based face matching strategy is used in our system [13]. In the first step, faces within shots are normalized based on the interocular distance and direction of the eye-to-eye axis to achieve scale and rotation invariance; Then, a set of four masks representing eyes, nose, mouth, and face (the region from the eyebrows downwards, as shown in Fig.8), is used the capture the features of the face. When attempting recognition, any two faces in the group are compared, returning a vector of matching scores computed through normalized cross-correlation among the mask areas of the faces. Those scores would be used to verify whether those faces belong to one or to several persons.

6. HIERARCHICAL VIDEO SUMMARIZATION AND SKIMMING

This section describes a hierarchical video summarization strategy based on video groups and their semantic classifications,. The flow chart of the algorithm is shown in Fig.9, which indicates that the hierarchical video summary consists of video groups at different levels. Video summary at the lowest layer (*Level 1*) is consisted with all video groups to uncover details among the videos. Then, the group merging is applied to all neighboring groups to determine the second layer summary. Afterwards, a group clustering operates recursively among each semantic category to construct the other layers. After the video clustering operation is completed, a model based semantic category selection is used to select video groups to form the highest-level video summary. The diagram in Fig.10 illustrates the corresponding steps.

6.1 Group merging

Because our shot grouping method may fail in grouping a scene into one group, a group merging method is necessary to discover correlations among neighboring groups. However, groups that belong to different semantic categories would have low correlation with each other, even if they have a relative high visual similarity. Hence, our group correlation is applied to those neighboring groups in the same semantic category (except “Unknown”, the “Unknown” group can be absorbed by any other group only if their correlation is large enough).

Assume G_i is any video group i, N_i is the number of shot in G_i and $S_{i,k}$ $(k\in[1,N_i])$ is shot k in G_i. Given a threshold $T_{merging}$ ($T_{merging}$=0.4 in our system) and two groups G_i or G_j, for any shot in G_i (or G_j), the maximal similarity between it and all shots in the group G_j (or G_i), $Max\{sim(S_{i,k},S_{j,l}),\ l=1,..N_j\},\forall k,\ k\in[1,N_i]$ is calculated first. If this value is larger than $T_{merging}$, these two shots are identified as similar shots; otherwise they are non-similar shots. After all shots in G_i and G_j have been processed in this way, the rate between average similarity of similar shots and non-similar shots, R_S and the rate between the number of similar shots and non-similar shots, R_N, are obtained as two factors for evaluating the correlation between G_i and G_j. The higher the values are, the larger the correlation. Hence, the product of R_s and R_N is used to determine whether neighboring groups should be merged in to one group or keep them unchanged, as shown in row 4 and 5 of Fig.10. Since a group with too many shots may result in an ambiguous semantic content, and in addition, may result in a higher probability to absorb other groups, any group containing more than 20 shots is no longer used for group merging.

The semantic category of each merged group remains the same as that of the old group, except for the “Unknown” group; any group merged with “Unknown” will maintain its semantic category as that of the merged group.

6.2 Hierarchical group clustering

In Section 6.1, the group that is merging is used to merge groups with large visual similarity into a new group; however, this operation is only applied to neighboring groups in the same semantic category. As we know, scenes in different parts of a video may be similar both in visual perception and semantic content; a compact video summary should eliminate those redundant groups efficiently. In this section, a group clustering method is used to address this problem by clustering those groups into a hierarchical structure.

The group clustering is also applied to groups in the same semantic category. For any two groups in the same semantic category, the group merging method described in Section 6.1 is used to calculate the correlation between them. After all correlations have been computed, those values are used to merge groups with large correlation. As in Section 6.1, any group containing more than 20 shots would not take part in the group clustering. The representative group for each cluster is selected by considering how many shots are contained in each group, with the group having the largest number of shots being selected as the representative group of the cluster.

According to our experimental results, a two layer hierarchical cluster is acceptable for most medical videos (that is, i=4 in Fig.10). Thus, the group clustering is applied recursively twice among each semantic category. Afterwards, those reserved groups (or representative groups) are used to construct the video summary at each layer i.

6.3 Semantic category selection

In general, medical video scenarios can be separated into three parts: (1) presenting subjects or topics, (2) showing evidence and details, and (3) drawing conclusions. These three parts are often shown separately at the front, middle, and back of the video; a simple model is shown in Fig.11. Hence, the summary at the highest layer (layer i+1) may be constructed by selecting the meaningful semantic categories from layer i to fit this model.

