AUTOMATIC CONSUMER VIDEO SUMMARIZATION BY AUDIO AND VISUAL ANALYSIS

Wei Jiang[1], *Courtenay Cotton*[2*], *Alexander C. Loui*[1]

[1] Corporate Research and Engineering, Eastman Kodak Company, Rochester, NY
[2] Electrical Engineering, Columbia University, New York, NY

ABSTRACT

Video summarization provides a condensed or summarized version of a video stream by analyzing the video content. Automatic summarization of consumer videos is an important tool that facilitates efficient browsing, searching, and album creation in the large amount of consumer video collections. This paper studies automatic video summarization in the consumer domain where most previous methods can not be easily applied due to the challenging issues for content analysis, *i.e.*, the consumer videos are captured with uncontrolled conditions such as uneven lighting, clutter, and large camera motion, and with poor-quality sound track as a mix of multiple sound sources under severe noises. To pursue reliable summarization, a case study with real consumer users is conducted, from which a set of consumer-oriented guidelines are obtained. The guidelines reflect the high-level semantic rules, in both visual and audio aspects, which are recognized by consumers as important to produce good video summaries. Following these guidelines, an automatic video summarization algorithm is developed where both visual and audio information are used to generate improved summaries. The experimental evaluations from consumer raters show the effectiveness of our approach.

Keywords— video summarization, consumer domain, audio summarization

1. INTRODUCTION

The proliferation of digital cameras has led to an explosion in the number of digital videos created, resulting in personal video databases large enough to require automated tools for efficient browsing, searching, and album creation. Video summarization is a mechanism to produce a condensed or summarized version of the original video by analyzing the underlying content in the entire video stream. Being an important tool to facilitate video browsing and search, video summarization has been largely explored in previous literatures. In general, all types of information have been used to help summarization, including text descriptions, visual appearances, and audio sounds. A relatively comprehensive survey can be found in [1]. Most previous works analyze videos with good quality, *e.g.*, with relatively high resolution, stable camera, low background noise in both audio and visual signals. Specifically, they mainly focus on certain video genres like sports, News, TV drama, movie dialog, or documentary videos. So far, very few work has been done to study consumer-quality videos, which are captured under uncontrolled conditions and have very diverse content.

One major reason that there lacks research on consumer video summarization is because of the challenging issues for content analysis in the consumer domain. First, in general there is no embedded text like subtitles or text captions, and methods relying on text features [2] can not be used. Second, different from sports videos or television drama, there usually lacks specific domain knowledge to guide summarization systems due to the diverse video content. Third, a consumer video typically has one long shot, with challenging conditions such as uneven lighting, clutter, occlusions, and complicated motions of multiple objects and the camera. The mixed sound track is also generated by multiple sound sources under severe noises. It is difficult to identify specific objects or events from the image sequences, and it is hard to identify semantically meaningful audio segments like nouns, exited/normal speeches, *etc.* Methods relying on object/event detection [3, 4], or special sound effect detection [4, 5] can not be easily applied. Also, it is hard to robustly assess distortion or detect object/camera motions, and other non-domain specific methods such as [6, 7] can not perform well either. Another difficulty of conducting video summarization in the consumer domain is that it is hard to assess users' satisfaction with the summaries. Previous studies [8, 9] show that both the structured content units, *e.g.*, the sequence of scenes, and special interesting events are important to users, and their evaluation is genre-dependent and context-dependent. Due to the subjective nature of the problem, the real consumer needs can only be obtained from consumer studies.

In this work, we explore video summarization in the consumer domain. We focus on four popular consumer categories: "birthday", "wedding", "show", and "parade". Videos from these categories usually have very diverse visual and audio content. For example, different parts of a wedding can look/sound very differently and there is even more diversity among different weddings. A case study is conducted over 50

[*] The work was conducted during Courtenay Cotton's internship in Eastman Kodak Company.

consumer videos from these categories, and from users' responses we obtain a set of consumer-oriented guidelines for generating video summaries. The guidelines reflect the high-level semantic rules, in both visual and audio aspects, that are recognized by consumers as important to produce good summaries. Following these guidelines a video summarization approach is developed where both audio and visual information are used to generate improved summaries. Specifically, we take into account the following factors: audio segmentation and classification, audio diversity, visual diversity, face quality, and overall image quality. In experiments, a total of seven consumer raters manually evaluate each of our summaries and compare it with an intuitive summary generated in the traditional way. The evaluation results show that our summaries can outperform the traditional summaries and better accommodate consumer needs.

