Doctoral Disseration

A Video Analysis for Camera Motion Estimation and its Application to Automatic Retrieval of Attractive Moments in Sport Videos

Pawin Prasertsakul

Supervisor: Prof. Hiroyuki Iida

School of Information Science
Japan Advanced Institute of Science and Technology

September 2018

Abstract

Videos are main sources of information and entertainment. They are presented in sequence of visual and audio information. For entertainment purposes, entertaining videos (e.g. sport videos) are made by video makers in order to entertain viewers whose cannot watch at the stadium because of their limitations. In order to access the video in a short time, several researchers begin to make an automatic system that can index attractive moments in the entertaining videos based on the human perspective. By the ideas of uncertainty in game information and motions in computer vision, a new research in the study area of information science is established. The contributions of this research are discussed in the two chapters of this thesis:

Chapter 3 presents a new algorithm of computer vision model to make computers understand the camera motions in each video frame automatically. To understand the camera motions, a 2D motion vector histogram is used instead of 1D motion vector histograms as described in existing works. The properties and behavior of the 2D motion vector histogram are analyzed in order to recognize the camera motions. Compare with 1D motion vector histograms, it shows that the 2D motion vector histograms can recognize more types of the camera motions.

Chapter 4 presents a mathematical model to show how the attractive moments can be retrieved by the camera motions. Based on the idea of changing in game information, the attractive moments are potentially occurred when the game information is changed. Since video makers notice the attractive moments, they operate the cameras in order to guide the viewers for attentions. From all camera motions, zooming camera motions potentially retrieve several attractive moments in soccer games. For example, score attempting, foul, player claim to the referee's judgment, etc. Finally, we generate a shortening video application based on this idea.

Keywords: Attractive images, Camera motions, Computer vision, Image processing, Video analysis.

Acknowledgements

First of all, I would like to express my appreciation to my supervisor Professor Hiroyuki Iida. He supported me several things during Ph.D. academic life in Japan Advanced Institute of Science and Technology, Japan. He gave me important guidelines how to write a good research article and also gave me instructions how to live in Japan.

Second, I would like to gratitude to all committee members including of: Associate Professor Ikeda Kokolo, Associate Professor Hasegawa Shinobu, Professor Kazunori Kotani, and Associate Professor Toshiaki Kondo. They gave several valuable suggestions in order to improve strengthens of this doctoral dissertation.

Third, I would like to thank you to Associate Professor Waree Kongprawechnon and Professor Thanaruk Theeramunkong whose introduced The Sirindhorn International Institute of Technology and Japan Advanced Institute of Science and Technology Ph.D. collaboration program to me when I was studying in Master Degree program at Sirindhorn International Institute of Technology, Thailand.

Finally, I would like to thank to my family for supporting me in both direct and indirect ways. I also would like to thank to all volunteers whose are joined in subjective evaluation to accomplish my research. All achievements cannot be done without them.

Contents

This dissertation was prepared according to the curriculum for the Collaborative Education Program organized by Japan Advanced Institute of Science and Technology and Sirindhorn International Institute of Technology, Thammasat University.

List of Figures

List of Tables

Chapter 1

Introduction

Videos are media sources that have plentiful information in the form of picture and sound. There are several purposes of usage. For example, education, entertainment, commerce, personal, etc. In the world, all videos are created by video makers in order to entertain viewers. The viewers can feel free to choose and to watch them based on their behaviours. After the viewers had watched the videos, they can feedback to the video makers "how do they feel about the videos?". Some videos may entertain the viewers while some videos may not entertain the viewers. From the first sight, several videos look very similar to each other but a few differences in contents may make the videos more entertaining.

