VIDEO QUALITY CLASSSIFICATION BASED HOME VIDEO SEGMENTATION

Si Wu[1*], *Yu-Fei Ma*[2], *Hong-Jiang Zhang*[2]

[1]Department of Computer Science, Jinan University, Guangdong, China, 510632
[2]Microsoft Research Asia, Beijing, China, 100080

ABSTRACT

Home videos often have some abnormal camera motions, such as camera shaking and irregular camera motions, which cause the degradation of visual quality. To remove bad quality segments and automatic stabilize shaky ones are necessary steps for home video archiving. In this paper, we proposed a novel segmentation algorithm for home video based on video quality classification. According to three important properties of motion, *Speed*, *Direction*, and *Acceleration*, the effects caused by camera motion are classified into four categories: *Blurred*, *Shaky*, *Inconsistent* and *Stable* using support vector machines (SVMs). Based on the classification, a multi-scale sliding window is employed to parse video sequence into different segments along time axis, and each of these segments is labeled as one of camera motion effects. The effectiveness of the proposed approach has been validated by extensive experiments.

1. INTRODUCTION

Parsing videos according to video quality is a desirable pre-processing for home video archiving. The bad quality segments, such as too blurred or shaking clips, should be removed or recovered, because these clips usually annoy the viewers. On the other hand, the different type of low quality segments should be managed in different way. For example, video stabilization is an important quality enhancement method, such as [1-3]. However, if a stable clip is processed by stabilization algorithm, the visual quality may be degraded. While the blurred frames caused by too fast camera motion maybe result in inaccurate stabilization. Therefore, we need to segment home video according to video quality classification before other processes, such as quality enhancement and archiving.

Traditional video parsing methods [4, 5] take shots as the basic elements of video, and parse video according to the feature change between adjacent frames. However, in home videos, the duration of physical shot (formed by camera start and stop) may be very long, such as 3-5 minutes, and the camera may change its status several times within one physical shot. So the physical shot in home video is meaningless for viewers. In general, home video can be segmented according to any ad hoc rules for different archiving purposes. However, before that, we often need segment home video according to video quality to facilitate home video quality enhancement, such as stabilization.

Some researchers have taken simple video quality based segmentation into account. For example, camera shaking is distinguished from zooming and panning based on the motion vectors in MPEG stream in [6]. Camera shaking segments are detected based on the discontinuity of target region motion trajectory in [7]. However, it is difficult to select the target region automatically. Although these methods have studied the detection of shaky video segments, they haven't considered other type of effects caused by irregular camera motion. In [8], the camera motion is classified into five categories. In their implementation, camera shaking is detected at first, then, regular camera motion is classified into panning, titling and zooming. Finally, camera vibration, actual still and irregular motion are determined. However, the video segmentation method in [8] is mainly based on the statistical classification of camera motion, which doesn't take the visual effects caused by different camera motions into account. So, the method in [8] doesn't suit the purpose of quality enhancement. The Hitchcock system in [9] classifies the amount of camera motion into five subjective categories of good, acceptable, tolerable, or unacceptable according to the amount of camera's horizontal and vertical pan and the amount of zoom, however, its purpose is to automatically detect suitable clips in the raw video material for the purpose of video editing.

In this paper, we propose a new home video segmentation method based on video quality classification. According to three primary properties of motion: *Speed*, *Direction*, and *Acceleration*, the effects caused by camera motion are classified into four categories: *Blurred, Shaky, Inconsistent and Stable*. Based on this definition, the statistical learning method SVMs are adopted as classifier

[*] The work presented in this paper was carried out in Microsoft Research Asia

and a multi-scale sliding window is used to parse video sequence into segments along time axis. The *Blurred* segments are detected and removed first. Then, the rest parts of video are further segmented by multi-scale sliding window scheme. Finally, a smoothing routine is carried out to merge those very short segments and determine the accurate boundaries of segments. In this way, different approaches to quality enhancement can be conducted on corresponding segments, such as deblur, stabilization [1-3], etc.

