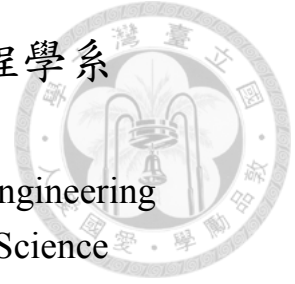國立臺灣大學電機資訊學院資訊工程學系
碩士論文
Department of Computer Science and Information Engineering
College of Electrical Engineering and Computer Science
National Taiwan University
Master Thesis

半管運動電視轉播影片之事件偵測技術
Event Detection in Halfpipe Sports Broadcasting Videos

聞浩凱
Hao-Kai Wen

指導教授：吳家麟博士
Advisor: Ja-Ling Wu, Ph.D.

中華民國 103 年 7 月
July, 2014

國立臺灣大學碩士學位論文
口試委員會審定書

半管運動電視轉播影片之事件偵測技術

Event Detection in Halfpipe Sports Broadcasting Videos

本論文係聞浩凱君（學號 R01922083）在國立臺灣大學資訊工程學系完成之碩士學位論文，於民國 103 年 7 月 26 日承下列考試委員審查通過及口試及格，特此證明

口試委員：

吳家麟

（指導教授）

陳文進

胡敏君　　　　　鄭文皇

系 主 任　　許永真

i

誌謝

首先，特別感謝我的指導教授吳家麟老師，提供美好的實驗室環境，給予我很大的自由可以做直排輪運動的題目，並且悉心教導我如何做研究，帶領我進入多媒體研究領域。我要感謝口試委員，陳文進老師、鄭文皇老師、胡敏君老師。您們指出口試中我敘述不清楚之處，並且給予未來研究方向的建議，使得論文更加完整，

再來，感謝實驗室學長姐們，張嘉祐學長、林裕訓學長、林映孜學姊。您們帶領我進行研究，不厭其煩的與我討論，我由衷地感謝。還有，感謝實驗室同學們，吳政陽、孫家豪、陳厚凱、謝宗廷、周俞荃、陳柏年。與你們討論，使我能夠檢視想法是否有矛盾，也感謝你們分享各自的專業知識，協助我程式寫作技巧技巧。

另外，感謝張偉哲同學在修課階段有機會可以一起合作，完成本研究的初步探勘。

感謝生機系與物治系同學與老師，從不同切入點討論研究，豐富我的看法。

感謝父母養育我，你們無微不至的照料，使我總是感到幸福快樂。也感謝我的大姊、二姐幫忙分攤家中事務，使我無後顧之憂地完成學業。感謝二姊夫在我小時候，教導我寫程式，對我影響甚甚大。

感謝女朋友柯逸凌，給予我鼓勵，時常陪伴我一起做研究。

最後，謹以本文獻給我的家人，還有謝謝所有關心我的人。

摘要

本文中，一個低成本和高效率的系統性自動分析半管運動轉播影片方法被提出來。除了場地顏色覆蓋比率，我們發現玩家區域使用顯著目標偵測機制可面對半管運動影片模糊場景的挑戰。此外，基於現有的 MPEG 壓縮的影片原生的運動向量（motion vector）我們提出一種可用於偵測旋轉（spin）事件的新穎且有效的方法。實驗結果表明，該系統能有效在半管影片中識別難以被檢測到的旋轉（spin）和滑行（grind）等事件。

Abstract

In this work, a low-cost and efficient system is proposed to automatically analyze the halfpipe sports broadcasting videos. In addition to the court color ratio information, we find the player region by using salient object detection mechanisms for facing the challenge of motion blurred scenes in HP videos. Besides, a novel and efficient method for detecting the spin event is proposed on the basis of native motion vectors existing in MPEG compressed video. Experimental results show that the proposed system is effective in recognizing the hard-to-be-detected spin and grind events in halfpipe videos.

Contents

List of Figures

Chapter 1

Introduction

1.1 What is a Halfpipe?

A half-pipe or a vert ramp is an U shaped structure used in gravity sports including snowboarding, skiing, skateboarding, bicycle motocross, inline skating, roller skating, and scootering.

A half-pipe can be categorized into two types: snow, non-snow, such as wood or metal. As shown in Figure 1.1, a common halfpipe includes flat section, two transition sections, two vertical sections or lips, and two decks or tables. The flat section assists a player to maintain balance and increases the transfering potential energy by lowering the body position. The transition section lets the player translating potential energy into kinetic engergy by "pumping" technique. The vert section helps the player control the direction that he or she wants to fly to. The deck is for the player to rest and to drop in the halfpipe. A rail or a coping is on the corner of a vertical section and a deck, which is an object used by players to slide along it by their "vehicles". But, there is no rail structure in a snow halfpipe.