Nowadays, there is a massive number of videos from several video makers. All video makers have to compete with other video makers to make videos which attract the viewers. For sport games, the video makers have to create a full video by recording all moments in order to prevent the missing important content. The full video is also used in TV live broadcasting. Some viewers are entertained by the full video but some viewers are not because of their behaviours. Thus, the video makers improve their videos in order to satisfy the viewers by editing the full video for a shorter video that summarizes only the attractive moments. In the past, an automatic system to find attractive moments in videos (e.g. sport videos) was developed [4]. This automatic system assists both video makers and viewers to find the attractive moments in videos automatically. For video makers, they can do video editing easier or spend a few time. For viewers, they are guided to access the attractive moments directly instead of watching the entire video. This automatic system follows the human perspective from simple questions such as "when do the

viewers feel exciting?" and "when do the viewers feel entertaining?". In details, it describes that motions in pictures have an effect to describe the videos [5, 6]. This system can be improved in order to perform a better solution to find the attractive moments.

To discuss the entertainment in several study areas, uncertainty has been widely used as a factor in the attractiveness following human perspective [7]. The uncertainty considers both unpredictable and hardly predictable moments to be attractive moments. In sport games, uncertainty in game outcome becomes an important element of all kinds of uncertainty which directly describe "how entertaining and attractive this match is?" [8, 9]. Basically, statistical data in the sport videos and recorded data is used to calculate the uncertainty in game outcome and used to measure the attractiveness. The viewers are not attracted by the entertaining games or videos which the outcome is easy to predict predictable [10]. Game scores are used to calculate the winning, losing, and draw probabilities [11–13]. These probabilities have the relationship to the uncertainty in game outcome. If a game has high winning or losing probabilities, it can refer to less entertaining because the viewers can predict the game outcome before the game is ended. In contrast, if a game has low winning and losing probabilities, it can refer to high entertaining because the viewers cannot predict the game outcome until the end. More examples, the number of score attempting and the number of game scores are used to measure the speed of game progress information in order to find the game progress patterns corresponding to each entertaining game [14, 15].

Generally, the uncertainty in game outcome expects that the viewers feel entertain because changing in game scores makes the viewers cannot predict the game result. Unfortunately, the attractive moments except changing in game score are not involved in the uncertainty in game outcome. We are curious to know that how to involve them as the attractive moments. Guironnet et al. mentioned that "we think that camera motion carries important information on video content" [3]. Therefore, this research is dedicated to develop a new approach which combines the concepts of uncertainty and motions in picture.

1.1 Background

In the past, there are several studies in order to understand the viewer's actual feeling when they are watching the videos. Basically, the viewers will be asked several questions directly, for example, "is this video enjoyable?", "when do you feel exciting?", "how attractive in video is?", etc. By these questions, the researchers can understand what kind of attractive moments can attract or entertain the viewers. It also helps the video makers to make more attractive videos and more videos. In this section, we discuss the background of this research direction. It divided in two directions: 1) attractive moments in pictures and 2) attractive moments in uncertainty.

1.1.1 Attractive Moment Identification in Pictures

In a past decade, a human emotional space "arousal-valence" were presented in the psychophysiology study area [16, 17], which arousal describes the strength of emotion while valence describes positive and negative emotions. This space was offered for psychologists to present a structure of human affective experience. The arousal-valence emotional space in Figure 1.1 roughly describes a set of human emotions in the arousal-valence space including of enjoy (high arousal and high valence), distress (high arousal and low valence), bored (low arousal and low valence), and relax (low arousal and high valence). The paradigm of the

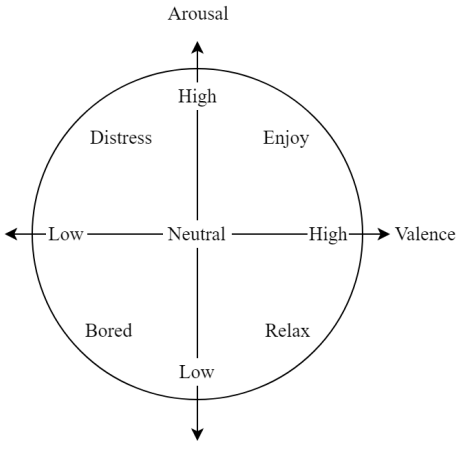


Figure 1.1: A set of human emotions in the arousal-valence space

arousal-valence spaces was synthesized in order to extract human emotions from videos [18–20]. The arousal-valence spaces are generated by motions features in