The rest paper is presented as follows. In Section 2, a new video quality classification approach is introduced first. In Section 3, video quality based segmentation method is discussed in detail. The evaluation results are reported in Section 4. Finally, Section 5 concludes the paper.

2. VIDEO QAULITY CLASSIFICATION

As camera motion is an important factor for video quality, we define video quality according to three primary properties of camera motion, say, *Speed*, *Direction* and *Acceleration*. Here, SVMs are employed as classifier due to its good general performance.

2.1. Definition of Motion Effects

As shown in Fig.1, the classification process is regarded as a decision tree.

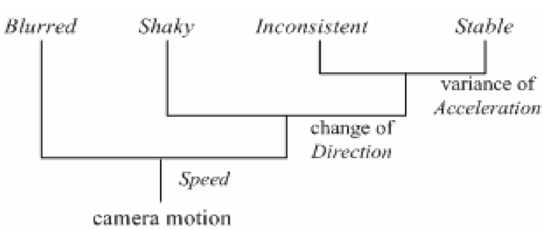

Fig. 1 Classification of Motion Effects

1) *Blurred*: If the speed of camera motion is high, the captured frames will be blurred. As the *Blurred* segments are difficult to be recovered and usually meaningless for viewers, these video segments can be discarded.
2) *Shaky*: If the speed is normal, but the direction of camera motion changes frequently, namely, the camera moves back and forth repeatedly, the captured videos are regarded as shaky. Usually, the *Shaky* segments can be stabilized by video stabilization methods.
3) *Inconsistent*: If speed is normal and direction is consistent, but the accelerations of camera motion in consecutive frames are uneven, say the variance of acceleration is large, the captured videos are inconsistent. The *Inconsistent* segments can also be improved by inserting frames.
4) *Stable*: The normal camera motion with rare direction changes and even accelerations is defined as stable motion. The *Stable* segments should not be touched by any processing to keep the original video quality.

2.2. Classification with SVMs

SVMs [10, 11], as a well-known non-linear classifier, have good generalization performance. Hence, SVMs are employed to classify motion effects defined in Section 2.1.

We extract feature vectors based on the three basic motion properties, say, *Speed*, *Direction* and *Acceleration*. Let S^x, S^y denote the average speed of translational motion on x axes and y axes, A^x, A^y denote the average acceleration on x axes and y axes, V^x, V^y denoted the variance of acceleration on x axes and y axes, and D^x, D^y be the frequency of direction change on x axes and y axes, respectively. The feature vector of a video clip in feature space is denoted by an 8 dimensional vector $v=(S^x, S^y, A^x, A^y, V^x, V^y, D^x, D^y)$, where

$$S^x = \operatorname*{avg}_{i=1}^{N-1}\left(\left|T^x_{i,i-1}\right|\right),\ S^y = \operatorname*{avg}_{i=1}^{N-1}\left(\left|T^y_{i,i-1}\right|\right) \tag{1}$$

$$A^x = \operatorname*{avg}_{i=1}^{N-2}\left(\left|T^x_{i,i-1} - T^x_{i+1,i}\right|\right),\ A^y = \operatorname*{avg}_{i=1}^{N-2}\left(\left|T^y_{i,i-1} - T^y_{i+1,i}\right|\right) \tag{2}$$

$$V^x = \operatorname*{var}_{i=1}^{N-2}\left(\left|T^x_{i,i-1} - T^x_{i+1,i}\right|\right),\ V^y = \operatorname*{var}_{i=1}^{N-2}\left(\left|T^y_{i,i-1} - T^y_{i+1,i}\right|\right) \tag{3}$$

$$D^x = \operatorname*{avg}_{i=1}^{N-2}\left(FD\left(T^x_{i,i-1}, T^x_{i+1,i}\right)\right),\ D^y = \operatorname*{avg}_{i=1}^{N-2}\left(FD\left(T^y_{i,i-1}, T^y_{i+1,i}\right)\right) \tag{4}$$

$$FD(T_1, T_2) = \begin{cases} 1 & if\ \mathrm{sgn}[T_1] = \mathrm{sgn}[T_2] \\ 0 & else \end{cases} \tag{5}$$

where, $T^x_{i,i-1}$ and $T^y_{i,i-1}$ are the translation between two adjacent frames I_i and I_{i-1} on x axes and y axes respectively, which are calculated by the camera motion estimation method proposed in [12].