1.2 What is a Trick?

For each player's turn, she or he starts from one side of the halfpipe, i.e. the vert section, passing through the flat section, and reaching to another vert section, and goes back and
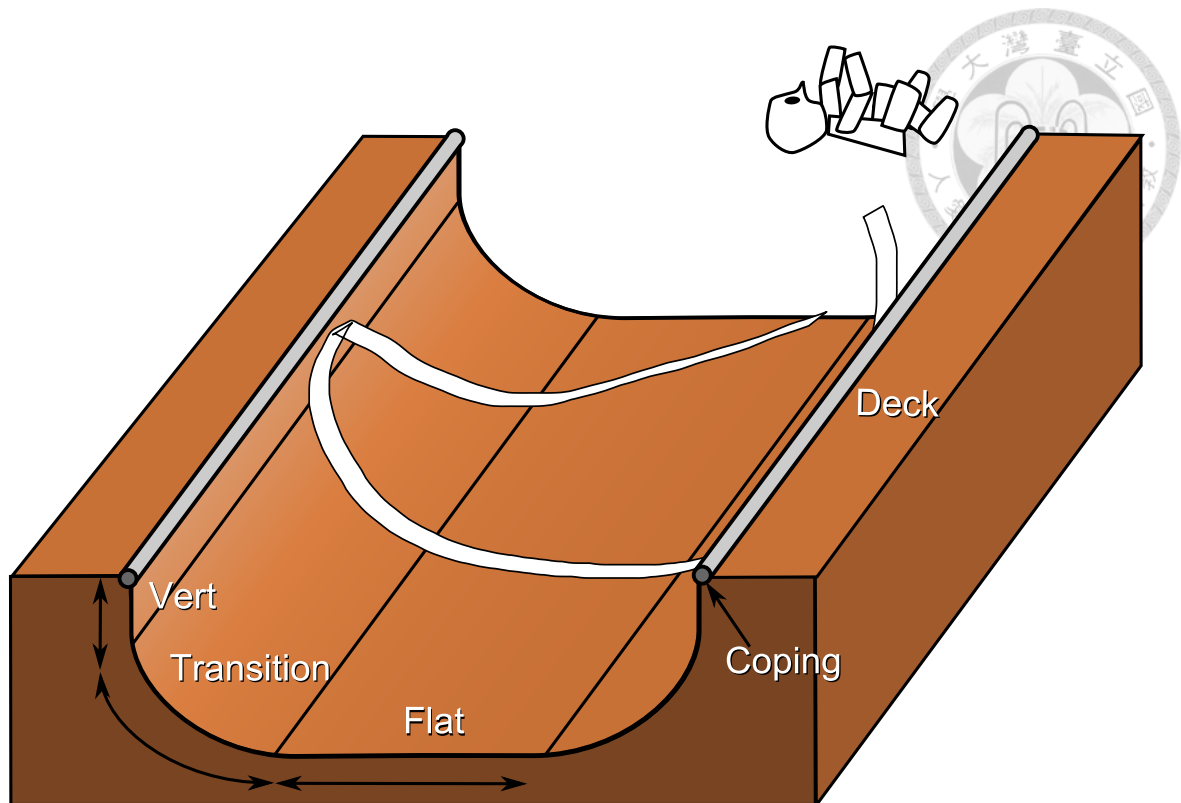
Figure 1.1: Structure of a halfpipe.

forth several times. Each time the player flies above the coping, she or he performs a trick. For convenient, we call that a player passes through the flat section and goes back to the falt section again is a trick segment.

A trick can be categorized into four major type: spins, grind, air/grab, and handplant.

Spin is a major type of tricks in halfpipe sports. As shown in Figure 1.2, a player starts to spin when he is above the coping.

A grind is a trick that involves a halfpipe player sliding along the coping using some part of his or her"vehicle" rather than wheels.

A grab/air is a kind of tricks that a player grabs his or her some part of vehicle while they reach the highest point of a trick. The trick name depends on which part a player grabs, as illustrated in Figure 1.3. Generally, a particular body position held by a player is called a grab/air trick.

A handplant trick looks like a form of a handstand. When a player reaches the coping, he or she flips the body upside-down, and puts his or her hand on the coping for a while, and then goes back to the halfpipe.

2

Figure 1.2: A spin trick.

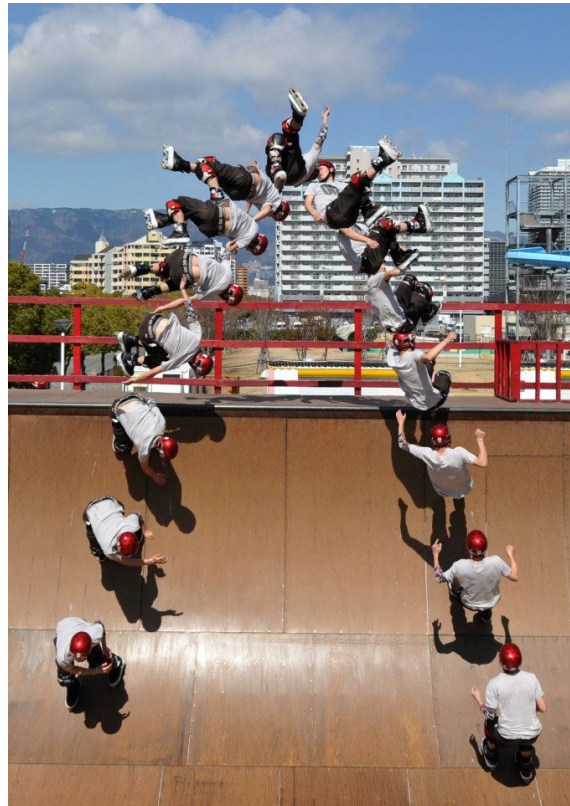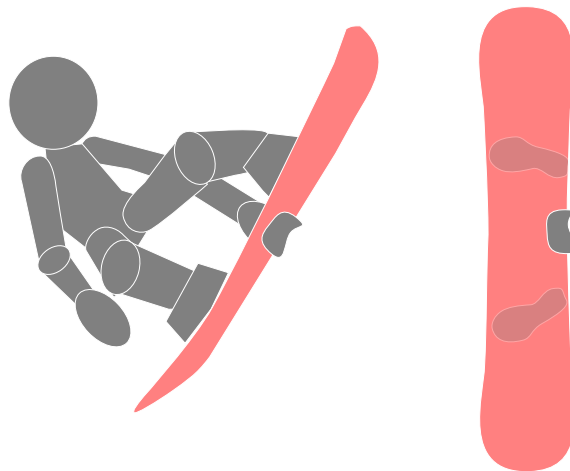Figure 1.3: A snowboard grab.
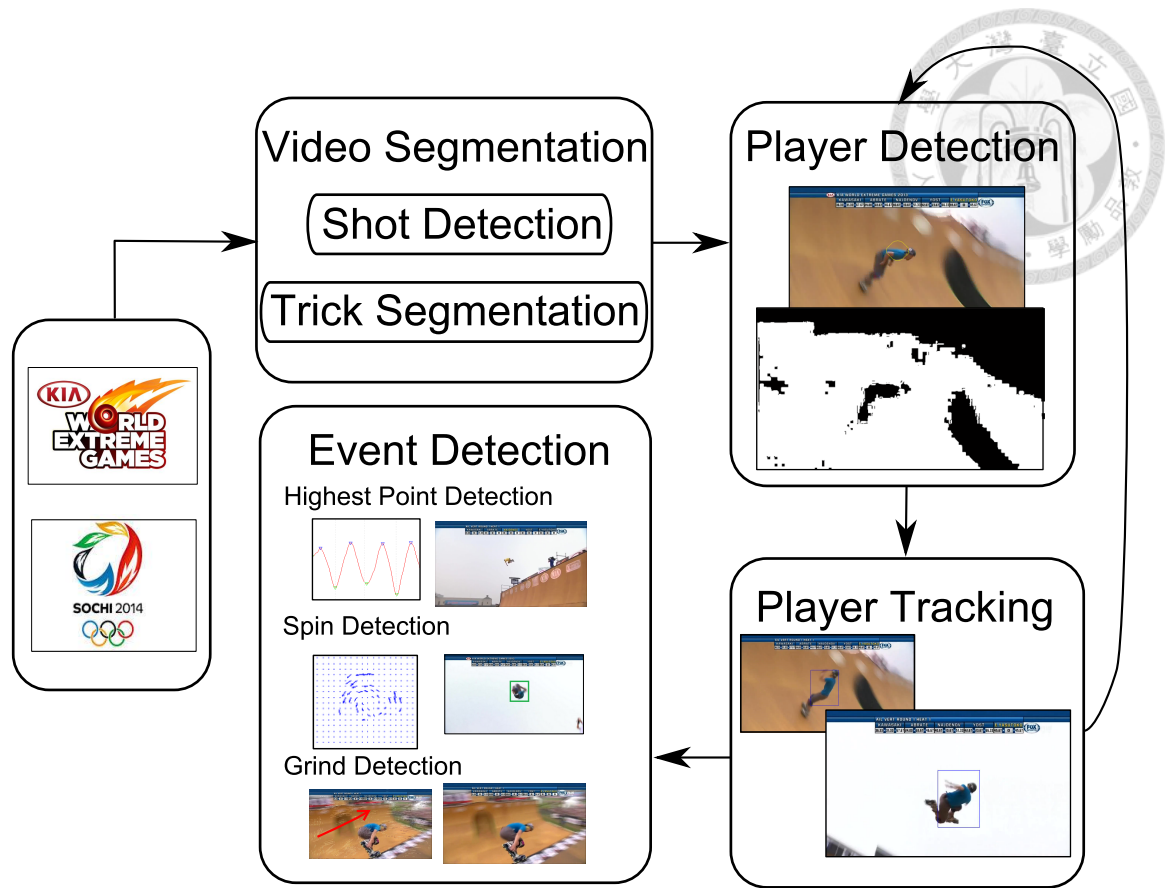
Figure 1.4: System diagram of the proposed approach.

