2014 International Joint Conference on Neural Networks (IJCNN)
July 6-11, 2014, Beijing, China

User-Generated-Video Summarization using Sparse Modelling

Yulong Liu, Huaping Liu, Yunhui Liu, Fuchun Sun

***Abstract*— A novel key-frame extraction method is proposed in this paper. Our method focused on user-generated-videos which were captured by smartphones or tablets or other smart devices which can record acceleration values and orientation values during video capturing. Our method use Dissimilarity-based Sparse Modeling Representative Selection(DSMRS) on orientation information to extract key-frames instead of visual features used by traditional key-frame extraction methods. Acceleration value is used in our method to exclude outliers.**

I. Introduction

SMART devices like smartphones, tablets and smart video cameras are getting popular in recent years. As a result, people are recording more videos than ever before. More than 30% of all videos are captured by smartphones in recent years according to statistical data. Such large amount of videos makes it difficult for users to find a specific video from their video library due to the amount of the videos and the small screens of the smart devices [1][2]. Therefore, how to efficiently summarize videos becomes a hot research area in the past two decades. To solve this problem, key-frame extraction ,which is defined as the most representative frame-set which can be used to describe the whole video's content precisely and concisely, is proposed. Hence, automatic key-frame extraction becomes a crucial problem to organize and retrieve videos which enables user to understand the whole content of a video at one glace.

Automatic key-frame extraction has been studied in the past two decades and a batch of methods have been proposed by researchers. The most simple key-frame extraction algorithm is Evenly Spaced Key-Frame(ESKF). The pros of this method are that it is fast, stable, and easy to change the amount of key-frames. However, it has huge cons: this method does not understand the video, so there is high probability that this method choose outliers or redundant frames. Although there are a lot of other key-frame extraction methods, most of which are focused on structured videos like sports games[3] or wildlife videos[4].

However, our paper focused on user-generated-videos which show greater diversity[5] and more difficult because of the following reasons:

1) User-generated-videos are captured from various scenes such as office, road, playground,schools etc., and have dissimilar interests. Therefore we have no prior knowledge of these videos.
2) Most user-generated-videos contain intentional camera motion (users are following something, or they changed their focus) and unintentional camera motion (hand shake). Meaningless frames may be introduced due to these motion.

Because these user-generated-videos are not well structured, a smart solution of key-frame extraction is to guess the intention of the user who captured the video. Typically user's intention is inferred from camera motion, which could be computed from the video frames by various methods [5]–[8]. Though the camera motion could be computed from video frames, the computational complexity is unbearable high, and the computed result is not accurate, especially when there are too many moving objects in the video.

User-generated-videos by smartphones has their own advantages: we can record not only visual information when capturing video. Smartphoness are equipped with a lot of sensors, including but not restricted to, accelerometer sensor, gravity sensor, geo-magnetic sensor, orientation sensor, light sensor, temperature, humidity, proximity, pressure, etc[9]. we reckon that orientation sensor and accelerometer sensor can be utilized in key-frame extraction.

Orientation sensor can describe which direction the device is orientated when capturing a specific frame. Similar orientation values indicates the corresponding frames are captured toward similar scene. So in this paper, we use orientation to infer the user's intention:

1) If some frames share similar orientation value, we could make the hypothesis that these frames are focusing on the same thing, in another word, they belong to the same scene, and one frame from those frames could represent these frames well. On contrast, if few frames' orientation values locate near a specific frame, we can assume this frame is captured during a camera motion and could be recognized as outlier.
2) If user captures the video toward a similar orientation for a long time, we can believe that this scene must contains something important to the user.

However, there are some outliers cannot be attenuate by using orientation value only, for example, if user switched from one scene to another, the frames near the start point and the end point have similar orientation as previous scene and trail scene respectively. But we can easily find these outliers by using acceleration value because these frames are likely to have larger acceleration value.

In this paper, we developed a key-frame extraction method for user-generated-videos, the main contributions of this pa-

All of the authors are with Department of Computer Science and Technology, Tsinghua University, State Key Laboratory of Intelligent Technology and Systems, TNLIST, Beijing P.R. China. hpliu@tsinghua.edu.cn.

This work is jointly supported by the National Key Project for Basic Research of China (2013CB329403), the National Natural Science Foundation of China (Grants Nos: 91120011 and 61210013), Tsinghua Self-innovation Project (Grant No: 20111081111) and Tsinghua University Initiative Scientific Research Program (Grant No: 20131089295).