In the remaining paper, we first describe the case study in Section 2, followed by our video summarization approach in Section 3 and 4. Section 5 and 6 give experiments and conclusion, respectively.

2. OBSERVATIONS FROM A CASE STUDY

Before designing an algorithm, it is important to understand what an ideal summary should be. The answers can only come from real consumers. Here we conducted a case study with a group of five users. We restricted videos to four popular consumer categories: "birthday", "wedding", "show", and "parade". Due to the uncontrolled content of consumer videos and the subjective nature of the task, such a restriction was necessary to make it possible that some common guidelines suitable for automatic summarization could be found. A total of 50 videos were collected, 32 with VGA quality from Kodak's consumer benchmark video set [10] and 18 with HD quality from Kodak's recent assets. The average length of these videos was about 80 seconds. Based on the rationale that it might be easier for users to decide what was wrong with a summary than to come up with a set of rules for an ideal summary, the study was conducted in the following way: we first generated automatic video summaries from these videos in an intuitive traditional way, and then provided these summaries to users to comment on.

The automatic summaries were constructed as follows. In the audio aspect, based on the "naive" assumption that sounds surrounding audio energy peaks were more interesting, n highest audio energy peaks (that were sufficiently separated from one another) were selected, and an m-second clip was taken, centered on each peak. These clips were ordered chronologically, which in combine gave the audio summary for the video. In the visual aspect, for each selected audio clip, we computed 5×5 grid-based color moments over image frames from the corresponding synchronized time window, and we grouped these frames into k clusters by K-means. Then the k frames that were closest to each cluster center were put together in chronological order as the visual summary. The audio and visual summaries were finally combined together into a video summary for users to analyze. In practice, we tried different combinations with $n=3,5$, $m=3,5$, and $k=3,5,10$. The responses indicated that the number of 5 clips with 3-second length each was the most favorable choice, and $k=3$ or 5 was better than 10. The rationale behind this summarization process was the importance of the audio signal in the video stream in consumer domain. As mentioned before, consumer videos usually contained single long shots, where visual appearances often did not change as dramatically as audio sounds. The importance of the audio signal was also confirmed by users in the case study where these audio-driven summaries were considered much more pleasant than alternative visual-driven ones (conducting keyframe selection first and then choosing audio clips surrounding keyframes).

Although there existed great disagreement among users, some common high-level semantic rules stood out from users' comments. In the audio aspect, audio clips where the name(s) of people were mentioned during birthday songs or wedding announcements should be included. Also, audio clips should start and end at phrase boundaries when they included speeches. In general, summaries should contain representative examples of all or many of the different semantic classes of sounds that appeared in each video. For example, if a video contained audio clips of music, speech, singing, and applause, the summary should include a reasonable mix of these sounds. In the visual aspect, clear shots of important people, such as the birthday person or the wedding couple, should be included. It was also important to avoid frames with poor qualities like blur, obstruction, or over/under exposure. If there were faces with reasonable sizes, the faces included should be clear with good quality. In addition, visual summaries should include representative examples of all or many of the different scenes that appeared in each video.