The semantic category selection is executed by selecting two groups that belong to different semantic categories from each part (front, middle, back) of the video, and then assembling those six groups to form a video summary at the highest layer. With our analysis from Section 5, different semantic categories have differing importance in addressing the video content and detail, hence, the order of the semantic category selection for each part is different and has been predefined. For groups at the front and the back, the order of selection is: “Presentation”, “Dialog”, “Clinical Operation”, “Unknown”. The selection order for the middle part is: “Clinical Operation”, “Presentation”, “Dialog”, “Unknown”. For any part of the video, two groups will be selected from layer *i* according to their selection order and semantic category. Then, those six groups will be used to form the video summary at the highest level.

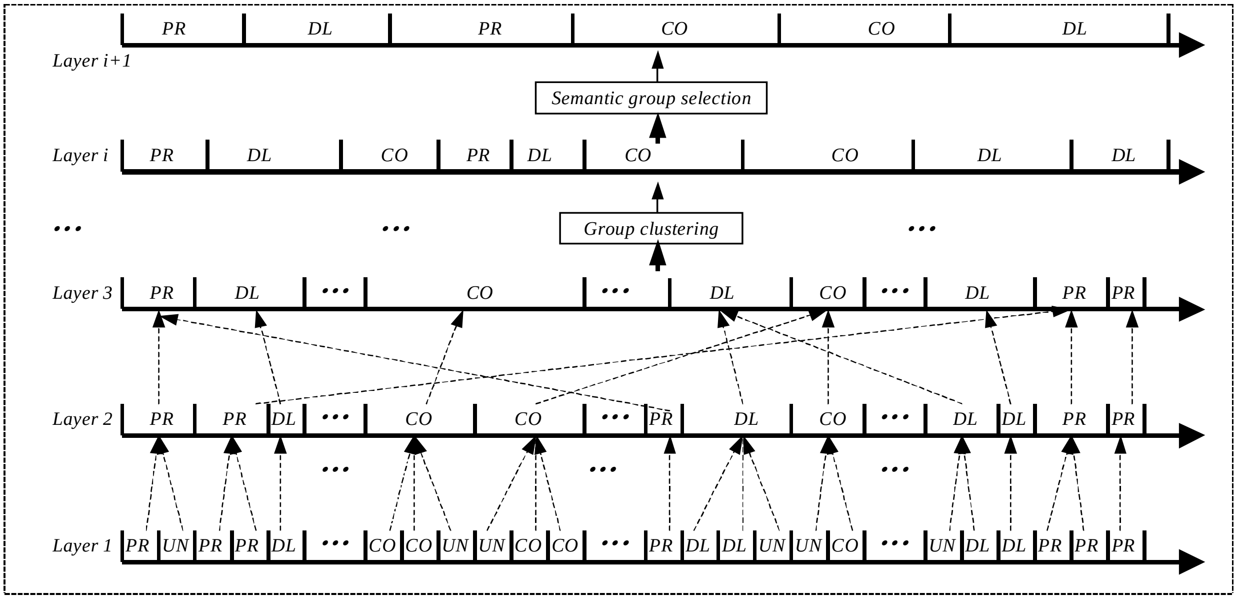
Layer i+1
PR DL PR CO CO DL
Semantic group selection
Layer i
PR DL CO PR DL CO CO DL DL
Group clustering
...
Layer 3
PR DL ... CO ... DL CO ... DL PR PR
Layer 2
PR PR DL ... CO CO ... PR DL CO ... DL DL PR PR
Layer 1
PR UN PR PR DL ... CO CO UN UN CO CO ... PR DL DL UN UN CO ... UN DL DL PR PR PR

Figure 10. Hierarchical video summarization strategy
(*PR*, *DL*, *CO*, *UN* represent “Presentation”, “Dialog”, “Clinical Operation”, and “Unknown” respectively)

6.4 Represent frame selection and video skimming

After the hierarchical summarization at each layer has been successfully generated, a series of representative frames for each group is selected using the strategy below to present a pictorial video summary:

1. The representative frames for a “Presentation” group are the key frames of the firsts two shots that contain the slide (or clip art, sketch drawing frame) or the close face, respectively.
2. The representative frames for a “Dialog” group are the key frames of the first two shots that contain the face of different persons, respectively.
3. The representative frames for a “Clinical Operation” or “Unknown” group are key frames of the first and last shots in the group.

After the video summary at a certain layer has been generated successfully, the shots with the key frames which have been selected as the representative frames for the video summary are also extracted as representative shots, and a hierarchical video skimming is produced by assembling those shots according to their temporal sequence.