978-1-4799-1484-5/14/$31.00 ©2014 IEEE

per are as follows: (1) Smart device sensor data is used to extract key-frame instead of visual feature. (2) A dissimilarity-based sparse modeling representative selection algorithm is used to extract key-frame. (3) We constructed a new dataset containing visual feature, acceleration data and orientation data. Also we reported extensive experimental results on that dataset.

The rest of this paper is organized as following: Section II give a introduction to smartphone's sensors and how we utilize them. Section III describes the detail around the proposed optimization model. Section IV describes our dataset used in the paper and illustrates the experiment detail and result. Section V gives the conclusions.

II. Smartphone Sensors

A. Sensor Introduction

Orientation value is defined as $\mathbf{o} = [o_x, o_y, o_z]$, where $o_x \in (-180°, 180°]$ is the degree of angle the device rotated around the x-axis. $o_y \in (-180°, 180°]$ is the degree of angle rotated around the y-axis, and $o_z \in (-180°, 180°]$ is the degree of angle rotated around the z-axis. Note: the coordinate system used in smartphone is different from the world coordinate system. Smartphone's coordinate system is shown is Fig.1

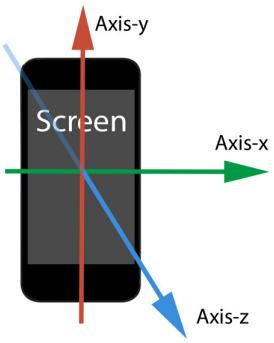

Fig. 1. The coordinate system used in smartphone. X-axis is horizontal, along the short side of the device and point to right; y-axis is horizontal, along the long side of the device and point to up; z-axis is vertical, points out of the device's front face of the device's screen.

Raw acceleration value is the acceleration value on x-axis, y-axis and z-axis. But the value of raw acceleration is influenced by gravity, so we use linear acceleration instead which removes gravity in this paper. In the following of the paper, acceleration means linear acceleration. Acceleration value is defined as $\boldsymbol{\alpha} = [\alpha_x, \alpha_y, \alpha_z]^T$, which coorespond to the linear acceleration value on x-axis, y-axis and z-axis.

B. Sensor Data

To illustrate how to utilize acceleration value and orientation value, consider the example shown in the Fig.2 which describes the orientation value's distribution of one video from our dataset. This video contains about 821 frames, but we downsample the video to 82 frames to make the figure shows more clearly. In this example, the orientation values are apparently belong to two classes, which represent the smartphone have two main orientation direction when capturing this video. So we can reasonably refer that this video contains two main scenes. Also there are some outliers caused by scene change exist in that video, outliers are marked as blue points and meaningful points are marked as red.

For this sample, we can easily separate the two scenes because they are significantly different in orientation value. Also it is not different to classify the outliers like outlier C or outlier D because there are few points have similar orientation as them. However, to classify outlier A is not that easy. This outlier locates at the start of the scene change, so its orientation value is close to useful data like meaningful frame B. Therefore, we cannot separate outlier A and meaningful frame B by only using orientation value.

Acceleration value is introduced in our paper to handle this problem. Because during the scene change period, smartphone's acceleration value is likely to be considerably larger than acceleration value under stable status. This phenomenon is clearly shown in the Fig.2's acceleration curve. The left yellow labeled interval of acceleration curve is plain, which correspond to meaningful frame B and its following 20 frames, and the right yellow labeled interval is not stable, which correspond to outlier A and its following 20 frames. Hence, we can use acceleration value to easily separate outliers like outlier A which could not be handle by using only orientation value.

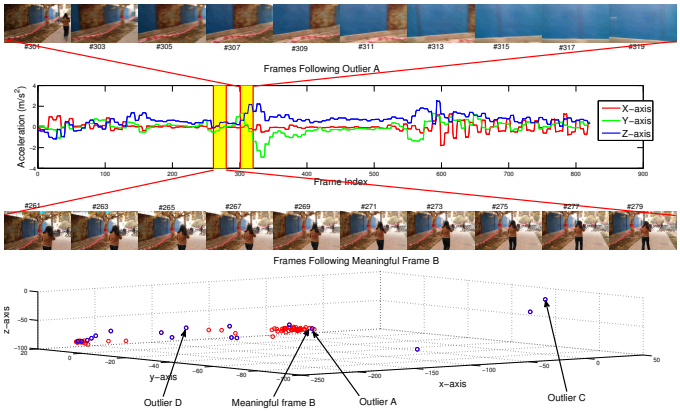

Fig. 2. The first row of this figure is 10 frames right following frame outlier A. The next row is the acceleration value on x-axis, y-axis and z-axis during the video. The third row is 10 frames right after meaningful frame B. The last row of this figure is the distribution of orientation of one sample video.