From users' responses above, we can obtain the following guidelines. First, we would like to include a varied subset of the different types of audio sounds present in a video. In general, the important audio types to include depend on video types. For the four consumer categories, four audio types are recognized as important by users: "singing", "applause", "speech", and "music". Therefore, we should include a mix of audio clips where these audio types present[1]. Second, we would like to start and end audio clips at reasonable boundary points, if not actual phrase boundaries, so that the result is not jarring to hear. Third, we should maintain the variety of audio sounds present in the video. For example, if there exist multiple stages in the audio such as different pieces of music, we need to include examples from these stages. Fourth, we would like to include keyframes with clear faces detected, and we would like to select keyframes with good overall quality. Finally, we should maintain the variety of visual scenes

[1] Intuitively, "singing" can be a subset of "music", but in the case study singing is quite distinctively (such as birthday singing) separated as an individual category. We retain this distinction.

in the video. If these exist multiple scenes we need to include keyframes from different scenes. According to these guidelines, we develop an automatic video summarization approach, as described in Section 3 and Section 4.

Other opinions from users are too high-level to follow, such as capturing people's whole names or capturing key sentences in a speech in the audio summary, and capturing the faces of key persons in the video summary. It is too difficult at the current stage to replicate such analysis in consumer videos by automatic summarization, *e.g.*, it is very hard to identify people's names from the noisy sound track or to identify the key persons from a single noisy video without additional training information.

3. OUR APPROACH: AUDIO SUMMARIZATION

3.1. Audio segmentation

As observed from the case study, it is important to automatically select start and end points of audio clips at reasonable boundaries so that the summary is not jarring to hear. To this end, we perform change detection using the *Bayesian Information Criterion* (*BIC*) [11]. This algorithm uses sliding windows at various scales to select points at which the audio on either side is better described by two separate Gaussian distributions than by a single one. Figure 1 shows an example segmentation (in black) on a spectrogram of a wedding video.

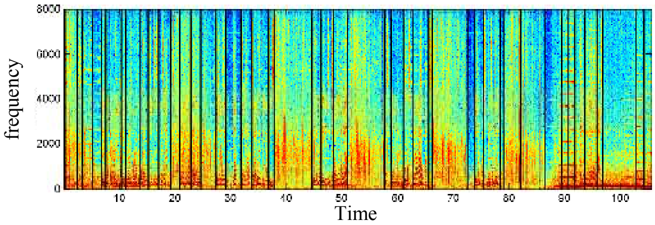


Fig. 1. Segmentation (in black) on a spectrogram of a video.

3.2. Segment classification

To identify the audio clips where the four important audio types ("singing", "applause", "speech", and "music") present, we adopt the supervised classification approach. That is, we train models to classify the automatically segmented audio clips into these four audio types. There are several caveats to this approach, most importantly the challenging conditions for audio classification in consumer-quality sound tracks, due to the differences in environment and background noise and the fact that many classes may appear concurrently. Therefore, it is necessary that the training sound tracks are also from the consumer domain with similar challenging conditions. In [12] a large-scale consumer audio set is collected, containing 2.25 hours of audio data from 203 consumer-captured videos gathered from both Kodak's assets and the YouTube video sharing site. These audio data are labeled to 10 audio classes, including the four audio types we are interested in.

The first audio features we use are the standard *Mel-frequency cepstral coefficients* (*MFCC*) [13] and their deltas, at a rate of 25 ms frames taken at 10 ms hops. Due to the poor discrimination ability of these features on "speech" and "singing", we also use some more specialized features. We calculate the 4-Hz modulation energy [14], which has been shown as a state-of-the-art feature for distinguishing "speech" from "music", and which should be more characteristic of "speech" than other classes. We also compute the harmonic coefficient [15], which is the maximum of the frame-level auto-correlation. This feature is expected to be high for "singing" and low for "speech" and other "music". All these features are concatenated together, based on which SVM classifiers are trained for each of the four audio classes.

Given a test sound track, we apply the four class models to the automatically segmented audio clips and generate four detection scores. We do not pursue hard classification because first it is very difficult to choose a threshold due to the mixing of different classes and second it is not necessary to know the exact labels. It is good enough to know which parts of a video are most likely to contain "applause", "music", "singing", or "speech". Specifically, the detection score for each class of an audio segment is the maximum score of frames in that segment under that class model.

