Summarization of Egocentric Moving Videos for Generating Walking Route Guidance

Masaya Okamoto* and Keiji Yanai

Department of Informatics, The University of Electro-Communications
1-5-1 Chofugaoka, Chofu-shi, Tokyo 182-8585 Japan
{okamoto-m,yanai}@mm.cs.uec.ac.jp

Abstract. In this paper, we propose a method to summarize an egocentric moving video (a video recorded by a moving wearable camera) for generating a walking route guidance video. To summarize an egocentric video, we analyze it by applying pedestrian crosswalk detection as well as ego-motion classification, and estimate an importance score of each section of the given video. Based on the estimated importance scores, we dynamically control video playing speed instead of generating a summarized video file in advance. In the experiments, we prepared an egocentric moving video dataset including more than one-hour-long videos totally, and evaluated crosswalk detection and ego-motion classification methods. Evaluation of the whole system by user study has been proved that the proposed method is much better than a simple baseline summarization method without video analysis.

Keywords: Egocentric Vision, Video Summarization, Walking Route Guidance.

1 Introduction

In this paper, we propose a system to summarize an egocentric moving video. The goal of summarization of an egocentric moving video in this work is to produce a compact summary video, assuming that an input video is an egocentric video recorded continuously from a departure place to a destination by a wearable camera that a walking person is wearing. Therefore, the generated summary video can be used as a guidance video which explains the walking route from a departure place to a destination. In general, making a route guidance video manually is a time-consuming job. With this system, we can generate compact guidance videos on walking path routes very easily by walking the routes with a wearable camera.

To produce a summarized video, we make use of ego-motion (motion of the person wearing a wearable camera) and detection of a pedestrian crosswalk as cues. Ego-motion is estimated based on optical flow of the scene, while crosswalks

* He is currently a graduate student at Graduate School of Information Science and Technology, The University of Tokyo.

R. Klette, M. Rivera, and S. Satoh (Eds.): PSIVT 2013, LNCS 8333, pp. 431–442, 2014.

are detected by Geometric Context [1] and a standard object recognition method based on bag-of-features representation and non-linear SVM classifier. Based on these cues, the system estimates importance of each section in the given video.

Actually, in the proposed system, summarized video files are not generated. Instead, a playing scenario is generated which instructs relative playing speed on each video section according to the importance scores of video sections. Basically, we play important scenes at a normal speed, while we play less important scenes at a high speed or skip them. With the playing scenario, the system dynamically controls video playing speeds. Since HTML5 video player can control video playing speeds even while playing a video, we implemented a viewing system on the Web in HTML5. That is, the proposed system does not generate a summary video in advance, but generates a summery video dynamically. Therefore, a user can adjust a total playing time of a summary video, when watching. This is the biggest difference from the previous works.

In the rest of this paper, we describe related work in Section 2. In Section 3, we describe the overview of the proposed system, and explain the methods to classify ego-motion, to detect crosswalks and to evaluate importance scores of video sections in Section 4. Section 5 shows experimental results and user study. We conclude this paper in Section 6.

2 Related Work

In this section, we review some works on video summarization related to egocentric vision and ego-motion classification.

To summarize a video, it is common to divide a given video into some shots, select important shots from them, and concatenate selected shots for generating a summarized video [2,3]. In contrast to them, our goal is to generate a route guidance video, and it does not erase any video shots from an original video because lack of shots confuse people who watch the video regarding a current location of the video. To achieve video summarization for route guidance, we control video playing speed according to the importance scores of video sections, and unrelated scenes for route guiding are fast-forwarded.

Some works on egocentric video summarization have been proposed so far. One of the works which is similar to ours is Lee et al.'s work [4]. In their work, they used no GPS or other sensors, and focuses on image-based detection of the most important objects and people with which the person wearing a camera interacts. The proposed method selects some important frames to create storyboards to explain the recorded egocentric video shortly. In contrast to our work, its targets are any kinds of egocentric videos, and they used not motion cues but only static image features. The results are represented by a set of selected frame images as a storyboard, while our objective is summarizing an egocentric video into a shorter video.