III. Model Description

Key-frames are a batch of frames extracted from one video and could represent the video's content completely and without outliers and overlaps. Because we used orientation information to extract key-frame, the extraction is equivalent to find the representative orientation values to express the whole orientation matrix well.

A naive solution using DSMRS[10] algorithm is like following: We assume the orientation value matrix is $\mathbf{O} = [\mathbf{o}_1^T, \mathbf{o}_2^T, ..., \mathbf{o}_N^T]^T$ where $\mathbf{o}_i^T$ is a 3×1 vector represents the orientation value for frame i, and N denotes the total frame count of the video. We used Euclidean distance to measure the difference between two orientation value in this paper. So we can compute a distance matrix $\mathbf{D}$:

$$\mathbf{D} = \begin{bmatrix} \mathbf{d}_1^T \\ \mathbf{d}_2^T \\ \vdots \\ \mathbf{d}_N^T \end{bmatrix} = \begin{bmatrix} d_{11} & d_{12} & \cdots & d_{1N} \\ d_{21} & d_{22} & \cdots & d_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ d_{N1} & d_{N2} & \cdots & d_{NN} \end{bmatrix} \tag{1}$$

where $d_{ij} = \|o_i - o_j\|_2$. Reference [11] introduces variables z_{ij} to indicate the possibility that sample i can represent sample j. So we have another matrix:

$$\mathbf{Z} = \begin{bmatrix} \mathbf{z}_1^T \\ \mathbf{z}_2^T \\ \vdots \\ \mathbf{z}_N^T \end{bmatrix} = \begin{bmatrix} z_{11} & z_{12} & \cdots & z_{1N} \\ z_{21} & z_{22} & \cdots & z_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ z_{N1} & z_{N2} & \cdots & z_{NN} \end{bmatrix} \tag{2}$$

There are two key aspects lie in key-frame extraction. From one side, we have to find the most representative frames, from another side, the amount of key-frames have to be controlled. In order to find the most representative frames which could illustrate the main content of the whole frame-set, we have to optimize the total encoding cost of key-frames represent all frames, the total encoding cost is

$$\sum_{i=1}^{N}\sum_{j=1}^{N} d_{ij} z_{ij} = tr(\mathbf{D}^T \mathbf{Z}) \tag{3}$$

where tr is defined as the trace of a matrix.

The second aspect is to control the number of key-frames. In the matrix $\mathbf{Z}$, if i is a representative frame of frame j, then $z_{ij} \neq 0$, otherwise, $z_{ij} = 0$. The larger the value of z_{ij} is, the better frame i can represent frame j. Hence, if frame i cannot represent any of other frames, then $\|\mathbf{z}_i\|_2 = 0$. So an ideal $\mathbf{Z}$ should contains a few non-zero rows, because we hope to represent the whole video using a few frames; and each non-zeros rows should have many nonzero entries because we want to maximize each frame's representative capability. In another word, we want this matrix $\mathbf{Z}$ to be a row-sparse matrix. As a result, we need to solve such a optimization problem:

$$\min_{\mathbf{Z}} \|\mathbf{Z}\|_{2,0} + tr(\mathbf{D}^T \mathbf{Z}) \quad s.t. \mathbf{Z} \geq 0, 1^T \mathbf{Z} = 1^T \tag{4}$$

where $l_{2,0}$-norm is defined as

$$\|\mathbf{Z}\|_{2,0} = \sum_{i=1}^{N} \mathcal{I}(\sqrt{\sum_{j=1}^{N} z_{ij}^2}) \tag{5}$$

where function $\mathcal{I}$ is

$$\mathcal{I}(\alpha) = \begin{cases} 0 & \text{if} \quad \alpha = 0 \\ 1 & \text{otherwise} \end{cases} \tag{6}$$

