

REDUCTION OF CAMERA MOTION ADJUSTMENTS UNDER A PLANNED VIDEO COMPOSITION WITH PAN-TILT CAMERA

Yoshinari Kameda¹†, Yasutaka Atarashi², Satoshi Nishiguchi³, and Michihiko Minoh¹

Academic Center for Computing and Multimedia Studies, Kyoto University, Japan¹
Graduate School of Informatics, Kyoto University, Japan²
Graduate School of Law, Kyoto University, Japan³

ABSTRACT

We propose a method to control a pan-tilt-zoom camera in order to realize a planned video composition by reduced number of camera motion changes. Video composition defines the way of filming an object in video image. It is described by position, velocity, size, and magnification rate of the object in addition to velocity of background region in video image. When we take video images, unexpected motion of the object makes difference between the actual video composition and a planned one. If we simply control the camera to realize the planned composition as precise as possible, the camera motion will be adjusted frequently and as a result obtained video images will not be comfortable to viewers. Therefore, we should consider the timing and the amount of camera motion adjustments. Our method can keep video composition within an acceptable range of the planned composition by reduced number of camera motion adjustments.

1. INTRODUCTION

Various approaches [1] [2] [3] have been proposed to film human activities including walking, lectures, concerts, games, etc. Automatic videography of moving objects is one of the essential technologies on filming human activities.

A **video composition** defines the way of framing a moving object in video images. When we utilize pan-tilt-zoom camera, video composition is described by position (p, q), velocity (\dot{p}, \dot{q}), size (z) of the object in video image, size magnification rate (\dot{z}) of the object, and velocity (\dot{r}, \dot{s}) of background region in video image. We call a specified video composition that should be realized **target video composition (TVC)** in this paper. Not all the elements of TVC should be specified on ordinary filming requests. For example, in order to film entire body of a lecturer in a classroom,

the location and the size of the object would be enough to realize sufficient video composition.

When filming is being executed according to the TVC, there may be a gap between the TVC and observed video composition (o-VC) due to the unexpected location of the object.

In various automatic videography methods [4] [3] [5] [6], a pan-tilt-zoom camera is controlled so as to minimize the gap between TVC and o-VC at each video image. As a result, the motion of the camera is changed frequently. This becomes more obvious when it is requested to eliminate the gap as much as possible and as soon as possible. However, this frequent motion adjustment will cause a serious problem of video quality for filming human activities because ordinary viewers may not stand such camera motions. Therefore, not only the minimization of the gap but also the timing and the amount of the camera motion adjustment should be carefully determined to provide a comfortable video to the viewers.

General video servoing approaches such that Tahri et al.[7] proposes do not pay much attention on video quality against viewers although they can track an object with pan-tilt or 6DOF cameras effectively.

Ozeki et al.[8] proposed virtual frame control to reduce unpleasant camera motion on tracking a object manipulated by an operator. In their approach, the direction of pan-tilt camera is fixed while the object is located inside a virtual frame that is set on a video image, and tracking the object starts just after the object goes beyond the virtual frame. When it comes to settle down, the camera is fixed again. This method works well because they think limited variations of parameter specification of video composition. In other words, their approach does not support other kinds of video composition such as the speed of the object inside the video image is specified.

In this paper, we propose a new method to control a pan-tilt-zoom camera in order to realize TVC by reduced number of camera motion adjustments. It does not only re-

† Current Address: Institute of Engineering Mechanics and Systems, University of Tsukuba, Japan, kameda@image.esys.tsukuba.ac.jp.

duce the gap between TVC and o-VC measured on video image but also determines the timing and amount of the camera motion adjustment to provide comfortable videos to viewers. Our approach uses optical flow both to identify foreground region and background region and to estimate current video composition. M-estimation is used to classify optical flows because it is sometimes affected by noise. We use Kalman filter to predict future camera-work precisely, which is needed to suppress unnecessary motion adjustments.

We introduce an **acceptable range** for every parameter of TVC. Each acceptable range is given by its lower limit and upper limit. If some parameters of TVC are not specified, corresponding acceptable ranges are set infinite. Video composition is thought to be realized if all the parameters of o-VC are within the acceptable range of TVC.

TVC could be determined based on the interest and situation of viewers. In this paper, we assume TVC and its acceptable range are given by other method[9] or set manually.