On the other hand, some other works used auxiliary information such as GPS and motion sensor logs to summarize life-log videos. Datchakorn et al. [5] proposed a system which summarizes life-log videos based on some sensor logs

including gyro, acceleration and GPS. In contrast to this work, we use only visual features extracted from videos, and need no GPS records.

As a method to analyze egocentric videos, ego-motion classification is common. For example, Kitani et al. [6] proposed a method to classify motions in several kinds of sports videos such as surfing and skiing employing non-parametric Bayesian model in a unsupervised way. In this work, as a feature to classify ego-motion, optical flow is used. In our work, we also use only optical flow to classify ego-motions as a visual motion feature. As another work on first-person activity analysis, Ogaki et al. [7] proposed a method on ego-motion recognition for an "inside-out" camera, which can capture both frontal view and eye movement of the person at the same time. In their work, they use the first-person eye movement as well as ego-motion in order to recognize indoor activities. In our work, we use normal wearable camera which can capture only frontal view, and target outdoor videos.

3 Overview of the Proposed System

In this section, we explain an overview of the proposed system. In the proposed system, we estimate importance of each video section of the given egocentric video based on ego-motion and existence of crosswalks, and generate playing scenario off-line. Based on the scenario, we summarize the given video dynamically by controlling playing speed of the HTML5 video player on-line. The off-line processing to generate a playing scenario is as follows:

1. Ego-motion classification based on optical flow with SVM.
2. Detection of pedestrian crosswalks with bag-of-features and non-linear SVM after estimation of road areas by Geometric Context [1].
3. Estimation of importance of each video section (The duration of each section is four second long.)

4 Method

In this section, we explain the detail of each processing step.

4.1 Ego-motion Classification

In general, a video taken by a wearable camera is called "an egocentric video" or "a first-person video". The most basic analysis on such a video is estimation of the motion of people wearing a camera, which is called "ego-motion".

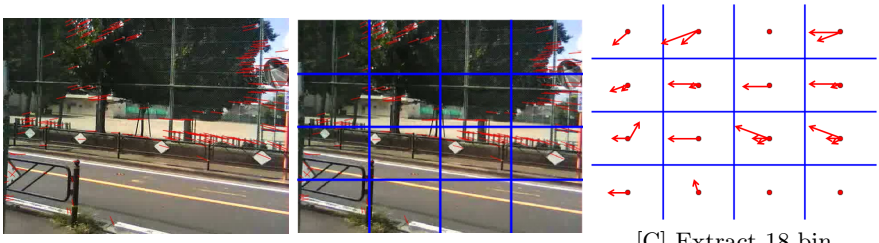
Since we assume that the given video is taken while walking in this work, we classify ego-motion of each section of the given video into one of the following four moving conditions: (1) moving forward (2) stopping (3) turning right (4) turning left.

Because we treat with only walking egocentric videos, we assume that the videos contain only forward, right and left moves, and do not contain backward

moves. If several kinds of moves are included in a unit shot, it will be classified into the majority move.

To classify ego-motion, firstly we extract twelve frame images per second from the given video for every four seconds, and secondly estimate optical flows for each interval of frame images using the Lucas-Kanade method [8], and then we convert them into histogram-based representation.

Figure 1 shows an example of calculating feature vector. We calculate coordinates of the feature points on the current video frame given the coordinates on the previous frame. The function finds the coordinates with sub-pixel accuracy. The red lines in (A) represent the obtained optical flows. To convert raw optical flows into a feature vector, we divide an image into 4×4 grids (B), and extract a 18-bin directional histogram from each block with dividing the optical flow direction by 20 degrees (C). Since we set temporal window size for motion feature extraction as four second long, we average feature vectors for four seconds and L1-normalize the averaged feature vector. Finally we obtain a 288-dim optical flow feature vector for every four-second sections of the given video.