However, the function in (4) is of little practical use, because the optimization problem is NP hard which cannot be solved in polynomial time. A natural and equivalent alternative solution is to use $l_{2,1}$-norm which is defined as

$$\|\mathbf{Z}\|_{2,1} = \sum_{i=1}^{N} \sqrt{\sum_{j=1}^{N} z_{ij}^2} \tag{7}$$

instead of the $l_{2,0}$-norm. Then the following convex optimization problem is like follows:

$$\min_{\mathbf{Z}} \lambda \|\mathbf{Z}\|_{2,1} + tr(\mathbf{D}^T \mathbf{Z}) \quad s.t. \mathbf{Z} \geq 0, 1^T \mathbf{Z} = 1^T \tag{8}$$

The parameter λ in the objective function is used to constraint the amount of key-frames. The larger λ is, the fewer key-frames will be selected, and vice versa.

This objective function is good enough to choose the representative frames, but it has its drawbacks. So we improved this method from two aspects. First, this method is prone to outliers. To attenuating outliers, acceleration value is used. Acceleration matrix $\mathbf{A}$ is like following:

$$\mathbf{A} = \begin{bmatrix} \alpha_1 & 0 & \cdots & 0 \\ 0 & \alpha_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \alpha_N \end{bmatrix} \tag{9}$$

where $\boldsymbol{\alpha}_i = \sqrt{\alpha_{ix}^2 + \alpha_{iy}^2 + \alpha_{iz}^2}$ denotes the acceleration value correspond to frame i, and $\alpha_{ix}, \alpha_{iy}, \alpha_{iz}$ denotes for the acceleration value on x-axis, y-axis and z-axis. So we change the second part of the objective function to $\lambda \|\mathbf{AZ}\|_{2,1}$.

The second drawback of this method is that it classifies all frames with similar orientation values into one category, regardless of their time difference. For example, if we are recoding a video toward scene A, then we changed our focus point to scene B, after a short while, we switch back to A. The naive DSMRS method will be failed in separating the two scene A which is not correct. Also, if we capturing the video toward one orientation for quite a long time, it is reasonable to select more than one key-frames from this time interval.

So we build a new matrix $\overline{\mathbf{D}}$:

$$\overline{\mathbf{D}} = \begin{bmatrix} \overline{\mathbf{d}}_1^T \\ \overline{\mathbf{d}}_2^T \\ \cdots \\ \overline{\mathbf{d}}_N^T \end{bmatrix} = \begin{bmatrix} \overline{d}_{11} & \overline{d}_{12} & \cdots & \overline{d}_{1N} \\ \overline{d}_{21} & \overline{d}_{22} & \cdots & \overline{d}_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ \overline{d}_{N1} & \overline{d}_{N2} & \cdots & \overline{d}_{NN} \end{bmatrix} \tag{10}$$

where $\overline{d}_{ij} = \sqrt{\|i-j\|}\,\|o_i - o_j\|$. So the new objective function should be:

$$\min_{\mathbf{Z}} \lambda \|\mathbf{AZ}\|_{2,1} + tr(\overline{\mathbf{D}}^T \mathbf{Z}) \quad s.t. \mathbf{Z} \geq 0, 1^T \mathbf{Z} = 1^T \tag{11}$$

We denote the method using only orientation value as DSMRS-O and the method introduced acceleration value and frame time distance as DSMRS-OAD. Fig.3 shows how our improvement performs on the same video we used in Fig.2. From the orientation distribution we can see that there are

two main orientation clusters in that video which located in three time interval:

- Frame #40-#296 are captured toward orientation cluster A
- Frame #342-#514 are captured toward orientation cluster B
- Frame #605-#821 are captured back to orientation cluster A

Other time intervals are outliers caused by hand shake and scene change. Method DSMRS-O chose 5 key-frames: #81, #256, #381, #541 and #546, DSMRS-OAD extracted 3 key-frames: #46, #381 and #721. It is clear that our improvement covered all 3 intervals using exact 3 key-frames. Instead, DSMRS-O failed to cover the time interval #605-#821.