In this four-class problem, we use one-against-all scheme to train 4 classifiers, separately. For each class, the positive training samples are samples in this class, while the negative training samples are all the other samples. In this work, we use $L=\{l_1, l_2, l_3, l_4\}$to stand for the whole set of motion effects, where l_1, l_2, l_3, l_4 denote *Blurred, Shaky, Inconsistent and Stable* respectively. Given a motion effect $l \in L$, the training sample set is $E=\{(v_i, u_i) \mid i=1, \ldots, n\}$, where v_i is the feature vector, $u_i \in \{+1, -1\}$, if v_i belongs to l, u_i=+1, otherwise u_i=-1. After the training of SVM, we can get a decision function f. For a given sample v, we first compute $z=\Phi(v)$, where Φ is the feature map, for example, the radial basis function can be adopted as the kernel function to implement the feature map. Then we compute the decision function $f(z)$. If $f(z)$=1, then v belongs to class l, otherwise, v is not in class l.

Therefore, for a given video clip c, it is classified by

$$F(c) = l_i,\ i = \arg\max_{i=1}^{4}(f_i(c)) \tag{6}$$

3. VIDEO QUALITY BASED SEGMENTATION

With the classification scheme defined in Section 2, each frame is labeled as one of motion effects. However, for automatic home video quality enhancement, to determine the motion effect type of a segment is more meaningful, because the video quality enhancement cannot be conducted on individual frames. So, how to segment video sequence and determine the accurate boundaries is the other important issue.

Based on the classification results, a sliding window is adopted for video segmentation. As shown in Fig.2, the boundary detection is performed in 3 steps. We first detect and discard *Blurred* segments from original video clip (Step 1). In this way, the video sequence is divided into a number of relative long segments. Secondly, each relative long segment is further segmented into sub-segments by multi-scale sliding window scheme, and labeled by *Stable* or *Shaky* or *Inconsistent* (Step 2). At last, those very short and noisy sub-segments are merged with there neighbors to smooth the detecting results (Step 3).

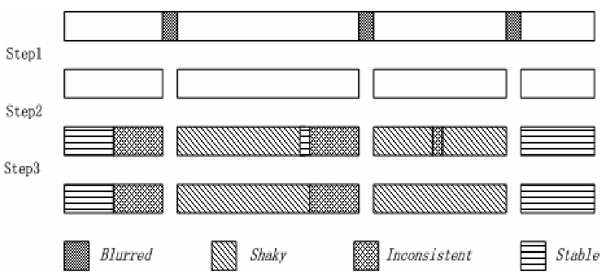

Fig. 2 Automatic Boundary Detection

As blurring motion usually lasts for a very short time in home videos, a short sliding window with length *len* is used, that is

$$if\ \ F(c(I_j, len)) = l_1\ \ then\ \ L(I_j) = l_1 \qquad (7)$$

where $c(I_j, len)$ represents a part of the clip with length *len* centered at frame I_j, $L(I_j)$ is the motion effect label of I_j. After the blurred frames are detected, the original video clip is divided into a number of segments. The segments with length less than *len* are regarded as *Blurred* segments and discarded because frames between too frequent blurring motions may be regarded as insignificant. In experiments, we set *len*=5. The relative long segments are then further segmented.