3.3. Audio summarization

Using the detection scores of four audio classes and the audio features, an algorithm for automatic audio summarization is developed in this section. As discussed in the case study in Section 2, it is important to include a nice mix of audio clips where the four audio classes present, and to include various examples reflecting different stages of the audio sound. Addressing these issues, we first cluster the audio frames (25 ms with 10 ms hops) into N clusters according to low-level audio features, by using the K-means algorithm. Then we keep the largest M clusters, where M is determined by the percentage P of audio frames in these M clusters. For each cluster, we select $K \leq 4$ audio frames. Each audio frame corresponds to the best identified frame for each of the four audio classes. There is an option of not selecting any audio frame for some classes if the detection score for these classes are too low. After that, we locate the candidate audio segments that contains the selected K audio frames. since some candidate segments for different clusters may be the same, we have $Q \leq K \times M$ candidate audio segments in total.

The chosen candidate audio segments are then expended into audio clips with greater than L-second length each (if possible), by appending the audio segments before and after alternatively. The majority of audio frames in an appended audio segment have to be from the same cluster as the majority of audio frames in the candidate audio segment, *i.e.*, the appended audio segments should sound similar to the candidate audio segment to avoid including annoying changes. In practice we use $N = 5$, $P = 60\%$, and $L = 3$ in our experiments. The resulting list of audio clips are sorted by chronological order, to preserve the original order of the video, since users typically want to hear the summary clips in the order in which they appear in the original video. Finally the clips are concatenated together with linear fades between them. From our experiment, each audio summary usually has 3 to 4 clips and the averaged length is about 16 seconds.

4. OUR APPROACH: KEYFRAME SELECTION

For each audio clip in the audio summary, a set of representative image keyframes are selected to accompany the audio clip and generate the final video summary. As discussed before, due to the challenging conditions of consumer videos, it is difficult to identify specific objects/events from the images, and domain-specific methods relying on object/event detection [3, 4] can not be easily applied. Non-domain specific methods such as [6, 7] also perform poorly since they do not address the issues in the consumer domain, that keyframes with clear faces are important to be included and that we should choose keyframes with good overall quality.

In this section, we develop a keyframe selection approach by addressing the issues of consumer videos found in the case study. We jointly consider three aspects: the overall quality of the keyframes, quality of detected faces in the keyframes, and the visual diversity of the selected keyframes.

4.1. Image quality evaluation

There has been some recent research on characterizing consumer photographs based on image quality as well as developing predictive algorithms [16, 17]. In particular, the work in [17] provided an empirical study where a set of visual features describing various characteristics related to image quality and aesthetic values were used to generate multidimensional feature spaces, on top of which machine learning algorithms were developed to estimate images' aesthetic scales. Their study was based on a consumer photographic image collection [16], containing 450 real consumer photographic images selected from a number of different sources: Flickrr, Kodak Picture of the Day, study observers, and an archive of recently captured consumer image sets. The ground-truth aesthetic values (ranging from 0 to 100) over the 450 images were obtained through a user study from 30 observers. Regression models were built based on various visual features to estimate aesthetic values of new images.

It worth noting that there exist significant differences between consumer photographic images and image frames from consumer videos, where the later generally have much worse quality, especially technical, from low resolution and motion. Therefore, models trained over consumer photographic images using technical quality related features can not generalize well to classify image frames. So, among the best performing features reported in [17], we use the features developed by Ke *et al.* in [18], including the spatial distribution of high-frequency edges, the color distribution, the hue entropy, the blur degree, the color contrast, and the brightness (6 dimensions). Specifically, given an audio clip in the audio summary, image frames are sampled at every 0.1-sec interval, and then the above 6-dim feature is computed for each image. The regression model is then applied to generate an aesthetic score roughly measuring the image's quality. Image frames are ranked based on the scores in descending order.

4.2. Face quality evaluation

In addition to measuring the overall image quality, a face detection tool from Omron® (http://www.omron.com/) is applied to the candidate image frames and detect faces. Then for images with detected faces, we compute the color contrast and the blur degree of the most confidently detected face region. The larger value of the color contrast and the lower score of the blur degree, the better the quality for the face region. For images without any face detected, the face quality is simply set to zero.