We take up a situation of filming a lecturer in a classroom as our application in this paper.

The rest of this paper is organized as follows. Camera control framework of automatic videography is described in Section 2. The camera adjustment reduction method is explained in Section 3. Our experiments are shown in Section 4 where we discuss the ability of our method. Finally, we conclude this paper in Section 5.

2. FRAMEWORK OF CAMERA CONTROL

2.1. Notation

We assume that at most one object may be moving in a scene. Pan and tilt axis of a pan-tilt-zoom camera are orthogonal to each other and both pass the focal point of the camera. Motion of the camera is controlled by feeding pan speed, tilt speed, and focal length (\dot{P} , \dot{T} , F). Hence, a change of camera motion is defined by feeding different (\dot{P} , \dot{T} , F) to the camera. These parameters are discrete in practical pan-tilt-zoom cameras.

Video composition is defined by position (p , q), velocity (\dot{p} , \dot{q}), size (z) of the object in video image, size magnification rate (\dot{z}) of the object, and velocity (\dot{r} , \dot{s}) of background region in video image. See Figure 2.1. Image size is normalized within -0.5 and 0.5 , and the center of the image is $(0.0, 0.0)$.

2.2. Camera Control Procedure

Camera control procedure is conducted by repeating process loop. Process loop consists of 3 phases; observation, prediction, and execution.

In the observation phase, current video composition is estimated based on the optical flows in the current video image. As the observed video composition may be affected by error and noise, we call it **o-VC** to distinguish it from actual current video composition.

In the prediction phase, future video compositions are deduced according to the current o-VC and preceding o-VCs. We call the estimated future video composition **e-VC**. In order to predict e-VCs for next L loops, we use Kalman filter that is designed for this kind of repetition procedure.

After obtaining L e-VCs, new camera command is issued if camera motion adjustment is needed in execution phase. Amount of the adjustment is determined so as to let e-VCs within the acceptable range for future L loops.

3. REDUCTION OF CAMERA CONTROL ADJUSTMENTS

3.1. Extraction of o-VC by M-estimation

In our approach, there are at most two regions each of which has unique horizontal speed, vertical speed, and scaling speed (H, V, S) in video image. As o-VC is easily calculated from (H, V, S), we focus on extraction procedure of (H, V, S) in this section.

If there are two regions, one should be foreground object and the other a background. We apply two stage M-estimation procedure to classify optical flows into two regions. In the first stage, (H_p, V_p, S_p) is estimated temporarily by inputting all the optical flows on video image to M-estimation method. **Primary region** is a region that has the flow (H_p, V_p, S_p). Then, in the second stage, (H_s, V_s, S_s) is estimated based on the M-estimation of the optical flows that are not included in the primary region. We call the region that has the flow (H_s, V_s, S_s) **secondary region**. The region of which (H, V, S) is consistent with (\dot{P}, \dot{T}, F) of the previous process loop is labeled as background region, and the other is labeled as foreground region.

The whole procedure to estimate foreground and background region is shown below.

Optical Flow Optical flow (u_i, v_i) at (x_i, y_i) ($i = 1, \dots, N_{opt}$) on video image is extracted by block matching method. N_{opt} is a fixed number of locations for optical flow evaluation.

Region Vector As the camera neither moves translationally nor rotates around optic axis of the camera, (u_i, v_i) can be written by region vector.

$$u_i = H + Sx_i, \quad v_i = V + Sy_i \quad (1)$$

To estimate the region vector of the primary region (H_p, V_p, S_p), M-estimation is applied. M-estimation

3.3. Determination of Camera Control Parameter

To minimize the number of the times of camera motion adjustments, new camera control parameter will be issued only if e-VC of the next loop denoted as $\vec{\alpha}_{n+1|n}$ will go beyond acceptable range.

The new camera control parameter $(\dot{P}_n, \dot{T}_n, F_n)$ is determined so as to let e-VCs within the acceptable range for the future L loops. If not possible, we adopt the parameter that ensures the e-VCs in the acceptable range as long as possible.