As a result, finding trick related segments and recognizing trick types automatically are highly wanted by and benifical to halfpipe players.

Recognizing the trick type in a halfpipe video is challenging because (i) the scenes in an HP video are motion blurred since players move fast, and (ii) the displacement of players is large in all directions of a 3D space. Due to (i), it is hard to measure the camera motion by using either pixel-based method (e.g., optical flow) or feature-based method (e.g., SIFT-matching). And hence, the 3D scene is hard to be reconstructed. Due to (ii), the background of a scene changes too rapidly to track the player, precisely. Most of the existed studies recognizing the actions of players with the help of sensors which, sometimes, affects the performace of the player seriously.

Therefore, in the proposed system, halfpipe contest videos will automatically be analysed effieciently and effectively, without using sensor information.

The system diagram of the proposed approach is illustrated in Figure 1.4. We will address how halfpipe sports videos become readable with the aid of video segmentation,

player detection, player tracking, and event detection mechanisms, presented in the following sections.

1.3 How to Learn a Trick

Sports learning usually starts from imitation. People find a way to practice a trick by watching sports videos. For halfpipe players, a common way to improve their skills is to learn by imitation. In order to learn the best tricks in the world, they imitate tricks performed by elite players through watching recorded broadcasting videos, such as Sochi 2014 Olympics and KIA World Extreme Games.

A novice player usually learns tricks by watching inline skate magazine and watching videos other than looking for a coach. For one thing, specific couch for halfpipe is hard to find. The other thing is, the skilled halfpipe players are sparse. As a result, readily available learning resources like magazines and videos are effective resource for novice players to quickly get in. Videos provide sequential actions of tricks, as illustrated in Figure 1.5, players can learn a trick step by step from watching the video.

1.4 The Problem of Watching Halfpipe Broadcasting Videos

However, watching an entire sports video is timely, and finding a particular trick in a video is a hard work. First, we do not know the start and the end points of a trick. Second, there are many tricks in a video. Third, people cannot watch the same kind of tricks intensively, because the same kind of tricks are usually very sparse in a video.

It is hard to identify a trick in a video by its name. A trick might be named as many different words of the same meaning, such as front flip 180, Mist flip, and Bio flip 540. Those tricks sound different but actually means the same trick. Another problem is that one name may indicate to more than one trick.

Figure 1.5: Sequential actions of a trick.

1.5 How to Solve Watching Halfpipe Broadcasting Videos Problems

In order to shortening the time lengthy video, we segment a halfpipe broadcasting video into tricks. And we label a trick as a spin or a grind, and people can know the attributes of a trick before watching it. Furthermore, watching the same kind of tricks intensively is affordable.

1.6 Challenges of Event Detection in Halfpipe Videos

Event Detection in halfpipe broadcasting videos is challenging because (i) the scenes in an HP video are heavily motion blurred since the camera moves very fast, and (ii) the displacement of a player is large in all directions of a 3D space. Due to (i), it is hard to measure the camera motion by using either pixel-based method (e.g., optical flow) or feature-based method (e.g., SIFT-matching). And hence, the 3D scene is hard to be reconstructed. Due to (ii), the background of a scene changes too rapidly to be used to track the player, precisely.

6

Chapter 2

Related Work

Most of the existing studies analyzed halfpipe sports by using wearable sensors[2][3]. They can recognize whether a player is spin and calculate the corresponding air-time. The data retrieved by wearable device is highly correlated to mechanics and it can be used to analyze the impact of human body joint during landing[9] and thus be used to prevent sports injury.

However, the best tricks performed by elite players are usually shown in the broadcasting videos. It is hard to let them wear sensors and perform tricks for population. Furthermore, although sensor data helps to injury prevention, one can not learn tricks from sensor data. So we plan to help players learn tricks from another aspect, the sport video content analysis.