As it is difficult to get accurate boundary using a single sliding window, a multi-scale sliding window scheme is adopted considering motions usually last for different durations. Let $len_1,\ldots, len_k$ be the length of k sliding windows, I_{Bi} and I_{Bi+1} are two adjacent blurred frames, for any frame I_j, $(B_i<j<B_{i+1})$, $F(c(I_j, len_i))$, i=1, …, k denote the detected results. The motion effect type of I_j is determined according to the priority of motion effect type in the decision tree as shown in Fig. 1. That is, the motion type with the highest priority in the detected results $F(c(I_j, len_i))$ is given as the label of I_j.

With multi-scale sliding window, the relative long segment is partitioned into a series of sub-segments. Each sub-segment is composed of consecutive frames with the same motion effect type. However, some very short sub-segments may appear between long sub-segments, which lead to over segmentation. For example, after the step 2, in Fig. 2, a very short inconsistent sub-segment appears between two long shaky sub-segments. The over segmentation interrupt the consistency of the segmentation results. Therefore, a smoothing routine is needed to remove these noisy short sub-segments. For the purpose of smoothing, say, merging over short sub-segments into neighboring sub-segments, a length threshold T_{len} is utilized to determine which sub-segment is over short. If the two neighboring sub-segments of the over short sub-segment have the same motion effect, the over short sub-segment and the two neighboring sub-segments are merged together. If the two neighboring sub-segments have different motion effects, the over short sub-segment is merged into the neighboring sub-segment with higher priority effect as defined in classification tree. In this manner, the boundaries of segments with the same motion effect are determined. In experiments, setting T_{len}=5 may lead to a satisfactory results.

4. EXPERIMENTS

As the proposed segmentation approach has two tasks, classification and boundary detection, we evaluate the performance of two tasks separately. The classification performance is easily evaluated by conventional objective methods. However, it is difficult to assess the accuracy of boundary detection objectively, because the accurate boundaries of motion effects cannot precisely determined, even by human. Therefore, we carried out a subjective experiment for boundary detection evaluation. We invited 5 subjects to assess the detected segments by assigning a satisfactory score to each segment, respectively.

In classification evaluation, total 659 video clips are used as training and testing samples. These clips are extracted from indoor and outdoor home videos, and labeled manually based on subjective perception, in which there are 160 *Blurred* clips, 169 *Shaky* clips, 169 *Inconsistent* clips and 161 *Stable* clips. In order to effectively utilize small scale sample set, a five-step random test is performed. Three-eighth samples in the sample set are randomly selected to construct testing set first. For constructing training set, one-fifth of the rest samples are randomly selected and accumulatively added into the training set in each step without repetition. After 50 times random tests, the average precisions and recall of classification at each step are listed in Table 1 and Table 2, respectively. From the experimental results, we can see that our classifiers can obtain satisfactory classification results as long as the training set is larger than the testing set. For example, at the fourth step, the average precision and recall are above 0.95. At the fifth step, the average precision and recall reach the highest value.

Table 1. Average Precision of 50 Random Tests

Type	step 1	step 2	step 3	step 4	step 5
Blurred	0.994	0.995	0.996	0.998	0.999
Shaky	0.867	0.930	0.963	0.974	0.978
Inconsistent	0.876	0.926	0.950	0.958	0.960
Stable	0.915	0.947	0.964	0.977	0.979

Table 2. Average Recall of 50 Random Tests

Type	step 1	step 2	step 3	step 4	step 5
Blurred	0.944	0.968	0.980	0.985	0.986
Shaky	0.890	0.919	0.942	0.953	0.957
Inconsistent	0.891	0.951	0.977	0.991	0.995
Stable	0.906	0.955	0.973	0.978	0.978

For boundary detection evaluation, 34 video clips, about one hour, are used as testing data. Five subjects are invited to rank the accuracy of detection with four levels: good, acceptable, bad and incorrect. The corresponding scores of the five ranks are: 1.0, 0.7, 0.4 and 0.0, respectively. Table 3. gives the evaluation results. The scores of *Blurred, Shaky, Inconsistent* segments are high, while the scores of *Stable* segments is relative lower. It is because some *Shaky* or *Inconsistent* segments are incorrectly detected as Stable.