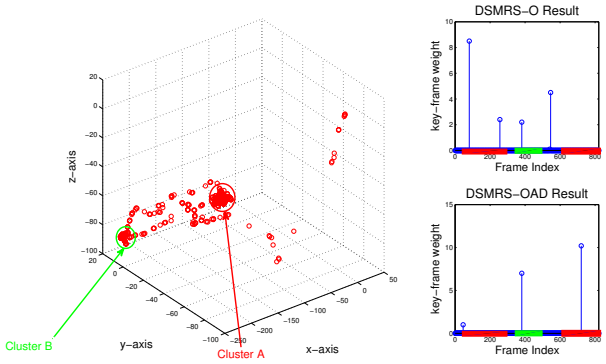

Fig. 3. The left part of the figure shows the orientation distribution of the video. The top-right part describes the result calculated by method DSMRS-O and the bottom-right part shows the method calculated by method DSMRS-OAD. The red color bar in the x-axis of DSMRS-O result and DSMRS-OAD result denotes Cluster A, and green color bar denotes Cluster B.

IV. Experiment Result

A. Dataset Introduction

As far as we know, there is no public available user-generated-video dataset which contains acceleration data and orientation data. So we build our own dataset in order to evaluate our method and compare the result with other methods. Our dataset contains 32 video clips with duration between 30 seconds to 1.5 minutes. The videos are recorded by 4 different person in order to get rid of the influence introduced by personal habits. And the videos are captured from different scenes including campus, playground, lab, etc. Most video contains scene changes, and some of them even contains intentional camera shake in order to increase the difficulty. The frame rate is 30 Hz, and the sample frequency of sensor data is 60 Hz.

B. Method Introduction

Four methods have been evaluated in this section: (1) SMRS: a dictionary method proposed by [12][13]. (2) SMRS-RSI, an improvement of method SMRS which RSI stands for Row-Sparsity-Index, its main purpose to make this improvement is to attenuate outliers. (3) Naive DSMRS-O proposed by [10], (4) DSMRS-OAD proposed in this paper. It use acceleration data to exclude the influence of outliers. Both SMRS and SMRS-RSI use visual feature to extract key-frames, the visual feature is proposed by [12], while DSMRS-O and DSMRS-OAD use orientation data and acceleration for key-frame extraction.

C. Qulitative Evaluation

Fig. 4. The results of the video Bike Repairing. From top to bottom: (1)Annotated results. (2)SMRS results. (3)SMRS-RSI results. (4)DSMRS-O results. (5) DSMRS-OAD results.

Fig. 5. The results of the video Preparation before Basketball Game. From top to bottom: (1)Annotated results. (2)SMRS results. (3)SMRS-RSI results. (4)DSMRS-O results. (5) DSMRS-OAD results.

In this section we list the key-frame extraction results of various methods on two representative video clips to show the advantages of our method over SMRS/SMRS-RSI. For the video Bike Repairing, we labeled 5 key-frames by hand and compare them to the each 5 key-frames extracted by the four methods in Fig.4. Annotate key-frames is not an easy task because key-frames are really subjective. We labeled 5 key-frames which could describe the whole story: (1) A student is pumping his own bike(#153). (2) An overview of

the whole bike repair shop(#372). (3) A couple passes by the bike repair shop, and the girl is checking their bike situation (#719). (4) The couple decides to have their bike repaired in the shop (#973). (5) A close-up of the shop's sign to illustrate the location of this story happened (#1381). These five key-frames can reflect the user's intention.

According to [12], there are two criteria to determine if a key-frame is a good match or not: (1)The extracted frame should be visually similar toward one hand labeled key-frame. (2)The extracted frame and its correspond hand labeled frame should occurred within a short period of time. We adopted this method to evaluate the four algorithm. In Fig.4 and Fig.5, we surrounded a color box outside of every hand labeled frame and each frame has distinct color. In the following four lines, every frame which is a good match, is surrounded by a box having the same color with its corresponding hand labeled key-frame. Those outliers are surrounded with blue box, and non-matches have no outlines. A good method should covers all annotated key-frames and with no outliers. In another word, the method's corresponding line should contain variety of colors and without blue color boxes.

As shown in the second row in Fig.4, SMRS gets 3 good matches (#167, #373 and #710) and one outlier (#563). In addition, it wrongly includes a frame #639, which is similar to #710 but does not provide more interesting information. As a result, it misses the key-frame #973, which shows that the couple decides to repair the bike.

The results of SMRS-RSI are shown in the third row in Fig.4. Although this method is proposed to attenuate outliers, in our scenario, it incorporates more outliers in the results (see the two blue boxes) and only get one good matches (#1385).