The characteristics of competitive diving videos are similar to halfpipe sports videos[5]. Since both players in diving video and halfpipe sports video perform acrobatics, the players' body are non-rigid deformed. The camera is always tracking on a player in both sports because the displacement of the player in the scene is large.

In [5] and [11], they segmented a player from background by using global motion estimation. And [5] further used object segments to recognize player's action by Hu moments (as shape descriptor) and continuous hidden Markov model.

For a diving video, the background is at the same scene depth. This fact is beneficial to the result of global motion estimation and thus the accurate player body segments can be obtained. Regrettably, for a halfpipe sports video, there are cuts in a trick segment, the

backgrounds are at different scene depth all the time, and the player's body take up the too much room of a frame. Even if a player's body is well segmented, the same tricks of a player may look different, because the relative positions and orientations of the camera and the player. So, the approach of [5] is hard to be applied to our work.
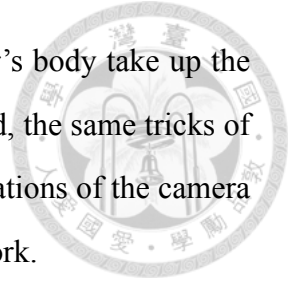
Chapter 3

Proposed Method

3.1 Highest Point Detection and Trick Segmentation

In a run, a player keeps performing tricks and moving back-and-forth on a halfpipe. When the player is on the flat section or transition section, the scene is mainly composed by the halfpipe. And once the player goes above the coping, the percentage of a halfpipe in the scene is getting less and less. Moreover, we observe that the percentages of the halfpipe's occupation are nearly the same when the player flies above drops below the coping. And once a player leaves the halfpipe, he or she will act as a projectile. So the highest point of a trick is occurred in the middle point of a flying and dropping below the coping.

We choose hue in HSV color space to represent the idea of the ratio of halfpipe in a scene. The halfpipe court color coverage ratio of frame t is defined as $r(t) = \sum_p \delta(hue(p) - t)$, where p is a pixel in a frame, and $hue(p)$ is the hue of the pixel p.

However, the contest videos were shot by deck view or ground view and the scenes may switch between them in a trick, which affects the distributions of the court color coverage ratio of a frame.

3.1.1 Noise Elimination

The peaks (i.e., the highest positions of players) may happen several times in between the two verts because the contest video could be shot by ground view or deck view when

the player passes through the flat, which affects the distributions of the peaks of the court color coverage ratio.

To solve the camera view switching issue and the objects with the same color as the halfpipe issue, we apply moving average filter to the halfpipe court color coverage ratio. The span s is chosen as the half duration of a trick to eliminate the noise induced by the two pre-discussed issues.

3.1.2 Peak and Valley Detection

The highest points lie in the valleys between the first peak and the last peak of the filtered result $r(t)$. The peaks are detected at frames x, when $m(x) > m(x+1), m(x) > m(x-1)$, and $m(x)$ denotes the corresponding halfpipe court color coverage ratio at frame x. Similarly, the vallies are located at $m(x) < m(x+1), m(x) < m(x-1)$.

3.2 Player Detection and Tracking

The task of player detection and tracking is quite challenging in halfpipe sports videos since the scene often changes very fast and both the player and the background are quite blurred. Nevertheless, there are always some scenes composed mainly of the player surrounded by halfpipe court during each trick segment, i.e. the scenes with high coverage ratio of court's color. As a result, the connected components in the non-court color parts of the scene stand for player region candidates. We then apply the multi-scale contrast saliency detection proposed by [6] on those scenes for detecting the players, since the player always be the most salient part of those scenes. Finally, the player region can be detected by combining the non-court color region and the salient object region in the scenes of high court color coverage ratio. For the other scenes, we apply the algorithm proposed in [7] to track the position of the player. Since there could be several shot changes during a trick segment, we will re-estimate the player region at each shot change detected based on the difference of color histograms [1][4]. Value in HSV color space is chosen to conduct shot detection.
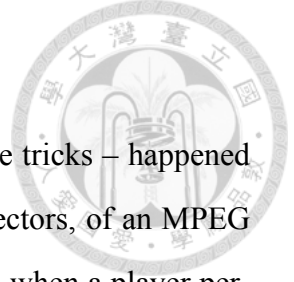
3.3 Spin Detection

We intend to recognize if there is a spin – one major type of halfpipe tricks – happened near the highest point. Since the movements of the native motion vectors, of an MPEG compressed video, will behave like a circle (as shown in Figure 3.1), when a player performs a spin trick. To recognize such a circle, the following mechanism is suggested. Let $A \in \mathbb{C}^{M \times N}$ be the matrix representing the motion field of the player region. We accumulate the vertical components of A along the horizontal axis, $u_j = \sum_{i=1}^{M} \text{Im}(A_{i,j})$, and likewise the vertical axis, i.e. $v_i = \sum_{j=1}^{N} \text{Re}(A_{i,j})$. As shown in Figure 3.1, vectors u and v will across the zero-valued axes, once they reach the center of the circle. However, motion vectors may not form a perfect circle because the rotating movements behind the body of the player cannot be captured by the camera when the rotation axis of the player is nearly perpendicular to the camera view. In such a situation, some part of the circle may be missed, but fortunately, either u or v could still capture the property. So we use the number of zero crossing points neighboring to the frame with the highest point to detect spin event. The frames with the fewer zero crossing points or no zero crossing points the less possible that there is a spin trick in those frames.