[A] Optical flow [B] Divide into grids [C] Extract 18-bin histograms

Fig. 1. Calculating feature vector

To classify ego-motion of each four-second video section into one of the above-mentioned four kinds of the ego-motion conditions, we use SVM as a classifier for the extracted ego-motion feature vectors. In this work, we use LIBSVM library and RBF kernel. We train four SVM classifiers in the one-vs-all way. We prepare hand-labeled training data for training in the experiment. In the step to estimate video section importance, we use the pseudo-probability values obtained by applying a sigmoid function to the output values of SVMs.

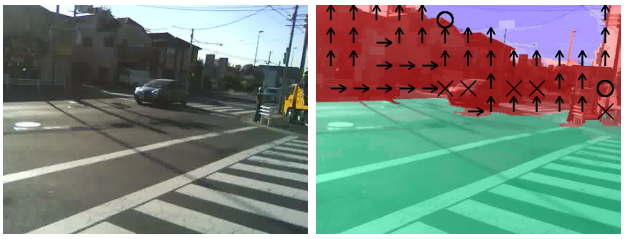
4.2 Crosswalk Detection

For walking route guidance in cities, crosswalk is an important and remarkable cue to explain walking routes. To prevent the scenes containing crosswalks from being skipped or played in a high speed, we detect pedestrian crosswalks in the given egocentric video as well as ego-motion conditions.

For crosswalk detection, we extract three frames per second, and detect the frames including crosswalks after road region estimation for each extracted frame.

We estimate road regions for each extracted frame. To do that, we use Geometric Context [1]. This work provides a multiple hypothesis framework for robustly estimating scene structure from a single image and obtaining confidences for each geometric label (See Figure 2). We regard ground regions as road regions which may contain crosswalks. The right figure in Figure 2 shows a final output of the Geometric Context method where ground (green), sky (blue) and vertical (red) regions are classified, subdivided into planar orientation (arrows) and non-planar solid ('x') and porous ('o').

We use results of boosted decision trees for main class, "support"(ground). The pixels that its likelihood is higher than the pre-defined threshold are regarded as spatial support pixels. In our work, the threshold is set 0.4.



Input Surface Layout

Fig. 2. Geometric context method

Ground regions estimated by Geometric Context method tend to fragment and have holes of non-ground regions. We integrate these regions and fill up the holes because fragmentation and holes make bad influence on the following feature extraction. We apply the closing method of morphological operations to input regions. Figure 3 shows an example of ground region integration. Green regions in these images correspond to the estimated ground regions. The regions applied the integration method are estimated as ground regions finally.

[A] Input [B] Ground Region [C] Output

Fig. 3. An example of ground region integration

We use a bag-of-features vector (BoF) with SIFT as an image feature for crosswalk detection. First, we extract keypoints using the default keypoint detector of SIFT which is the difference of Gaussian (DoG) from the ground region

estimated in the previous step, and describe local patterns around the keypoints with SIFT descriptor. Next, we vector-quantize extracted SIFT features against the pre-specified codebook and generate a bag-of-features (BoF) vector regarding each of the extracted frames. After that, all the BoF vectors are L1-normalized. In the experiments, we built a 500-dim codebook by k-means clustering with SIFT features sampled from training data.

We use non-linear SVM as classifier to examine if the given frame contains a crosswalk or not. In the experiments, we trained SVM with 80 positive images and 160 negative images.

4.3 Estimation of Importance of Video Sections

Based on the results of ego-motion classification and crosswalk detection, we estimate a importance score of each video section of the given egocentric video. We divide the given video into video sections every four seconds. The importance score varies between 0.0 and 1.0, which decides playing speed of the corresponding video section.