On the n th loop, $\vec{\alpha}_{n+k}$ ($k = 1, 2, \dots, L$) can be calculated according to Eq.(20) because $\vec{\alpha}_{n|n}$ has been obtained in the observation phase.

$$\vec{\alpha}_{n+k|n} = \vec{A}^k \vec{\alpha}_{n|n} + \vec{A}^{k-1} \vec{c}_n \quad (21)$$

Therefore, \hat{c}_n that realizes acceptable $\vec{\alpha}_{n+k}$ for $k = 1, 2, \dots, M$ ($M \leq L$) can be estimated as follows.

$$\vec{c}_n = \vec{A}^{-(k-1)} \left(\vec{\alpha}_{n+k} - \vec{A}^{-k} \alpha_{n|n} \right) \quad (22)$$

With Eq.(10) and Eq.(22), $(\dot{P}_n, \dot{T}_n, F_n)$ can be determined so that it realizes $\vec{\alpha}_{n+k}$ for all M loops within the acceptable range.

4. EXPERIMENT

We have implemented our method and conducted experiments.

Our system uses SONY EVI-G30 as a shooting camera and Hitachi IP-5000 image processing board to calculate optical flows. Size of input gray image is 320 by 240 pixels. Optical flows are extracted at 16 rows by 12 columns. One loop takes 66ms on average in this system and L is set to 10.

Although our method can deal with object size s and s' , we assume these two parameters to certain values and fix the focal distance F in the experiments because our prototype system has rough resolution for object size estimation that is determined by the positional interval of optical flow extraction. This limitation will be eliminated if the system can handle many optical flows.

We also omit tilt control because pan control and tilt control can be done independently and the method itself is completely the same for both controls.

In the experiments, the system filmed a person walking. The distance between the pan-tilt camera and the person was about 6 meters. The person was asked to walk freely for 25 seconds on a line that is almost orthogonal to the optic axis of the camera.

We presented four different filming in this paper. TVC is specified to set the walker in the center of the video image, and acceptable ranges are varied. In the first two examples, we set the same acceptable range of horizontal position while we set different sizes. These examples are labeled as

Data-1 (position range 0.25 and size 0.2) and Data-2 (position range 0.25 and size 0.3).

The object sizes in video are shown in Figure 2, and corresponding results are shown in Figure 3 and Figure 4.

Graph (a) indicates the position along with the time and graph (b) the speed. In graph (a), horizontal axis indicates time in seconds and vertical axis indicates x value in video image. Black line shows e-VC which was estimated 5 loops (0.33 seconds) before and dotted line means true camera-work that is calculated based on a color extraction of the walker. The '+' mark shows o-VC and two horizontal dotted lines show the acceptable range of this filming. In graph (b), vertical axis indicates speed of x direction. As no acceptable ranges of the speed are set in Data-1 and Data-2, there are no horizontal dotted lines in graph (b).

In Data-3, the acceptable range of the speed is specified to 0.15 and the size to 0.3. In Data-4, both the position and the speed are specified. The values are position range 0.25, speed range 0.15, and size 0.3). Figure 5 and Figure 6 show the results of Data-3 and Data-4 respectively.

The number of times of camera control adjustments is shown in Table 1. The total amount of absolute adjustment value of \dot{P} is also shown in the table. All the adjustments are ± 1 unit in Data-2, 3, and 4.

Table 1. Camera motion adjustment

Data	Data-1	Data-2	Data-3	Data-4
Number of adjustment	14	17	14	21
Total amount	21	17	14	21

We also conducted an experiment that clarify the advantage of our method against conservative automatic videography approach. The conservative method uses the same processing loop except that camera adjustment is always issued when it is effective to minimize the position gap in the next loop. The environment of the experiment is the same as those of Data-1 to Data-4. We conducted filming twice, one by our method and one by the conservative method. Note that the situations were not exactly the same for the two experiments although the walker tried to walk in the same way as much as possible. Figure 7(a) shows the positional result of our method while Figure 7(b) shows the same of the conservative method. The acceptable range is set to be ± 0.25 for the position of the object and $L = 10$. These video clips last 25 seconds.