the videos. From psychophysiological studies, they had shown that motions in video frames have a significant impact on individual affective responses [5, 6]. In details, the arousal in the arousal-valence spaces is directly related to the motions in picture. For example, high motion activity expects for strong emotion while low motion activity expects for weak emotion. In video content analysis, the arousal is used for an application such as highlight video extraction [4] and video ranking [21]. Their applications mainly expect that high motion activity has the attractive moments in the videos.

The viewers can be attracted by the visual features in spatial and temporal domains. In spatial domain, human eyes are strongly attracted to the still image by contrast in color (e.g. red-green and blue-yellow) [22, 23] and special image texture such as human face [24, 25], while the human eyes are attracted to the image sequences by the contrast in motion and moving objects in temporal domain [26]. Shih et al. used these concepts in order to rate the attention score in each video frame [27, 28]. From this idea, the attention score in each video frame can be used for applications, such as video summarization [29] and video keyframes selection [30].

Comparing between the spatial and temporal domains, the temporal domain is more interested than the spatial domain. The viewers can acquire attractiveness passively from the camera motions in videos which are made by video makers. The video makers often guide the viewers by operating cameras. For example, the video makers operate the cameras to follow a soccer player who is dribbling the ball in soccer game. Them, the viewers are attracted by this moment that what will happen in the future. From this viewpoint, a guide of viewer attention was presented by using the camera motions as attractive magnifiers [31]. The camera motions affect toward viewer's attention since the human eyes focus on an area of video frame partially. A study on the effect of panning, tilting, and zooming camera motions was presented [32]. In this study, viewers' eyes are tracked to monitor how much effort do the viewers pay attention while they are watching videos in several situations. From this study, the zooming camera motions make the viewer pay attention more than the panning and tilting camera motions.

1.1.2 Video Contents Analysis for Attractive Moments Identification

In other study areas, the attractive moments in sport videos are examined by analyzing contents in videos. For example, game scores in sport videos are used to calculate three probabilities of game outcomes: 1) probability that Team A will win, 2) probability that Team A will lose, and 3) probability that the game is drawn [11–13]. By using these probabilities, the attractive moments can be predicted when the value in the probabilities is changed. There are three game progress patterns which are described in [14, 15]. They are called

- Balanced game: Both of the teams have no goal through the game.
- Seesaw game: One team leads, then the other team leads, and this happen repeatedly alternate. However, it is necessary that the difference in game score between the two teams should be small.
- One-sided game: The game score of one team is always greater than that of the other team. However, it is further divided into complete one-sided and incomplete one-sided games. If the difference in game score is very high, it is the complete one-sided game. Otherwise, it is the incomplete one-sided game.

1.2 Problem Statements

As mentioned at above, the uncertainty in game outcome expects that the viewers feel entertain with attractive moments in games when the game result is changed. It is a natural sense in the human perspective that the viewers can enjoy and attract to the videos because of unpredictable results. However, there are other attractive moments which are not involved by the uncertainty in game outcome. These non-involved attractive moments can be found by the motions in picture. The attractive moments, which the pictures contain high motion activity, are expected to entertain the viewers based on the synthesized human emotional space [4]. There is another way to use the motions in picture to find the attractive moments. A camera motion is one of the motions in picture which the global motions (i.e. motions in background image) are translated into more meaningful

factors. The camera motion can be involved to find the attractive moments in the videos. As described in [32], a study on the effect of camera motions shows that the zooming camera motions make the viewers pay more attention when they are watching the videos. In sport games, it is curious that why the video makers operate zooming camera motions rather than moving the camera to follow a player. It may consider that the zooming camera motions are related to the attractive moments due to an uncertainty situation. Thus, the camera motions are investigated in order to find the attractive moments following problem statement and research questions.