The importance scores of the i-th video section S_i in terms of ego-motion is calculated in the following equation:

$$S_i = c_f v_{f[i]} + c_s v_{s[i]} + c_r v_{r[i]} + c_l v_{l[i]}, \tag{1}$$

where $v_{f[i]}$ $v_{s[i]}$, $v_{r[i]}$ and $v_{l[i]}$, are the sigmoid function values of the SVM outputs regarding "moving forward", "stopping", "turning right" and "turning left", respectively. c_f, c_s, c_r and c_l represent weighting factors. We set those constants as follows: c_f = -2, c_s = 1, c_r = 2, and c_l = 2. Because we regard the scenes classified as "turning right or left" as being important, c_r and c_l are positive values. Because the scene classified as "moving forward" is less important, we set c_f as a negative value.

After S_i is obtained, we normalized it so that it varies between 0 to 1 by using maximum S_{max} and minimum S_{min} within the given video in the following equation:

$$S'_i = \frac{S_i - S_{min}}{S_{max} - S_{min}} \tag{2}$$

The sections where crosswalks are detected are important cues for walking route guidance. We sum all the outputs of the SVM classifiers for crosswalk detection in each video section. The sections where total output values become more than the pre-defined threshold are regarded as being crosswalk sections. In the experiments, we set the threshold as 7, which are decided in the preliminary experiment. The normalized importance scores of the crosswalk sections are added 0.5 by the following equation:

$$S''_i = min(S'_i + 0.5, 1.0) \tag{3}$$

Note that the scores of the first and last four sections are set as 1.0 regardless of ego-motion and crosswalk, since the sections representing the starting place and destination are important.

The list containing importance scores on each video section over the given video is called as "play scenario".

4.4 Calculation of Playing Speed

Finally, we decide speeds of playing video sections based on play scenario. Only this step is carried out on-line when playing, since a user can provide the maximum playing speed. We define playing speed of the i-th section s_i using the following equation:

$$s[i] = \frac{1}{S_i''(1 - (1/(s_{max} - 1))) + (1/(s_{max} - 1))} + 1, \tag{4}$$

where s_{max} is a scaling factor which is expected to be given by a user when playing. s_{max} defines the maximum speed of playing video. s_{max} is set as 7 in the experiments.

When playing practically, the smoothing regarding playing speed is carried out except for the first and last section by the following equation:

$$s'[i] = 0.1(s[i-1] + s[i+1]) + 0.8s[i] \tag{5}$$

5 Experiments

5.1 Data Collection

To collect egocentric videos, we used Looxcie2 (Figure 4). We can wear it on an ear like Figure 5. Figure 6 shows an example of the recorded egocentric video frame.

Fig. 4. Looxcie2

Fig. 5. Video recording style

We collected nine egocentric videos for the experiments as shown in Table 1.

Fig. 6. Examples of the video frame

Table 1. Data collection

Activity	Num	Average duration (min.)
walk	9	9:12

5.2 Evaluation on Ego-motion Classification

First, we evaluate performance of ego-motion classification. To learn four kinds of SVM classifiers, we extract hand-labeled sections from four videos for training. The detail of training data is shown in Table 2. The values in the Table represent the number of video sections.

For this experiment, we extract video sections for test from three videos. The results are shown in Table 3. Figure 7 shows the recall-precision curves of four-kinds of ego-motion classification. The classification rate over four-kinds of ego-motion was 83.8%.

Figure 8 shows a typical failure of ego-motion classification. Since it is a scene of waiting for a traffic light, its video section should be classified as "Stop". However, it was classified as "Go right" under the influence of optical flows from motions of the car. Ego-classification tends to fail in the video sections including many moving objects such as people or bikes.