4.3. Keyframe selection

The face quality score and the image quality score computed above are linearly combined to generate the final overall quality score for keyframe selection. The relative importance of these two quality scores depends on the type of videos. For example, for "birthday" or "wedding", detecting clear faces of the birthday person or the wedding couple may be more important than in "parade" videos. In our experiments, we just use one empirical weight setting for all four video categories.

To maintain the diversity of the selected keyframes, we extract 5×5 grid-based color moments from the image frames. From the list of candidate best-quality image frames, the ones with large-enough distances measured by the color moments feature are selected as keyframes. These keyframes are ranked in chronological order and are put together with the audio summary to generate the final video summary.

5. EXPERIMENTS

The experiments are conducted over the 50 consumer videos described in Section 2. We create two summaries for each of the 50 videos, one using our proposed approach and the other using the intuitive method described in the case study of Section 2. The average length of summaries generated by our algorithm is about 19 seconds, which is slightly longer than that of the intuitive summaries (16 seconds).

5.1. Experiment setup

The summaries are given to a total of seven consumer raters for manual evaluation. There are two runs of manual evaluation. In the first run, audio summaries (without accompanied keyframes) are provided to raters so that the evaluation is only based on the audio sound. In the second run, the entire video summaries are given to raters for final evaluation. The reason for conducting two runs is because of the observations from the case study, that users' understanding of audio content in the video varies according to whether they see the visual images or not. In each run, the raters are asked to assign a score ranging from 0 (very poor) to 10 (perfect) to each of the two summaries for each of the videos. The following are the instructions given to raters for their evaluation.

Instruction for run 1 – Please listen to the original sound track first, and then assign a score to each of the two summaries. There are some factors to consider:

1. Does the summary capture the main content of the sound track? There can be multiple interpretations of the term "content", here are three examples:
   (a) Overall semantic: if the sound track is about a wedding, can you tell from the summary that it is about wedding?
   (b) Overall diversity: if you recognize different stages (segments) in the sound track, does the summary capture these stages or most of them?
   (c) Special interests: besides the overall semantics, if some audio segments carry other semantic information that you think is important, *e.g.*, person's name mentioned in a birthday song or wedding announcement, does the summary capture them?
2. Does the summary sound pleasing? This can be very subjective. A common concern is whether you think the way the summary cuts the sound track is annoying.

Instruction for run 2 – Please view the original video first, and then assign a score to each of the two summaries. There are some factors to consider:

1. The quality of the audio summary (this is the same with the previous task in run 1)
2. The quality of visual keyframes:
   (a) Do the keyframes capture the main content of the image sequence? Some possible interpretations of "visual content" are:
   - Overall semantic: if it is a wedding video, can you tell from keyframes that it is wedding?
   - Overall visual diversity: if you recognize different scenes (subshots), do the keyframes cover all or most of them?
   - Special interests: anything you think is semantically important, do the keyframes cover them. For example, if the video has nice shots of the main person(s), such as the birthday person or the wedding couple, do the keyframes capture them?

   (b) Do the keyframes look pleasing? This can be measured from two aspects:
   - Do you think the keyframes are technically and aesthetically pleasing?
   - Do the keyframes have too much redundancy?

There exist significant disagreement among the raters in terms of the absolute scores assigned to individual summaries. Some raters are very strict and assign low scores to most of the summaries, while some others are more forgiving and assign much higher scores to many summaries. Tables 1 (a) and (b) give the overall scores averaged across different videos and cross different raters for run 1 and run 2, respectively, where the number in the parenthesis is the standard deviation. The averaged results tell us that our approach is rated better than the intuitive method in general. However, due to the disagreement among the raters, the per-video rating scores are very noisy for us to analyze.

