A Method of Automatic Video Synthesis for Comparing the Player Movements in Ski Competition

Tatsuya OKUYAMA[1)] **Takafumi KOJIMA**[1)] **Tomokazu ISHIKAWA**[1)]

Masanori KAKIMOTO[1)] **Tomoyuki NISHITA**[2)]

1) Tokyo University of Technology 2) UEI Research / Hiroshima Shudo University

G311400735@edu.teu.ac.jp

Abstract

Records of sporting event are being updated every year. We are sure that clarifying the difference between other player movements can assist athletes who want to make a new record. However, it takes time and requires plenty of user efforts to synthesize two video sequences to clarify differences of motions. We address this problem and propose a method to realize an automatic video synthesis. The proposed method provides a new video representation for not only sport athletes but also spectators. Although our ultimate goal is to provide a tool for live coverage of a broad range of sports, we focus, in this research, on broadcast videos of alpine ski competitions. In the proposed method, we first make a rough image adjustment using mask images corresponding to the terrain covered with snow. Then the system synthesizes the two input videos of competing skiers by detecting the gates or flags on the course and aligning the image position precisely for each frame. For exact timeline matching of the two videos, we used the lap time shown as a TV caption. As a result, we were able to compound two videos with a desired timing and position for each frame. A future work is to cancel errors of the synthesizing positions caused by the difference of camera zooming ratios in shooting two competing players of a similar lap time range.

1. Introduction

Recently, video synthesis and image composition have been widely researched in the field of Computer Graphics (CG), which has facilitated the production of interesting visual products. We predict that, by clarifying differences between the movements of athletes, it will be possible to compare player' motion and optimize motion in each sport easily. Conventionally, images are processed frame by frame and composed manually. However, composing images is difficult because video synthesis is a time-consuming process. In addition, there are physical limitations in implementing cameras and setting up apparatuses, such as chroma key composition, because such devices can obstruct players in sport competitions. In this research, we focus on alpine skiing. We assume that players hoping to set a new record by improving their motion, and video producers who need interesting visual effects in presenting the sport will be the primary users of our tool. We aim to realize a fully-automatic video synthesis, without physical limitations due to the camera during the shooting, for use in live coverage. In video synthesis, two videos overlap to form one semi-transparent video.

2. Previous work

Video synthesis has been actively researched. Rüegg et al. [1] proposed a method for cutting out two videos in a seam in order to synthesize the videos. This approach can be applied in many cases moving, there is no need for special movie shooting equipment such as chroma keying. However, rather than being fully automatic, two videos are combined by a user's stroke on an object that is shown in both videos. Because a synthesized video is completely different through the boundary, the video is adjusted by color blending near the boundary. While we set three restrictions that are described later in this research, the proposed method, i.e., brush stroke, is not introduced, rather, the synthesis is performed automatically.

In research involving separating moving objects from a background, Ohta et al. [2] separated a movie into foreground and background by classifying the directions of movement calculated using optical flow [4]. Optical flow is widely used to trace objects [5, 6]. However, it has some problems. Estimation using optical flow can yield false recognition in case of animation wherein intensity barely changes. Further, the method in which images are divided into a grid and feature points are compared exhibits a low processing speed. In the method proposed by

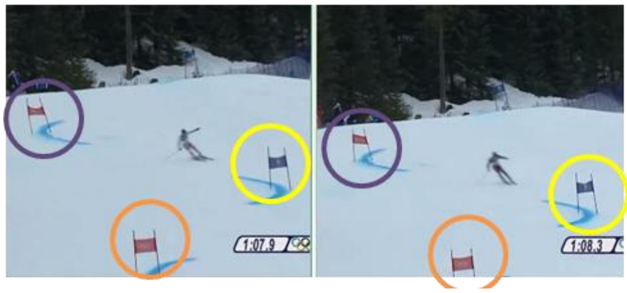


Figure 1. Example of identical objects.

Asaoka et al. [3], which uses robust statistics [7, 8] that fix errors due to outliers using a novel weighted least squares method, moving objects were separated accurately, and the background was estimated from sequential images by determining correspondence among feature points dispersed in the sequential images. However, in our research, corresponding feature points in videos must be extracted because the composition is the same but one is two different videos and they are not the sequential images.

3. Proposed method

3.1 Target videos

We set the following three restrictions on videos rather than setting camera limitation.

(a) Videos are filmed at the same angle.

(b) Videos have a time indicator.

(c) Videos capture the same objects.