3.4 Grind Detection

As shown in Figure 3.2, when a player just goes back and forth, he or she has zero velocity at the highest point, but when a player performs a grind trick, he or she still has horizontal velocity. We use motion intensity[10] and standard deviation of motion intensity features to see whether there is camera motion and the player has velocity at the highest point. Moreover, if a player performs a trick, the halfpipe can be seen all the time in the scene. So, the halfpipe color coverage ratio should not be low at the highest point. The motion intensity is defined as $\frac{1}{||\Phi||} \sum_{\Phi} \sqrt{v_x^2 + v_y^2}$, where Φ is the set of inter-coded macro-blocks and $||\Phi||$ denotes the cardinality of the set Φ. And the variance of motion vector magnitudes is defined as $\frac{1}{||\Phi||} \sum_{\Phi} (\sqrt{v_x^2 + v_y^2} - i)^2$, where i is the motion intensity
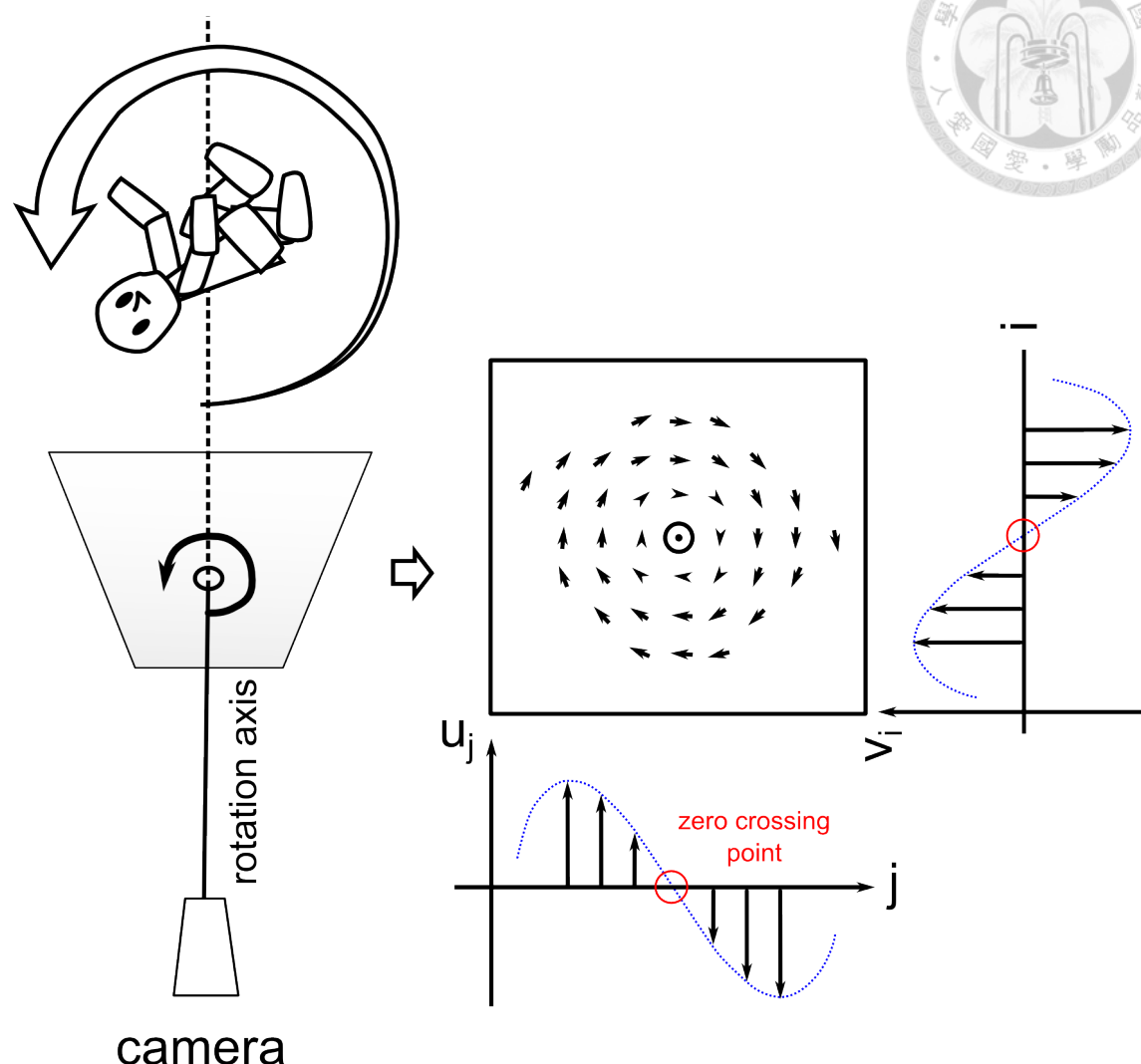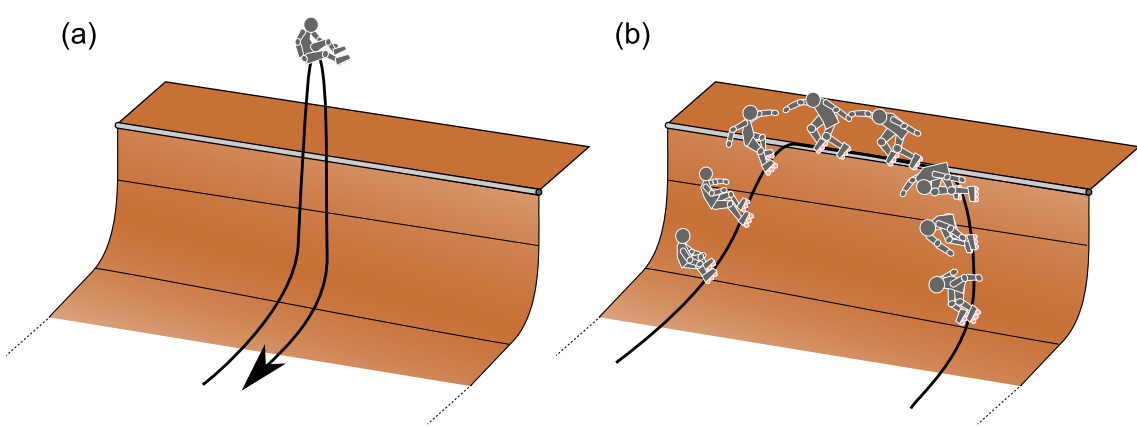
Figure 3.1: Proposed spin event detection.

Figure 3.2: Illustration of (a) air trick and (b) grind trick.

Chapter 4

Results and Discusion

For testing the proposed system, we downloaded "KIA World Extreme Games 2013 In-line Vert Heat 1'' video from Youtube. The duration of the video is 15 minutes. There are 5 players in the contest and each player performed 3 runs. Each run is at most 45 seconds. The video contains 129 cuts, 116 successful tricks, 45 successful spin tricks, and 13 successful grind tricks. The ground truth tricks cuts, and the highest points have been manually annotated.