Table 1. Rating scores for different runs, averaged across different videos and different raters

(a) run 1

Intuitive summary	Our Approach
6.7375 (±0.9732)	7.465 (±1.2175)

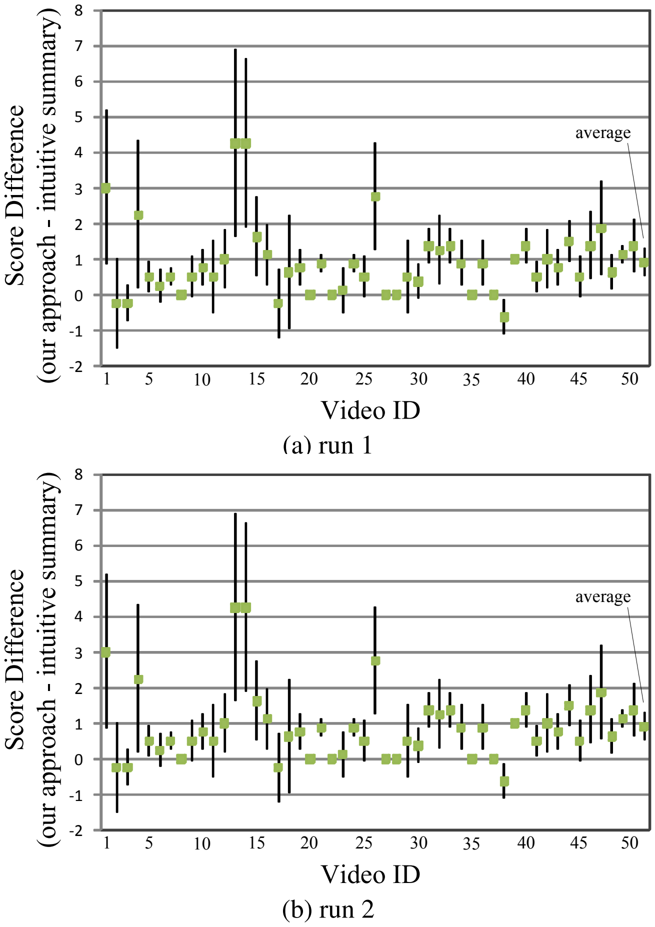
(b) run 2

Intuitive summary	Our Approach
?? (±??)	?? (±??)

To accommodate the issue of disagreement among the raters, we compute the rating differences between our summaries and the intuitive summaries and show the per-video results (averaged across different raters) for run 1 and run 2 in Figures 2 (a) and (b), respectively. The green squares are averaged score differences and the black vertical lines are the standard deviations. The figures clearly show the advantage of our approach that over most videos, in both run 1 and run 2, most raters agree that it outperforms the intuitive method. In run 1, by listening to the audio summaries alone, out of the 50 videos, the intuitive summaries are better than ours over 4 videos, where the general complaint is that sentences are cut off in our summaries. One typical case is that the video has several short and semantically not meaningful sentences, and the intuitive summary happens to capture one or two of such short sentences. Our method, on the other hand, deliberately finds more speech segments to include in the summary, and ends up with some broken sentences. When we combine visual signal and audio signal together, there are less confusion about the content of the videos, and the raters agree with each other more. Almost all of our final video summaries are rated better than the intuitive video summaries where ?? summaries have significant improvements, *i.e.*, improved by more than ?? points. Figures 3 (a) and (b) give some example keyframes selected by the intuitive method and our approach, respectively. This "wedding" video records the bridesmaid procession. The intuitive summary only captures the first part where there is loud background music, while our summary includes three segments representing different stages of the whole process. Especially, when the camera focuses on the bridesmaids with close-up shots, there exists large camera motion. Through assessing the face quality and the overall image quality, our method is able to pick out clear keyframes.

6. CONCLUSION

We studied automatic video summarization in the consumer domain by analyzing the visual and audio content of the video stream. A case study was conducted to obtain a set of consumer-oriented guidelines that reflected the high-level semantic rules of generating good summaries of consumer videos under challenging conditions in both image sequences and audio sound tracks. Following the guidelines, an automatic consumer video summarization system was developed, which took into account the following aspects to generate improved video summaries: audio segmentation and classifica-



(a) run 1

(b) run 2

Fig. 2. Per-video rating score differences (score of our approach minus that of the intuitive summary) for different runs.

tion, audio diversity, visual diversity, face quality, and overall image quality. Evaluation from consumer raters confirmed that our approach better accommodated consumer needs than the traditional method.