(a) is necessary because measurement at different angles causes emptiness. (b) is necessary to determine synthesis frames. (c) is necessary to perform position adjustment using the same object (エラー! 参照元が見つかりません。). We selected alpine skiing because this sport fulfill all three conditions required for synthesizing video. The movement of alpine skiers, and consequently, the panning speed of the camera, is considerably fast. Therefore, detecting the same objects is difficult. Track events are often filmed using a stationary camera and a camera with slow pan speed. Therefore, if we can synthesize alpine ski videos, our method should be applicable to other sports. In this research, we use videos of the 2010 Winter Olympic in Vancouver available on YouTube [9].

Figure 2 shows a flow chart of our method. The input data are two sets of video sequence (video A and B), each of which represents a competing skier for a similar range of lap time. The outline of the flow is to adjust the frame timing of video B, to align each frame of video B, and then to synthesize video A and B with some translucency. The frame timeline matching is fulfilled using lap time recognition for each video, and described in Section 3.3. The image position of each frame in video B is aligned using two steps of object matching between the frame B and the corresponding frame from video A. A detail of the position alignment technique is described in Section 3.4.

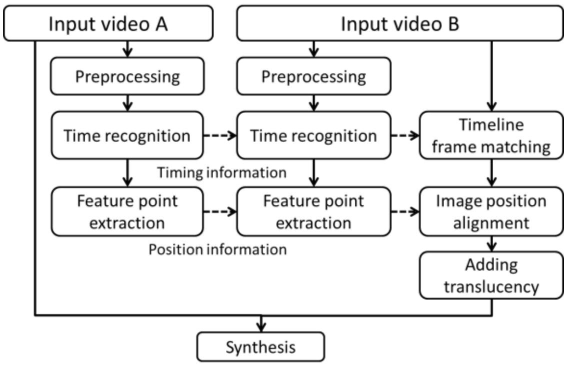


Figure 2. Processing flow.

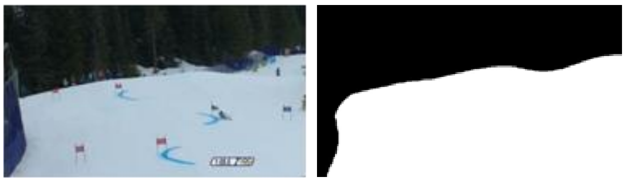

Original image
Mask image

Figure 3. Example of mask image.

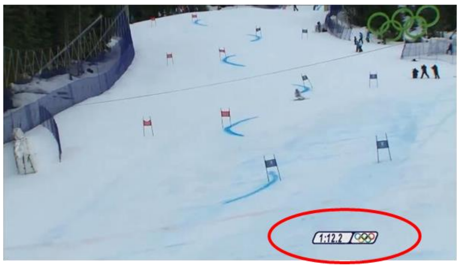


Figure 4. Example of lap time display as a caption.

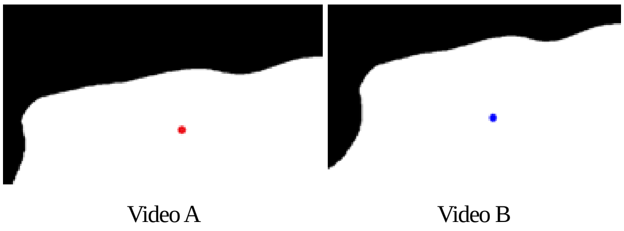
3.2 Preprocessing of videos

Sports videos are large file size. Therefore, we resize videos to manage them with greater efficiency. We calculate a mask image (**Figure 3**) using binarization after applying a median filter to clear noise, excluding the object from the filter.

3.3 Timeline matching

In order to achieve an exact comparison between two players of a speed competition, we need to synchronize their lap times in synthesizing the motion pictures of the players. Fortunately, in relay broadcast of an alpine ski competition, they always provide a lap time displayed as a caption at a corner. Our system uses the lap time for the timeline matching between the two input video sequences of competing skiers. Each number in the caption is recognized with a template matching method [10-12] using small rectangular templates for numbers 0 to 9.

Since the place for the caption is fixed in a relay broadcast, we focus on a specific region of interest to recognize the lap time for each frame, as shown by a red ellipse in **Figure 4**. Our image synthesis is then carried out for frames with the same lap time in the two input videos.



Video A

Video B

Figure 5. Mask median points for images from two input videos.