Table 3. Subjective Evaluation Results

segment type	*Blurred*	*Shaky*	*Inconsistent*	*Stable*
subject1	1.0	0.96305	0.99053	0.88553
subject2	1.0	0.93625	0.98343	0.83900
subject3	0.97014	0.82598	0.87108	0.84708
subject4	0.88209	0.96887	0.89533	0.82068
subject5	0.85022	0.83603	0.79862	0.82373
average	0.94049	0.90604	0.90779	0.84320

With the video quality classification based segmentation, the home video quality can be separately improved according to different motion effects. The *Stable* ones are kept as the original and the *Blurred* ones may be discarded. The *Shaky* ones can be stabilized by video stabilization routines, such as [1-3]. While the *Inconsistent* ones can be enhanced by interpolation of frames or other more complex methods.

5. CONCLUSION

In this paper, an automatic home video segmentation algorithm based on motion effect classification is proposed. According to three primary motion properties: *Speed*, *Direction* and *Acceleration*, the motion effects are classified into four categories: *Blurred, Shaky, Inconsistent* and *Stable.* In this work, SVMs are adopted as the classifier. Based on the classification results of each frame, a sliding window scheme is used to detect boundaries of motion effects. The effectiveness and robustness of the classification and boundary detection have been validated by an objective and a subjective experiment, respectively.

6. REFERENCES

[1] Z. Zhu et al, "Camera stabilization based on 2.5D motion model estimation and inertial motion filtering," *IEEE International Conference on Intelligent Vehicles*, Stuttgart, Germany, pp. 329-334, 1998.

[2] A. Litvin, J. Konrad and W. C. Karl, "Probabilistic video stabilization using kalman filtering and mosaicking," *Proceedings of SPIE Conference on Electronic Imaging, Image and Video Communications and Proc.*, Santa Clara, CA, vol. 5022, pp. 663-674, 2003.

[3] S. Erturk, "Translation, rotation and scale stabilisation of image sequences," *Electronics Letters*, vol. 39(17), pp. 1245-1246, 2003.

[4] R. Lienhart, "Comparison of automatic shot boundary detection algorithms", *Proceedings of SPIE Storage and Retrieval for Image and Video Databases VII*, vol. 3656, pp. 290-301, 1999.

[5] A. Hanjalic, "Shot-boundary detection: unraveled and resolved?" *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 12(2), pp. 90-105, February 2002.

[6] K. Dobshi, A. Kodate, H. Tominaga, "Camera working parameter extraction for constructing video considering camera shake," Proceedings of International Conference on Image Processing, vol. 3, pp. 382-385, 2001.

[7] W.Q. Yan and M.S. Kankanhalli, "Detection and removal of lighting & shaking artifacts in home videos," In: *Proc. ACM Multimedia* 2002, Juan Les Pins, France, pp.107-116, 2002.

[8] Dong-Jun Lan, Yu-Fei Ma, Hong-Jiang Zhang, "A systemic framework of camera motion analysis for home video," *Proceedings. of IEEE International Conference on Image Processing*, vol. 1, pp. I-289-I-292, 2003.

[9] Girgensohn, A., et al, "A semi-automatic approach to home video editing." *CHILetters: Symposium on User Interface Software and Technology (UIST)*, vol. 2(2), pp. 81–89, 2000.

[10] C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Data mining and knowledge discovery*, vol. 2, pp.121-167. 1998.

[11] J. Weston and C. Watkins, "Multi-class support vector machines," Tech. Rep. CSD-TR-98-04, Royal Holloway, university of London, 1998.

[12] J. Konrad, F. Dufaux, "Improved global motion estimation for N3," *Meeting of ISO/IEC/SC29/WG11*, No. MPEG97/M3096, San Jose, 1998.