7. REFERENCES

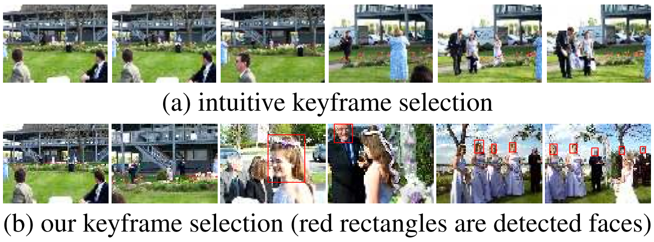
[1] A.G. Money and H. Agius, "Video summarisation: A conceptual framework and survey of the state of the art," *Journal of Visual Communication and Image Representation*, 19(2008):121–143, 2008.

[2] D. Tjondronegoro, Y. Chen, and B. Pham, "Highlights for more complete sports video summarization," *IEEE Trans. Multimedia*, 11(4):22–37, 2004.

[3] F.N. Bezerra and E. Lima, "Low cost soccer video summaries based on visual rhythm," *Proc. ACM Multimedia*, pp. 71–77, 2006.

[4] Y. Song, G. Marchionini, and C.Y. Oh, "What are the most eye-catching and ear-catching features in the video? implications for video summarization," *ACM WWW*, 2010, Raleigh, North Carolina.

[5] I. Otsuka and *et al.*, "A highlight scene detection and video summarization system using audio feature for a personal video recorder," *IEEE Trans. Consumer Electronics*, 51(1):112–116, 2005.



(a) intuitive keyframe selection

(b) our keyframe selection (red rectangles are detected faces)

Fig. 3. Examples of keyframe selection.

[6] Z. Cernekova, I. Pitas, and C. Nikou, "Information theory-based shot cut/fade detection and video summarization," *IEEE Trans. CSVT*, 16(1):82–91, 2006.

[7] Z. Li, G.M. Schuster, and A.K. Katsaggelos, "Minmax optimal video summarization," *IEEE Trans. CSVT*, 15(10):1245–1256, 2005.

[8] L. Agnihotri and *et al.*, "Study on requirement specifications for personalized multimedia summarization," *Proc. IEEE ICME*, pp. 757–760, 2003.

[9] C. Forlines, K.A. Peker, and A. Divakaran, "Subjective assessment of consumer video summarization," *Proc. SPIE Conf. Multimedia Content Analysis, Management and Retrieval*, vol. 6073, pp. 170–177, 2006.

[10] A.C. Loui and *et al.*, "Kodak consumer video benchmark data set: concept definition and annotation," *ACM Workshop on MIR*, 2007.

[11] S. Chen and P. Gopalakrishnan, "Speaker, environment and channel change detection and clustering via the bayesian information criterion," *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, pp. 127–132, 1998.

[12] C. Parker, "An empirical study of feature extraction methods for audio classification," *Proc. IEEE ICPR*, pp. 4593–4596, 2010.

[13] P. Mermelstein, "Distance measures for speech recognition, psychological and instrumental," *Pattern Recognition and Artificial Intelligence*, pp. 374–388, 1976.

[14] E. Scheirer and M. Slaney, "Construction and evaluation of a robust multifeature speech/music discriminator," *Proc. IEEE ICASSP*, pp. 1331–1334, 1997.

[15] W. Chou and L. Gi, "Robust singing detection in speech/music discriminator design," *Proc. IEEE ICASSP*, pp. 865–868, 2001.

[16] C.D. Cerosaletti and A.C. Loui, "Measuring the perceived aesthetic quality of photographic images," *IEEE QOMEX*, 2009.

[17] W. Jiang, A. Loui, and C. Cerosaletti, "Automatic aesthetic value assessment in photographic images," *Proc. IEEE ICME*, pp. 920–925, 2010.

[18] Y. Ke, X. Tang, and F. Jing, "The design of high-level features for photo quality assessment," *Proc. IEEE CVPR*, vol. 1, pp. 419–426, 2006.