5.3 Evaluation on Crosswalk Detection Performance

In this subsection, we evaluate the performance of crosswalk detection. For this experiment, we selected four videos for training and five videos for test from

Table 2. Training data for ego-motion classification

Motion	Positive	Negative	Num
Move forward	512	216	728
Stop	74	146	220
Turn right	70	147	217
Turn left	68	131	119

Table 3. Ego-motion classification result

Motion	# section	recall	precision	F-number
Forward	244	0.943	0.697	0.801
Stop	72	0.694	0.893	0.781
Go right	84	0.738	0.969	0.838
Go left	88	0.795	0.972	0.875

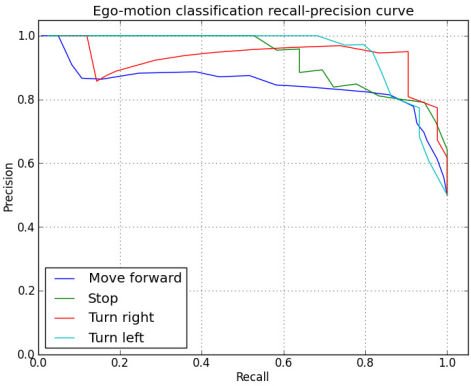


Fig. 7. Recall-Precision curves of ego-motion classification

datasets. We extract 250 images from the learning videos, and extract 200 images from the testing videos. To evaluate the effectiveness of ground region estimation based on Geometric Context, we compared the results with and without ground region estimation. In case of the method without ground estimation, we extracted SIFT keypoints from whole a frame image.

Figure 9 shows the recall-precision curves of crosswalk detection with and without ground region estimation. These results proved that ground estimation could improve the crosswalk detection performance.

Figure 10 shows typical success and failure examples on crosswalk detection. Green regions in these images correspond to the estimated ground regions. In the case of [A], a crosswalk was recognized successfully with ground estimation, while it failed without ground estimation. In this case, ground estimation correctly removed noise regions on the upper part of the given image. On the other hand, in the case of [B], a crosswalk detection failed with ground estimation, while it succeeded without ground estimation. This is because ground estimation recognized a part of the crosswalk region as a vertical region incorrectly.

5.4 User Study

In this subsection, we show the results of user study employing ten subjects. To evaluate the proposed method on walking egocentric video summarization,

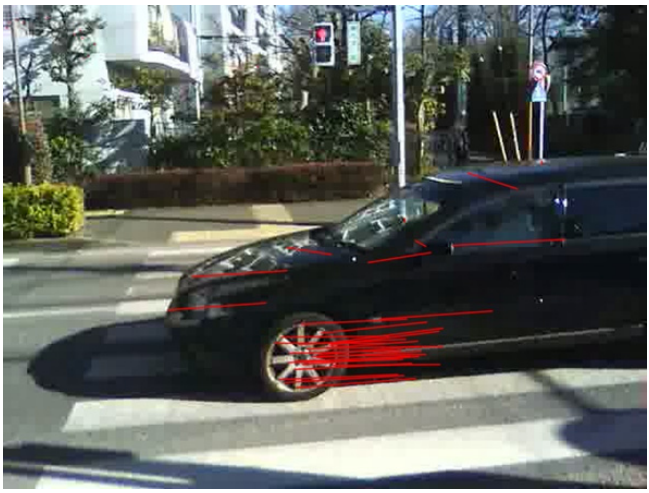


Fig. 8. A typical failure of ego-motion classification

we compared the proposed summarization method with two baseline methods. The first baseline is just playing videos in fast-forwarding at a uniform speed. The second one is a storyboard-style summarization which displays uniformly sampled frames from the video. The numbers of storyboard frames are proportional to the length of the given video. We sampled a frame every five second from the given video. To evaluate the effectiveness of crosswalk detection, we carried out a method using only ego-motion classification without crosswalk detection as well. The summarization methods are experimented as follows:

1. Ego-motion classification only
2. Crosswalk detection + ego-motion classification
3. Baseline 1 (fast-forwarding at a uniform speed)
4. Baseline 2 (storyboard-style (sampled frames at even interval))

We asked the subjects to vote the best summary as a walking route guidance among the storyboard and the videos generated by three kinds of summarization methods for each of the egocentric video. The maximum playing speed s_{max} of an input summary video are fix to 7. The details of input and summary video are shown in Table 4.