The fourth row shows that DSMRS-O obtains 3 good matches (#501, #656, #796/#1056) and no outliers. The last row shows our improved method DSMRS-OAD extracted 4 good matches (#226, #761/#1176, #1371), which is the best among the four methods, but does not include any outliers.

For the video Preparation before Basketball Game, the 5 annotated key-frames and the results of various methods are shown in Fig.5. The key-frames should summarize such a plot: (1) the blue team is warming up (#3). (2)A close-up to the captain of the blue team (#376). (3) The green team is making preparations (#607). (4) The referee is checking the official ball for the game (#766). (5) The blue team is still warming up (#979).

As shown in the second row in Fig.5, SMRS gets 2 good matches (#91 and #580) and one outlier (#134). The results of SMRS-RSI are shown in the third row of Fig.5. It shows that SMRS-RSI still does not provide any improvements on outlier rejection. The 4-th row shows that DSMRS-O get 4 good matches, and one outlier, the last row shows that our method DSMRS-OAD get 4 good matches and no outlier. This indicates that our method generates most representative key-frames on this video and performs best in outlier attenuating.

D. Quantitative Evaluation

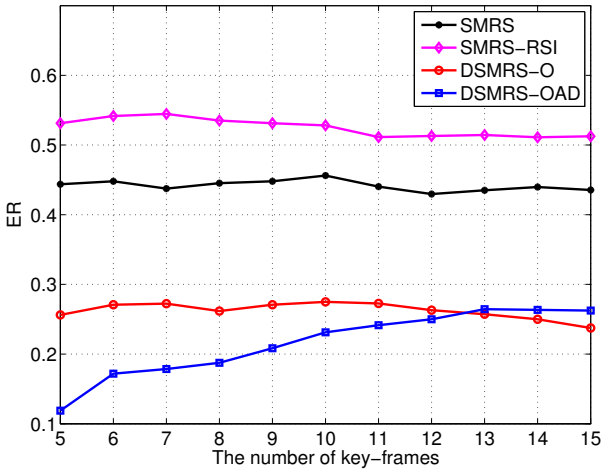

Fig. 6. The average ER for all four methods on 32 videos.

The quantitative evaluation of the quality of the selected key-frame is not easy to determine because the key-frames are subjective for different person. In this paper, we defined two criteria to evaluate the quality of key-frames. First, we use error rate to describe how many outliers are falsely selected. Outliers are a set of frames, which have motion blur, captured during scene change or camera shaking. We define error rate as $ER = \frac{N_e}{K}$, where K is the number of selected key-frames and N_e is the number of outliers exist in these key-frames. A smaller ER value indicates a better method.

All of the four methods can generate ranked key-frames. So we select $\{5, 6, 7, \cdots, 15\}$ key-frames using each method and calculate the ER for the four methods. The result is shown in Fig.6. The figure shows that method SMRS-RSI extracts most outliers in the four methods, because its ER value is around 0.5 regardless how many key-frames are selected. SMRS is slightly better which has an ER value at around 0.45. DSMRS-OAD performs best under most circumstance, however, it is a bit worse than DSMRS-O when the key-frame number is larger than 13. Thus, we can say our improvement performs well in most situation.

It is widely accepted that the meaning of key-frame extraction is to use a small amount of frames to describe the whole video. To evaluate the degree of key-frame's representative capability of the whole video, we introduced reconstruction error on visual features proposed by [12]. We define reconstruction error as $RE = \frac{1}{N} \|\mathbf{D} - \mathbf{D}_K\mathbf{X}\|_F^2$, where $\mathbf{D}$ is the visual feature matrix of the whole video, $\mathbf{D}_K$ is the feature matrix of key-frame set, and $\mathbf{X} = (\mathbf{D}_K^T\mathbf{D}_K)^{-1}\mathbf{D}_K^T\mathbf{D}$ is the coefficient matrix.

Because we have already labeled the outliers, we can define another reconstruction error on only meaningful frames. This criteria is more reasonable because we do not need to represent the outliers. Reconstruction error on meaningful frames is defined as $RE_m = \frac{1}{N_m} \|\mathbf{D}_m - \mathbf{D}_K\mathbf{X}_m\|_F^2$, where N_m is the number of meaningful frames, $\mathbf{D}_m$ is

the feature matrix of all meaningful frames, and $\mathbf{X}_c = (\mathbf{D}_K^T\mathbf{D}_K)^{-1}\mathbf{D}_K^T\mathbf{D}_c$ is the coefficient matrix for meaningful frames.