7. EXPERIMENTAL RESULTS

Two kinds of experimental results, semantic classification and hierarchical video summarization, are presented in this section. About 8 hours of medical videos (*MPEG-I* encoded) which describe face repair, nuclear medicine, laparoscopy, skin examination and laser eye surgery, etc. are used as our test bed.

First, we use the strategy in Section 4 to assign video shots into groups. Then, we assign each group with a semantic category using human judgment, and use the results to verify the efficiency of our semantic category classification strategy. The experimental result of semantic classification is shown in Table 1 (*P* and *R* represent precision and recall, respectively). It can be found that, the classification results for “Presentation” and “Dialog” are much better than ”Clinical Operation”, since those two categories have distinct visual features related to semantic definition. However, without the context information, a “Clinical Operation” may be taken as an “Unknown”, since our definition of “Clinical Operation” cannot cover all situations. Hence, some groups have been falsely classified into other groups, due to the failure of special frame detection, ambiguity of the semantic content, or the inherent insufficiency of visual features.

Fig. 12 presents the experimental results of the hierarchical video summarization. As stated before, we specify *i*=4 in our experiments, so that 5 layer video summaries are produced for each video. Three questions are introduced to evaluate the quality of the video summaries at each layer: (1) How well do you think the summary addresses the main topic of the video? (2) How well do you think the summary covers the scenario of the video? (3) Is the summary concise? For each of the questions, a score from 0 to 5 (5 indicates the best) will be specified by five student viewers after viewing the video summary at each level. Before the evaluation, viewers are asked to browse the entire video to get an overview of the video content. An average score for each level is computed from the students’ scores (shown in Fig.12). A second evaluation process used the rates between the numbers of representative frames at each layer and the number of all key frames (*RC)* to indicate the compression rate of the video summary. In order to normalize this value with the scores of the questions, we multiply *RC* by 5, and use this value in Fig.12.

From Fig.12, we see that as we move to the lower levels, the ability of the summary to cover the main topic and the scenario of the video is greater. The conciseness of the summary is worst at the lowest level, since as the level decreases, more redundant groups are shown in the summary. At the highest level, though the video summary exhibits an inferior ability to describe the scenario of the video, it can supply the user with a concise summary and relatively clear topic information. Hence, this level can be used to show differences between videos in the database. It was also found that the fourth level acquires a relatively optimal scores for all three question, thus, this layer is the most suitable for giving the user an overview of the video selected from the database for the first time.

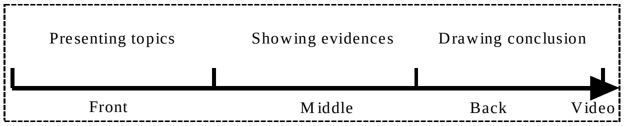

Figure 11. Common model of the medical video.

Table 1. Semantic classification result

Semantic category	Real groups	Detected groups	*P*	*R*
Presentation	10	12	91.2	90.1
Dialog	24	20	83.5	80.0
Clinical operation	33	26	66.8	71.0

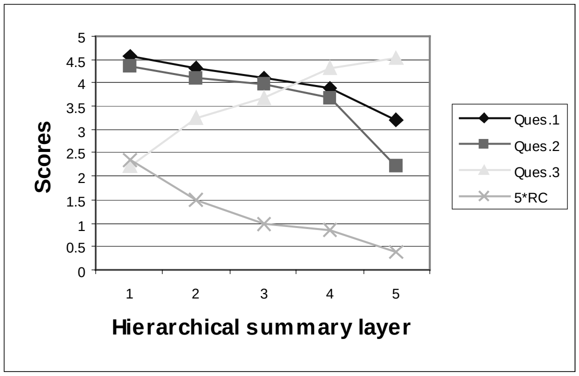

Figure 12. Hierarchical video summary evaluation

The system is implemented in *C++* with an *MEPG-I* decoder which was developed by our group. Since we need to generate the video skimming at each level, *MPEG-I* edit tools have also been developed to assemble several video clips into one *MPEG-I* stream with integrated audio signal.

8. CONCLUSION

In this paper, we have addressed the problem of automatic hierarchical video summarization for medical data. Unlike other strategies which select the important video units first and then assemble these units into a video summary, we present a hierarchical video summary strategy which integrates the semantic analysis, group correlation and clustering together. The hierarchical video summary generated by our strategy considers both the semantic content structure, group based visual perception and redundancy of the video. To produce the summary, methods for video shot grouping, knowledge based semantic classification, visual feature based group merging, and clustering are also introduced.