4.1 Cut Detection

A cut is counted detected if it is within 3 frames difference to the ground truth cut. We choose value in HSV to count the color histogram, and there are 32 bins. The threshold is automatically determined in each run by Otsu's threshold selection [8]. The precision of cut detection is 1.0, the recall is 0.984, and F1 score is 0.992. The corresponding result is illustrated as Figure 4.1.

4.2 Highest Point Detection

A highest point is counted detected if it is within $\pm$ 10 frames difference to the ground truth because the highest point is hard to be determined even by human. And we only count successful tricks. If a player falls and stops moving, the hypothesis of highest point
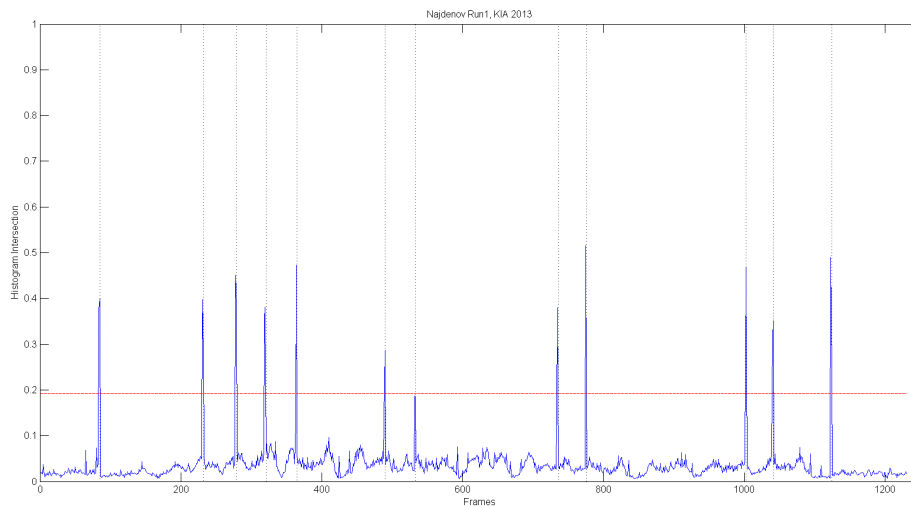
Figure 4.1: The cut detection results in Najdenov run1.

Figure 4.2: A player falls.

detection "a player is always go back and forth" is no longer fulfilled as shown in Figure 4.2. The precision of the highest point detection is 0.956, the recall is 0.931, and F1 score is 0.943. The result of highest point detection is good, but there are still some cases failed as discussed in the following.

As shown in Figure 4.4, the player of this run is the one in the right side of the first picture, but the camera operator did not track him until he performed first trick. This situation happens frequently when the running order of the contest rearranged due to weather or other issues.

Our hypothesis is that once a run starts, the player will go back and forth regularly until the end of a run. But, if a player falls and indeed stops for a while, it does not fit the hypothesis, and the highest point detector will miss the highest point of this trick. We, therefor, chose successful tricks for performing the highest point detection here.

When the shots changed from a ground view to a deck view and back to a ground view within a trick, the color coverage ratios changed from a low level to a high level and then
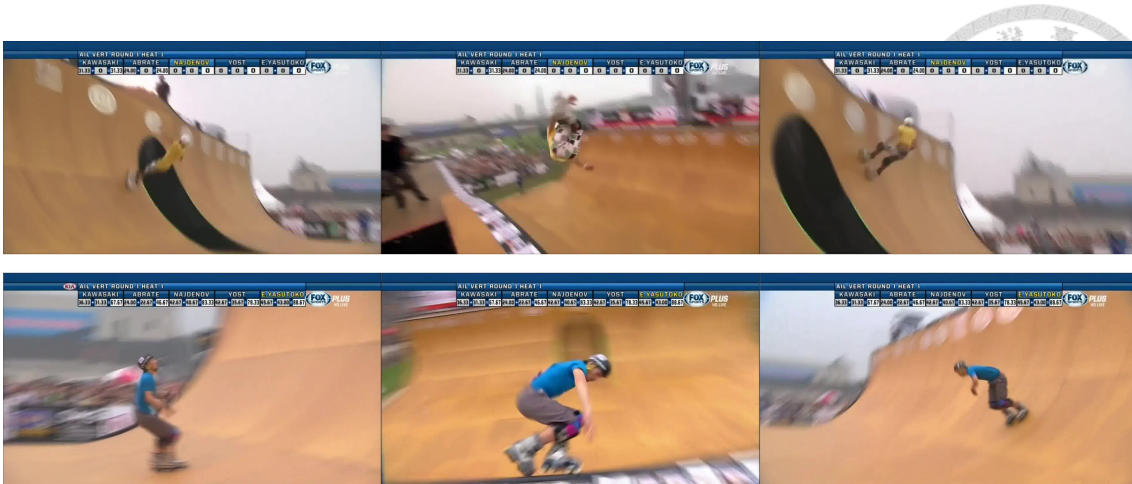
14

Figure 4.3: The highest point of a grind event, and the player make a curve in the air.

Figure 4.4: Camera operator does not track the correct player.

back to a low level. And thus, the valley of the color coverage ratio is far from the highest point of a trick. In our experiment, most of the cases can be solved by using a moving average filter, but some of the highest points could not be annotated at the exact matched frame numbers.

As shown in Figure 4.3, the player performed a grind, the highest point could not be detected. Likewise, as shown in figure, the player made a curve in the air above the coping, and the highest point is failed to be detected. In other cases, if a player flies high enough, the color coverage ratio is low in the deck view shot, and the highest point can be detected. However, making a curve in the air naturally causes the flying height lower than flying straight. So, these cases need be solved in the future.

4.3 Spin Detection

The $\pm$ 10 frames near to the highest point are used to detect a spin event. And only if there are more than 56% of these frames have circles, a spin event is detected. The precision is 0.444, the recall is 0.444, F1 score is 0.444, and the accuracy is 0.569.

Figure 4.5: a slow grind.