The stastical figures of the experiments are shown in Table 2. Although the position becomes three times looser, it is within the acceptable range for most of the time in the result of our method. Note that the number of the camera motion adjustments is drastically reduced from 23 times to only 4 times. Therefore, the generated video by our method

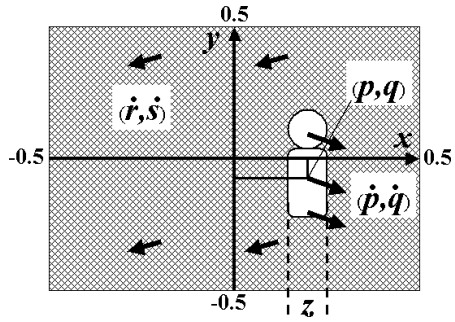
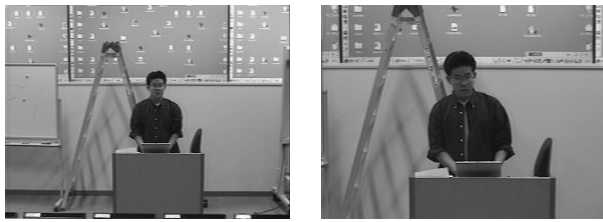


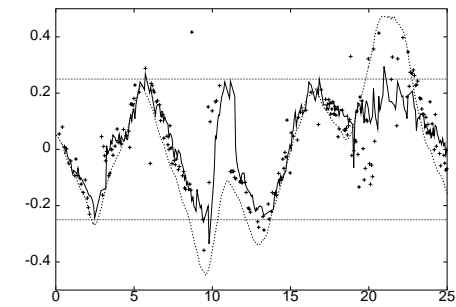
Fig. 1. Vide Composition



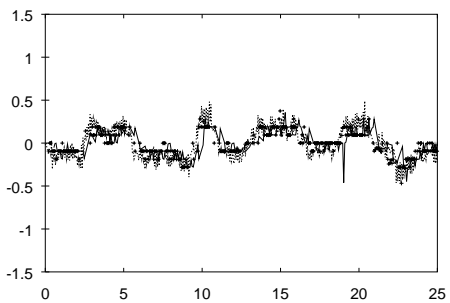
(a) size 0.2

(b) size 0.3

Fig. 2. Object size

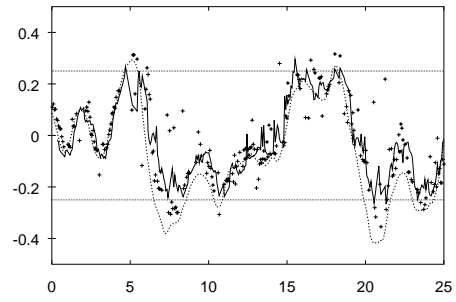


(a) Horizontal position

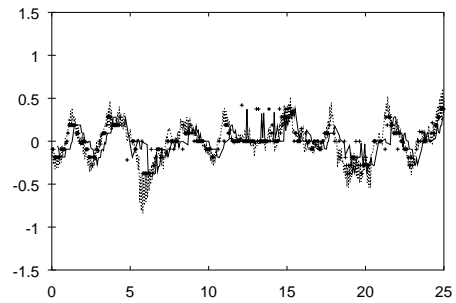


(b) Horizontal speed

Fig. 3. [Data-1] Position range ± 0.25 ; size 0.2

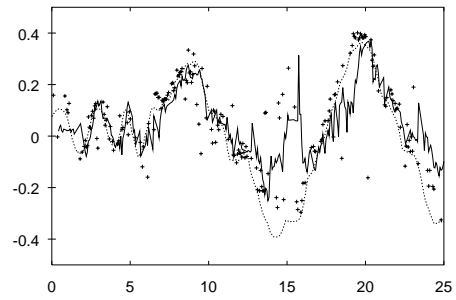


(a) Horizontal position

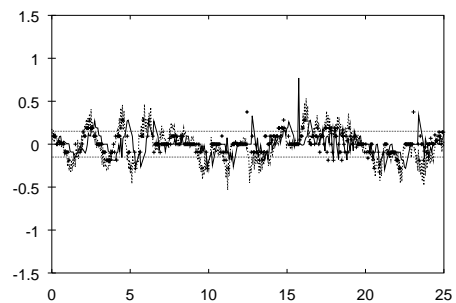


(b) Horizontal speed

Fig. 4. [Data-2] Position range ± 0.25 ; size 0.3



(a) Horizontal position



(b) Horizontal speed

Fig. 5. [Data-3] Speed range ± 0.15 ; size 0.3

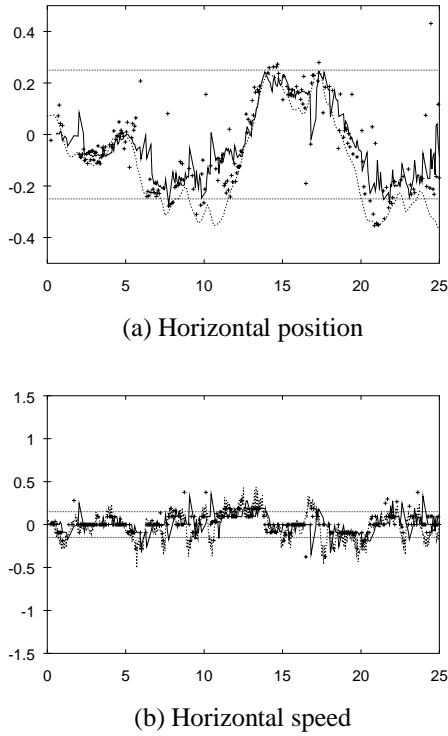


Fig. 6. [Data-4] Position range ± 0.25 ; Speed range ± 0.15 ; size 0.3

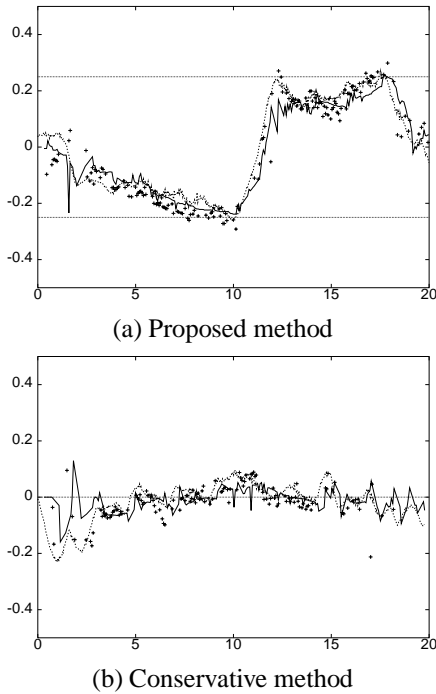


Fig. 7. Adjustment control

is more comfortable than that of the conservative method.

Table 2. Reduction of adjustments

Method	Our method	Conservative method
Time	4	23
Amount	4	24
Mean of Error	0.15	0.05
Deviation of Error	0.066	0.052

In our current implementation, we set Θ_{seg} and Θ_{reg} manually. However, we think optimal values of these thresholds could be determined based on the object size s in TVC.

5. CONCLUSION