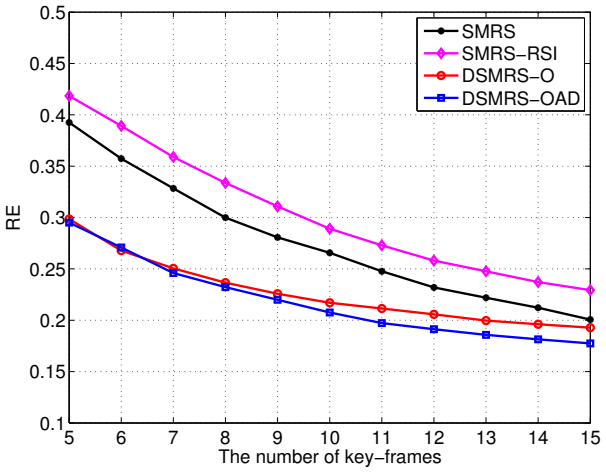

Fig. 7. The reconstruction error of all four methods.

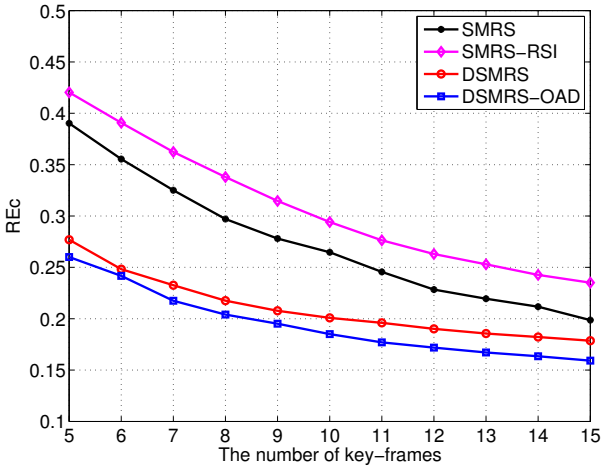

Fig. 8. The reconstruction error on only meaningful frames of all four methods.

We calculated the RE and RE_m for all four methods, the result is shown in Fig.7 and Fig.8. From these three figures and previous description, we can say that our method has three advantages over SMRS/SMRS-RSI: (1) Our method performs good for outlier attenuating because the outlier rate in the result generated by our method is significantly smaller than SMRS/SMRS-RSI. (2) Our method obtains more representative key-frames than SMRS/SMRS-RSI. The reconstruction error of DSMRS-O/DSMRS-OAD is ranging from 0.3 to 0.2 depending on the key-frame number selected, while these two values of method SMRS/SMRS-RSI is from 0.4 to 0.3.The difference on reconstruction error of clean frames is more evident. This proves our method extracts more representative frames both on all video frames and clean frames. (3) We extract key-frame using orientation and acceleration value, which could be directly obtained from smartphone, instead of visual features which have to calculate based on the video. Our method saved the time of feature calculating.

We express the detail result of 5 randomly chosen videos in Fig.9, where weight means the weight of the corresponding frame is considered as a key-frame. The left part shows the vector $[\|\mathbf{x}_1\|_2, \|\mathbf{x}_2\|_2, \cdots, \|\mathbf{x}_N\|_2]^T$, where $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N \in \mathbf{X}$ is the coefficient matrix optimized by method SMRS. The right part shows the vector $[\|\mathbf{z}_1\|_2, \|\mathbf{z}_2\|_2, \cdots, \|\mathbf{z}_N\|_2]^T$, where $\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_N \in \mathbf{Z}$ is the representative capability matrix $\mathbf{Z}$ calculated by method DSMRS-OAD. Fig.9 vividly shows that our method generates much more sparse result comparing to method SMRS. This means we can choose every non-zero rows as key-frame in our method, while we have to set a threshold for method SMRS. Also it explains why method SMRS's result performs worse than our method on reconstruction error: we have to omit most nonzero rows from the computed result, which will destroy the representative capability of the result.