There are some miss detected cases are disussed in the following: When a player is relatively small in a scene , the number of motion vectors is not large enough to represent a spin. When the bounding box is smaller than a player, a spin trick cannot be detected. When the rotation axis is nearly parallel to the camera view, motion vectors may not form a perfect circle because the rotating movements behind the body of the player cannot be captured by the camera. When a player spins fast, the player's region will be blur. In this case, the corresponding macroblocks are intra-predicted, so there are nearly no motion vectors in the bounding box for being detect as a spin trick.

There are false alarmed cases as discussed in the following: When the bounding box is larger than a player, motion vectors of background areas may be included. Sometimes, it causes a false alarm. When a player waves his or her hand like a circle at the highest point, the motion vectors form a circle, and it also causes a false alarm.

4.4 Grind Detection

The highest points are assumed to be known in this experiment. The threshold of court color coverage ratio is set to 0.15, the threshold of motion intensity is 4 pixels per macroblock, and the threshold of variance of motion vector magnitude is set to 36. The span of moving average filter for motion intensity and the variance of motion vector magnitude are set to 10 frames, and the span for court color coverage ratio is 41 frames. In the end, the precision for grind detection is 0.430, the recall is 0.889, F1 score is 0.458, and the accuracy is 0.845.

As shown in Figure 4.5, when a player performs a grind but looks like merely locked-in, the grind was not detected, because he did not move to the highest point and thus the
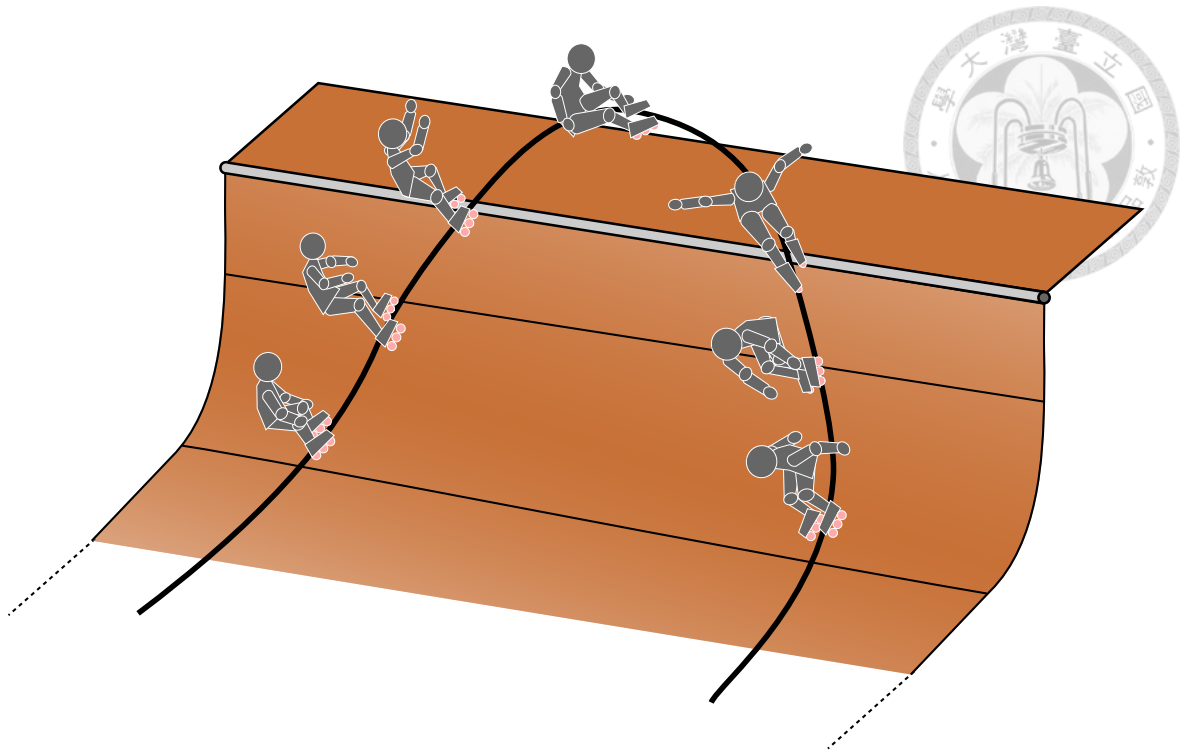
Figure 4.6: The trajectory of a player making a curve in the air.

camera has no motion. However, from learning perspective, a locked-in should be detected because one can learn grind tricks from locked-ins. A slow but long grind could not be detected because the grind detector did not take overall players' motion into consideration but only players' motion around the highest point.

Although the result seems not good enough, we have observed a useful phenomenon from those failed cases, which is the reason why our accuracy severely reduced.

When a player makes a curve in the air, the trajectory looks like the one shown in Figure 4.6; when a player performs a grind trick, the trajectory looks like the one shown in Figure 4.7. If a player flies in the air but not so high when making a curve or a player grinds not long enough, the camera motion of them has no difference, and the grind detector can not distinguish them. This fact explains why the precision is low.