In this paper, we have proposed new camera control method that realizes a planned video composition with less adjustments of camera motion. We introduced 'acceptable range of video composition' to ensure the video quality while we succeeded in reducing the adjustments. The results with various acceptable ranges and the comparison with conservative approach reveal the availability of our approach.

6. REFERENCES

- [1] Q. Cai and J. K. Aggarwal, "Real time tracking for enhanced tennis broadcasts," in *CVPR*, 1998, pp. 68–72.
- [2] Q. Liu, Y. Rui, A. Gupta, and J.J. Cadiz, "Automating camera management for lecture room environments," in *CHI*, 2001, pp. 442–449.
- [3] M. Greiffenhagen et al, "Statistical modeling and performance characterization of a real-time dual camera surveillance system," in *CVPR*, 2000, pp. 2335–2342.
- [4] R. Collins et al, "A system for video surveillance and monitoring: Vsam final report," in *Tech report CMU-RI-TR-00-12, Robotics Institute, CMU*, 2000.
- [5] S. A. Hutchinson, G. D. Hager, and P. I. Corke, "A tutorial on visual servo control," *IEEE Trans. Robot. Automat.*, vol. 12, no. 5, pp. 651–670, 1996.
- [6] J. Adachi and J. Sato, "3d visual servoing from uncalibrated cameras for uncalibrated robots," in *ACCV*, 2002, pp. 267–272.
- [7] O. Tahri and F. Chaumette, "Application of moment invariants to visual servoing," in *ICRA*, 2003, pp. 4276–4281.
- [8] M. Ozeki, Y. Nakamura, and Y. Ohta, "Camerawork for intelligent video production – capturing desktop manipulations," in *ICME*, 2001, pp. 41–44.
- [9] Y. Kameda, K. Ishizuka, and M. Minoh, "A study for distance learning service - tide project -," in *ICME*, 2000, pp. 1237–1240.