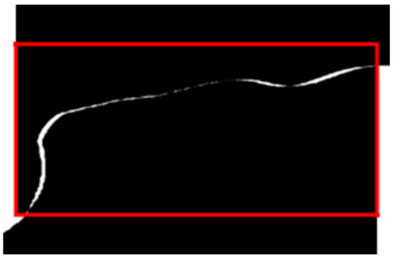


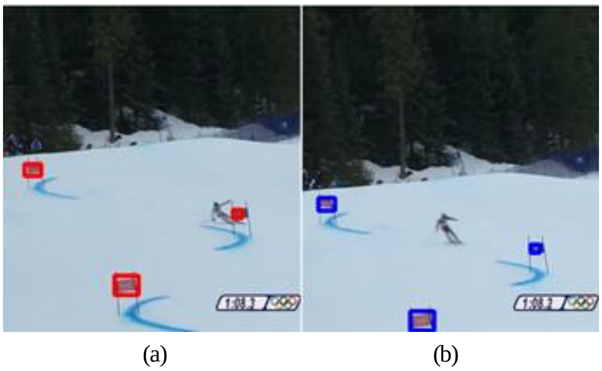
Figure 6. Masks of two input images after a rough shifting operation. The white area represents the difference of the two mask images. The red rectangle is the ROI used for the difference evaluation for image shifting.

3.4 Image position alignment

In our method, each pair of timeline-matched frames from the two input videos is aligned by extracting and recognizing several corresponding feature points. The alignment process is divided into two steps, which are a rough shifting operation using a terrain region as a mask, and a more precise alignment using objects on the terrain.

In order to exclude objects and spectators out of course, we use the mask images generated in preprocessing as described in Section3.3. In addition, the mask image is used to align the two frames to synthesize. First, for each image (A and B), a median point is computed for the mask or the snow terrain region as represented in white in **Figure 5**. Then one of the compared images (Image B) is shifted such that the two median points coincide with each other. The image is further shifted within a certain limited bound, minimizing the difference of the mask images. To appropriately evaluate the difference, we focus on a region of interest (ROI), which is simply the overlapping rectangular region after the shift operations. Within the ROI, the relative difference is computed by dividing the different area (number of pixels) by the ROI area. **Figure 6** describes an example of the ROI area (red rectangle) and the different area (white area in the ROI).

After roughly shifting the image, the system proceeds to a more precise alignment step, which finds corresponding objects in the two images incorporating a simple feature extraction method. For the feature extraction we use gates or poles on the course of the alpine ski competition. A set of gates used in giant slalom events consists of red and blue ones. The same is true for the poles used in slalom events. For giant slaloms, a couple of feature points are extracted using color template matching technique for the gates. As described previously, the search range is limited within the mask region, excluding out-of-course objects such as trees or spectators. Then our system compares the two sets of the extracted feature points from the two video sequences. For each feature point, we find the nearest feature point in the paired image, and evaluate the total distance accumulating for each corresponding feature point pair. Finally, the precise alignment vector is found while minimizing the total distance.



(a)

(b)

Figure 7. Extracted feature points.

4. Results

To implement our method, we used C++ and OpenCV for CPU programing on a standard PC (CPU: Intel® i7-4770 CPU 3.40GHz, RAM: 12.0 GB). All videos used in the experiment had a resolution of 1280 × 720 pixels, a frame rate of 30 fps, and duration of 10 s. The calculation time is 1 min 38 s, 80% of which involved the calculation of feature point extraction. We show the comparison of processing time in case of the proposed method and a manual method using "Adobe After Effect" in Table1. Examinees for synthesis in manual were 10 students studying video processing technique and they studied beforehand operating procedures to obtain a quality equivalent to our synthesis result. We have visually checked the quality of synthesis result produced in manual. We can see that the processing time of our method is shorter than the average time of manual method.

Table 1. Processing time comparison

	Processing time
Our method	1min 38sec
Manual average	2min 51sec

(a) **(b)**

Figure 7 shows results of feature point extraction by template matching. In Fig. 7 (a), the red bounding square indicates a flag extracted as feature point in case of input video A. Similarly, in Fig. 7 (b), the blue bounding square is shown in case of input video B. We can see that the feature points are extracted accurately, and the size of the frame is also supported.

Figure 8 shows the results of synthesis video based on the extracted feature points. Figs. 8 (a) to (c) are three sequential images from 48 frames to 50 frames, (d) to (f) are three sequential images from 148 frames to 150 frames. In Figs. 8 (a) to (c), we succeeded in clarifying the movements of the athletes using the synthesis videos because the positions of the flag in each frame match. The 3 frames shown in Figs. 8 (d) to (f) have some minor deviations caused by the difference in the zoom factor among the frames, although the position adjustment was completed successfully. It is difficult to apply the proposed method in case where there is a significant difference in the zoom factor.