Although making a curve in the air and performing a grind tricks are different, there is a strong correlation between them. A player learns to grind by making several curves just above the coping. After the player feels confidence, he or she put his or her vehicle slightly on the coping. Finally, the player grinds when he or she put his or her vehicle completely on the coping. This means successful detection of making-curve player in the
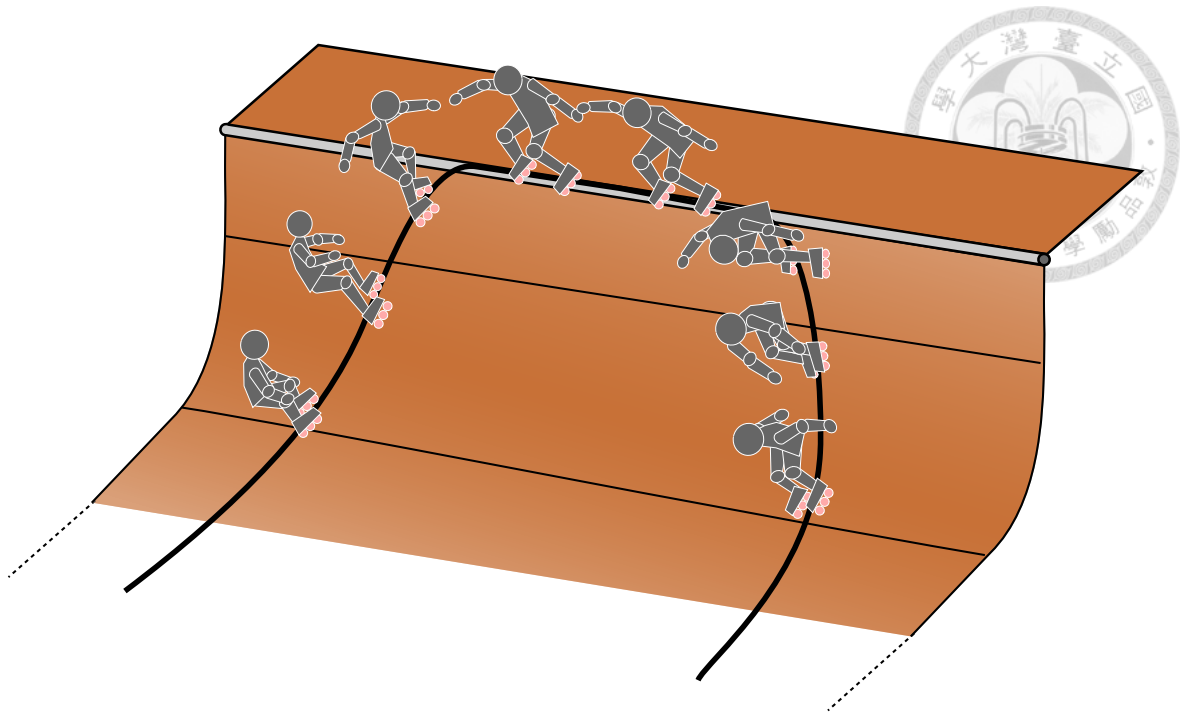
Figure 4.7: The trajectory of a player performing a grind trick.

air is useful for a player who wants to learn a grind trick.

Chapter 5

Conclusion and Future Work

To the best of our knowledge, our approaches about event detection in halfpipe broadcasting sports videos are novel. The main difference between our work and existing one is that we detect halfpipe sports event by analyzing videos instead of using wearable sensors.

In this thesis, we have presented issues in learning halfpipe sports and methods for detecting events in halfpipe sports broadcasting videos.

The proposed method successfully segments runs into tricks, and detects the highest point event, the spin event, and the grind event. With the aid of our system, when a player watches halfpipe sports videos, he or she can jumps between tricks without the need of seeking manually, and knowing the attributes of a trick before hand.

Because the analysis of halfpipe sports videos is very challenging and brand-new, there are lots of directions left for further investigation and we summarize them as follows.

Detection for more complex events: There are still important events in halfpipe sports which do not covered in this thesis, such as player fall, handplant, and grab/air event.

A bit more precise trick attributes. A watcher may not want only to find where spin tricks are but also to know if the rotation axis of the human body of a spin trick is the longitudinal axis, the median axis, or the transverse axis.

Recognition of combined tricks. A player may combines several tricks into a trick. For example, a player may spins into the coping, then grinds on the coping, and again spins off the coping. Some actions are hard to be imagine before watching them in the
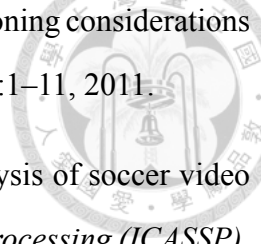
videos. For instance, a player may handplant on the coping, then jump by his hand, and grind on the coping. How to detect this series of tricks successfully is very challenging but useful for a halfpipe sports learner.
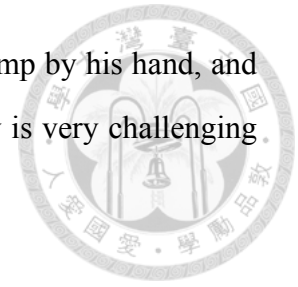
Bibliography

[1] U. Gargi, R. Kasturi, and S. H. Strayer. Performance characterization of video-shot-change detection methods. *Circuits and Systems for Video Technology, IEEE Transactions on*, 10(1):1–13, 2000.

[2] J. Harding, C. Mackintosh, A. Hahn, and D. James. Classification of aerial acrobatics in elite half-pipe snowboarding using body mounted inertial sensors. In *The Engineering of Sport 7*, pages 447–456. Springer Paris, 2008.

[3] T. Holleczek, J. Schoch, B. Arnrich, and G. Troster. Recognizing turns and other snowboarding activities with a gyroscope. In *Wearable Computers (ISWC), 2010 International Symposium on*, pages 1–8. IEEE, 2010.

[4] I. Koprinska and S. Carrato. Temporal video segmentation: A survey. *Signal processing: Image communication*, 16(5):477–500, 2001.

[5] H. Li, J. Tang, S. Wu, Y. Zhang, and S. Lin. Automatic detection and analysis of player action in moving background sports video sequences. *Circuits and Systems for Video Technology, IEEE Transactions on*, 20(3):351–364, 2010.

[6] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, and H.-Y. Shum. Learning to detect a salient object. *IEEE trans. PAMI*, 33(2), 2011.

[7] J. Ning, L. Zhang, D. Zhang, and C. Wu. Robust mean-shift tracking with corrected background-weighted histogram. *Computer Vision, IET*, 6(1):62–69, January 2012.

[8] N. Otsu. A threshold selection method from gray-level histograms. *Automatica*, 11(285-296):23–27, 1975.

[9] J. Turnbull, J. W. Keogh, and A. E. Kilding. Strength and conditioning considerations for elite snowboard half pipe. *Open Sports Medicine Journal*, 5:1–11, 2011.

[10] L. Xie, S.-F. Chang, A. Divakaran, and H. Sun. Structure analysis of soccer video with hidden markov models. In *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, volume 4, pages IV–4096. IEEE, 2002.

[11] J. Zhang, J. Qiu, X. Wang, and L. Wu. Representation of the player action in sport videos. In *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2013 Asia-Pacific*, pages 1–4. IEEE, 2013.