We obtain 30 votes from 10 subjects. The experimental results are shown in Table 5. The proposed method gathered the best votes on average over three test videos, which means the proposed method based on ego-motion and crosswalk detection was effective compared to the two baselines and the method without crosswalk detection. The method with only ego-motion received some votes for A and B videos. It means the method based on only ego-motion was some effective. The two baseline methods received few votes for all the videos.

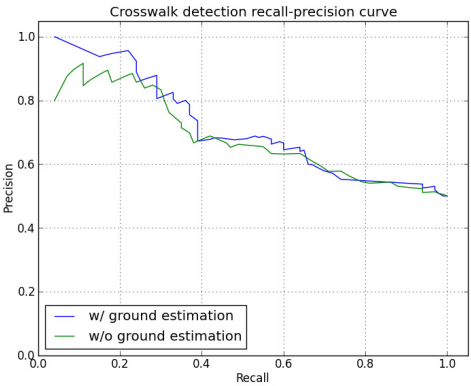


Fig. 9. Recall-Precision curves of crosswalk detection

[A] Success only w/ ground estimation [B] Success only w/o ground estimation

Fig. 10. Crosswalk detection examples of success and failure

Table 4. User study dataset

Video	Duration	After duration	Average speed	Storyboard size
Video A	7:47	1:45	4.5	21
Video B	9:17	2:20	3.9	28
Video C	11:26	2:40	4.3	32

Table 5. User study result

Video	Ego-motion	Ego. + crosswalk	baseline	storyboard
Video A	4	6	0	0
Video B	3	6	1	0
Video C	1	7	1	1
total	8	19	2	1

6 Conclusions and Future Work

In this paper, we proposed a new method to summarize walking ego-centric video for generating walking route guidance. From the user study, it has been proved that the proposed method is better than simple summarization method. About ego-motion classification, it has achieved the classification rate 83.8%. About crosswalk detection, the proposed system has achieved the 63.5% detection rate with ground region estimation.

For future work, we plan to extend target egocentric videos. Although we focused on only walking egocentric video in this paper, we plan to treat with a bike and a car egocentric video. In such cases, other object cues are expected to be importance for egocentric video summarization. We plan to add detection methods on other important objects to the system.

References

1. Hoiem, D., Efros, A., Hebert, M.: Recovering surface layout from an image. International Journal of Computer Vision (2006)
2. Chong-Wah, N., Yu-Fei, M., Hong-Jiang, Z.: Video Summarization and Scene Detection by Graph Modeling. IEEE Transactions on Circuits and Systems for Video Technology 15 (2005)
3. Arthur, G., Harry, A.: Video summarization: A conceptual framework and survey of the state of the art. Visual Communication and Image Representation 19 (2008)
4. Lee, Y.J., Ghosh, J., Grauman, K.: Discovering important people and objects for egocentric video summarization. In: Proc. of IEEE Computer Vision and Pattern Recognition (2012)
5. Tancharoen, D., Yamasaki, T., Aizawa, K.: Practical experience recording and indexing of life log video. In: Proc. of ACM SIGMM Workshop on Continuous Archival and Retrieval of Personal Experiences (2005)
6. Kitani, K.M., Okabe, T., Sato, Y., Sugimoto, A.: Fast unsupervised ego-action learning for first-person sports videos. In: Proc. of IEEE Computer Vision and Pattern Recognition, pp. 3241–3248 (2011)
7. Ogaki, K., Kitani, K.M., Sugano, Y., Sato, Y.: Coupling eye-motion and ego-motion features for first-person activity recognition. In: Proc. of CVPR Workshop on Egocentric Vision (2012)
8. Lucas, B., Kanade, T.: An iterative image registration technique with an application to stereo vision. In: International Joint Conference on Artificial Intelligence, pp. 674–679 (1981)