5. Conclusion

In this research, using sports videos, we realized fully automatic video synthesis by template matching. By combining the video of two players, it is possible to compare the players' movements and suggest better motion or technique. The method is also beneficial for people who watch sports because it allows for interesting video effects.

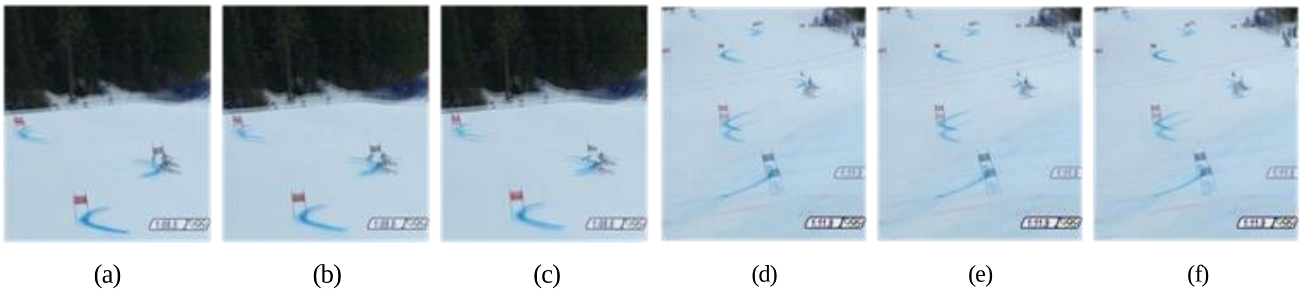


Figure 8. Synthesized results.

Future work will involve increasing the precision of matching and decreasing the speed of calculation so that the proposed method can be used in live broadcast. Furthermore, because the position of the feature point is shifted if the zoom rate is different among the frames being compared, it is necessary to correct the resulting distortion. We will consider straight-line detection as a universal feature value using the Hough transform [13], because we plan to apply the method to competitions other than skiing.

References

[1] J. Rüegg, O. Wang, A. Smolic, M. Gross. "DuctTake: Spatiotemporal Video Compositing" Computer Graphics Forum, vol.32, no.2, pp.51-61, 2013.

[2] N. Ohta. "Structure from Motion with Confidence Measure and Its Application for Moving Object Detection" The IEICE Transactions, vol.76, no.8, pp.1562-1571, 1993.

[3] T. Asaoka, N. Yokoya, H. Takemura, K. Yamazawa. "Motion Detection from Image Sequences with a Moving Camera Using Robust Statistics" IEICE Technical Report. vol.96, no.492, pp.33-40, 1997.

[4] J. L. Barron, D. J. Fleet, S. S. Beauchemin. "Performance of Optical Flow Techniques" International journal of computer vision. vol.12, no.1, pp.43-77, 1994.

[5] T. Iwasaki, T. Yokoyama, H. Koga, T. Watanabe. "Human Extraction Using Optical Flow in Complex Background" IEICE Technical Report. vol.104, no.669, pp.73-77, 2005.

[6] R. Okada, Y. Shirai, J. Miura, Y. Kuno. "Object Tracking Based on Optical Flow and Depth" The IEICE Transactions. vol.80, no.6, 99.1530-1538, 1997.

[7] Y. Sato. "Recent Progress on Using Statistical Models in lmage Processing: Robust Estimation and Minimum Description Length Criterion" Medical imaging technology. vol.12, no.1, pp.48-58, 1994.

[8] P. J. Rousseeuw, A. M. Leroy. "Robust Regression and Outlier Detection" Wiley Series in Probability and Mathematical Statistics, 1987.

[9] "olympicvancouver2010" YouTube. http://www.youtube.com/user/olympicvancouver2010. 2013-01-14.

[10] T. Kaneko, O. Hori. "Update Criterion of Image Template for Visual Tracking Using Template Matching" The IEICE Transactions. vol.88, no8, pp.1378-1388, 2005.

[11] Y. Shinohara, N. Funabiki, J. Kawahima, J. Takeuchi, M. Ishizaki. "A License Plate Recognition Algorithm Using Cross Count and Template Matching Methods" IEICE Technical Report. vol.104, no.573, pp.39-44, 2005.

[12] K. Arai, K. Morimoto, H. Yamana. "Similar Object Detection Using Template Matching Focused on Positional Relationship of Feature Regions" The Special Interest Group Technical Reports of IPSJ, CVIM. vol.172, no.4, pp.1-8, 2010.

[13] D. H. Ballard. "Generalizing The Hough Transform to Detect Arbitrary Shapes" Pattern recognition. vol.13, no.2, pp.111-122, 1981.