Video summaries that only take into account the fairly low-level features of the audio, video or closed-captioned tracks put a great deal of emphasis on the detail but not on the content. Most video that has been produced for a particular purpose (as opposed to informally shot, raw footage) has a deliberate structure. Video content or structure analysis would be necessary before the video summarization, since the most useful summary may not be just a collection of the most interesting visual information. Hence, our hierarchical summarization strategy should obtain a more reasonable result. With this hierarchical video summarization strategy, the video data can be parsed into a hierarchical structure, with each node consisting of one or several video groups. As a result, the structure can also be used for video indexing and hierarchical browsing.

ACKNOWLEDGEMENTS

This work was supported by NSF under 9972883-EIA, 9974255-IIS and 9983249-EIA, a grand from the state of Indiana 21th Century Fund, and by grants from HP, IBM, Intel, NCR, Walmart, and Telcordia.

REFERENCES

1. M. M. Yeung and B.-L. Yeo, “Video visualization for compact presentation and fast browsing of pictorial content”, *IEEE Transaction on CSVT, Vol.7, pp.771-785, Oct. 1997.*
2. N. D. Doulamis, A. D. Doulamis, Y. S. Avrithis, K. D. Ntalianis, S. D. Kollias, “Efficient Summarization of Stereoscopic Video Sequences”, *IEEE Transactions on CSVT. Vol.10, No.4, June. 2000.*
3. D. DeMenthon, V. Kobla, D. Doermann, “Video Summarization by Curve Simplification”, *Proceedings of the sixth ACM international conference on Multimedia, ep.13-16, Bristol, U.K. 1998.*
4. S. Uchihashi, J. Foote, A. Girgensohn, J. Boreczky, “Video *Managa*: Generating Semantically Meaningful Video Summaries”, *Proceedings of the seventh ACM international conference on Multimedia,* Oct. 1999, Pages 383 – 392, *Orlando, FL, U.S.A.*
5. R. Lienhart, “Abstracting Home Video Automatically”, *Proceedings of the seventh ACM international conference (part 2) on Multimedia (Part 2), 1999, Pages 37 – 40.*
6. L.W. He, W. Sanocki, A. Gupta, J. Grudin, “Auto-summarization of audio-video presentations”, *In Proceedings of the 7th ACM international conference on Multimedia, p.489-498, Oct.30-Nov.5, Orlando, FL, 1999.*
7. M. G. Christel, A. G. Hauptmann, A.S. Warmack, S.A. Crosby, “Adjustable Filmstrips and Skims as Abstractions for a Digital Video Library”, *IEEE Advances in Digital Libraries Conference, Baltimore, MD, May 19-21, 1999.*
8. Girgensohn, J. Foote, “Video Classification Using Transform Coefficients” *In* Proceedings of the International *Conference on Acoustics, Speech, and Signal Processing (Phoenix, AZ), vol. 6, pp. 3045-3048, 1999.*
9. N. Dimitrova, H. Elenbaas, T. McGee, E. Leyvi, L. Agnihotri, “An Architecture for Video Content Filtering in Consumer Domain”, *In proceedings of the international conference on information technology: coding and computing (ITCC'00), 2000.*
10. J.P. Fan, X.Q. Zhu, L.D. Wu, ``Automatic model-based semantic object extraction algorithm'', *IEEE Trans. on Circuits and Systems for Video Technology, vol.11, no.10, pp.1073-1084, Oct., 2001.*
11. T. Lin, H.J. Zhang “Automatic Video Scene Extraction by Shot Grouping”, *Proc. ICPR 2000.*
12. K. Ratakonda, M. I. Sezan and R. Crinon, “Hierarchical video summarization”, *IS&T/SPIE Conference on Visual Communications and Image Processing’99, Vol. 3653 pp.1531-1541, San Jose, January, 1999.*
13. R. Brunelli, T. Poggio, “Face Recognition: Features versus templates”, *IEEE Transactions on Pattern Analysis and Machine Intelligence, 15:1042-- 1052, 1993.*
14. R. Lienhart, S. Pfeiffer and W. Wffelsberg, “video abstracting”, *Communication of ACM, vol.40, Issue 12, pp.54-62, Dec, 1997.*
15. C. Kim and J. N. Hwang, “An integrated scheme for object-based video abstraction”, *In proceedings of the 8th ACM international conference on Multimedia, pp.303-311, Los Angeles, 2000.*
16. D. Ponceleon, A. Dieberger “Hierarchical brushing in a collection of video data”, *Proceedings of the 34th Hawaii international conference on system sciences, 2001.*