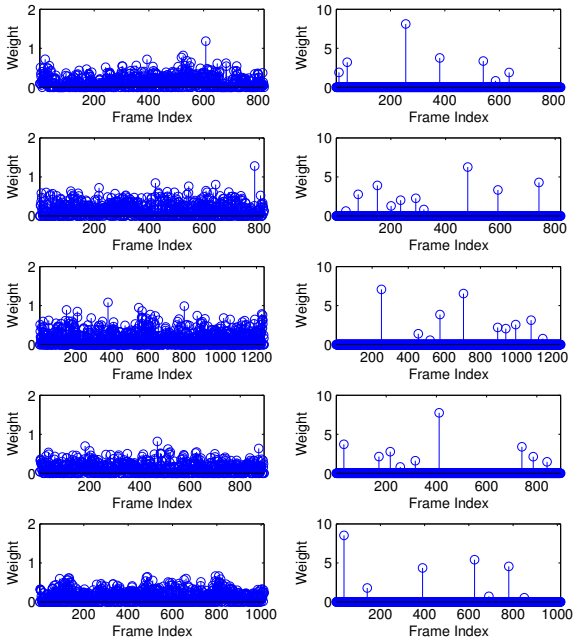

Fig. 9. The key-frames extracted by method SMRS(left) and method DSMRS-OAD(right).

V. Conclusion

In this paper, we proposed a new method DSMRS to solve key-frame extraction, and we use orientation data, which is easy to obtain, instead of traditional visual features to extract key-frame. Also we introduced acceleration data to attenuating outliers. At last, we build a real dataset to test the proposed method.

REFERENCES

[1] G. Hua, Y. Fu, M. Turk, M. Pollefeys, and Z. Zhang, "Introduction to the special issue on mobile vision," *Int. J. of Computer Vision*, vol. 96, no. 2, pp. 277-279, Feb. 2012.

[2] Y. Tian, W. Wang, X. Gong, X. Que, and J. Ma, "An enhanced personal photo recommendation system by fusing contextual and textual features on mobile device," *IEEE Transaction on Consumer Electronics*, vol. 59, no. 1, pp. 220-228, Feb. 2013.

[3] W. Abd-Almageed, "Online, simultaneous shot boundary detection and key frame extraction for sports videos using rank tracing," *IEEE International Conference on. Image Processing(ICIP)*, pp.3200-3203, 2008.

[4] S. P. Yong, J. D. Deng and M. K. Purvis, "Key-frame extraction of wildlife video based on semantic context modeling," *In The 2012 International Joint Conference on Neural Networks (IJCNN)*, pp.1-8, 2012.

[5] J. Luo, C. Papin, K. Costello, "Towards extracting semantically meaningful key frames from personal video clips: From humans to computers," *IEEE Trans. Circuits and Systems for Video Technology*, vol.19, no.2, pp.289-301, 2009.

[6] G. Abdollahian, C. M. Taskiran, Z. Pizlo, and E. J. Delp, "Camera motion-based analysis of user generated video," *IEEE Trans. on Multimedia*, vol. 12, no. 1, pp. 28-41, Jan. 2010.

[7] J. C. Huang and W. S. Hsieh, "Automatic feature-based global motion estimation in video sequences," *IEEE Trans. on Consumer Electronics*, vol. 50, no. 3, pp. 911-915, Aug. 2004.

[8] S. K. Kim, S. J. Kang, T.-S. Wang, and S.-J. Ko, "Feature point classification based global motion estimation for video stabilization," *IEEE Trans. Consumer Electronics*, vol. 59, no. 1, Feb. 2013.

[9] N. D. Lane, E. Miluzzo, H. Lu, D. Peebles, T. Choudhury, and A. T. Campbell, "A survey of mobile phone sensing," *IEEE Communications Magazine*, vol. 48, no. 9, pp. 140-150, Sept. 2010.

[10] E. Elhamifar, G. Sapiro, A. Yang, et al. "A Convex Optimization Framework for Active Learning," *IEEE International Conference on Computer Vision(ICCV)*, 2013.

[11] E. Elhamifar, G. Sapiro, and R. Vidal. "Finding exemplars from pairwise dissimilarities via simultaneous sparse recovery," *Neural Information Processing Systems(NIPS)*, pp.19-27, 2012.

[12] Y. Cong, J. Yuan, J. Luo, "Towards scalable summarization of consumer videos via sparse dictionary selection," *IEEE Trans. Multimedia*, 2012, 14(1): 66-75.

[13] E. Elhamifar, G. Sapiro, R. Vidal, "See all by looking at a few: Sparse modeling for finding representative objects," *IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, pp.1600-1607, 2012.