Problem statement: When making an automatic system to retrieve the attractive moments in the videos, it is needed to find features that have a relationship of the attractive moments. In the past, high motion activity in picture is used to find the attractive moments. However, it may not cover all attractive moments base on the human perspective. Thus, we investigate more about the motions in picture in order to improve the performance. There are hints from [3, 32] which mention that the camera motions have a relationship to the attractive moments. In human eyes, the camera motions can be realized easily but it is difficult for computers. Moreover, the human can understand the attractive moments while the computers cannot understand. Therefore, we give two research questions for this study.

Research question 1: How to design a model that can extract the camera motions from the video?

Research question 2: How can the attractive moments be retrieved by the camera motions?

1.3 Structure of the Thesis

In Chapter 2, related works of this study are reviewed and explained in details. In Chapter 3, steps of camera motion extraction are introduced. We use several video categories which have different behaviour in contents. In results, both single and multiple camera motions are extracted from the video. In Chapter 4, an automatic retrieval of attractive moments in sport videos is presented. We use soccer video for this study. Chapter 5 we give the final conclusions and future works of this dissertation that come from all related experiments.

Chapter 2

Literature Review

2.1 Introduction

In order to find attractive moments in the entertaining videos, there are two questions: First, "How to design a model that can extract the camera motions from the video?". Second, "How the attractive moments can be retrieved by the camera motions?".

For the first question, there are two types of models that are used for extracting the camera motions. First, parametric models compute projective transformation parameters between two consecutive video frames [33–38]. The parameters refer to horizontal and vertical velocities of image pixels from one video frame to its next video frame. In MPEG video domain works [39–41], the motion vectors (MVs) in predicted frames (P-frames) and interpolated bi-directional frames (B-frames) are accessed directly to retrieve the projective transformation parameters which can reduce computational times.

Second, non-parametric models are more flexible than the parametric model because they use textures or edges [42] instead of parameter values of image pixel to estimate the camera motions. Template matching in optical flow simply find the camera motions by dividing the video frames into sub-images equally [43], or by observing four sub-image regions at the corners [44, 45]. The dominant MVs in each sub-image lead to the camera motion pattern. From the past until now, 1D MV histograms are popular and useful tools in non-parametric models to extract the camera motions from the videos [1, 2, 46–51].

For the second question, Guironnet et al. mentioned that "We think that camera motion carries important information on video content. For example, a zooming camera motion makes spectator attention to focus on a particular event" [3]. There is a study on the camera motion effects [32]. The viewer's eyes are tracked to monitor how much effort do viewers pay attention while they are watching the videos. From the study, the zooming camera motions make the viewer pay attention more than the panning and tilting camera motion.

In this chapter, we review existing works that are related to this research.

2.2 Camera Motions Extraction Using 1D Motion Vector Histograms

2.2.1 Prerequisite

In order to extract the camera motions, it is essential to retrieve motion features from the videos. Basically, MVs can be extracted or can be accessed to the videos directly since the videos are represented in visual distribution, such as color, shape, texture, etc. The block-based motion estimation is a simple and fast approach for estimating the MVs.

To estimate the MV in the video frame at current time t, the previous video frame at time $t-1$ is used as the reference frame in order to investigate how the image pixels move from video frame $t-1$ to video frame t. Figure 2.1 summarizes the process of the block-based motion estimation. First, both current video frame t and its previous video frame $t-1$ are divided into blocks without overlapping equally. Then, the block-based motion estimation starts to find the MV in each block position. At the n-th block position of video frame t, the search area is set on the previous video frame $t-1$ where the center of the search area is at the same block in video frame t. Finally, a search window, which has size the same as the block, is used to scan inside the search area in the raster-scan order. Then, the image in the search window and the image in the n-th block position are compared. If they have the best matching, the MV is estimated. Otherwise, look for the next location. From the explanation above, it spends long times if it is performed by full search algorithm. Thus, we give an example of fast

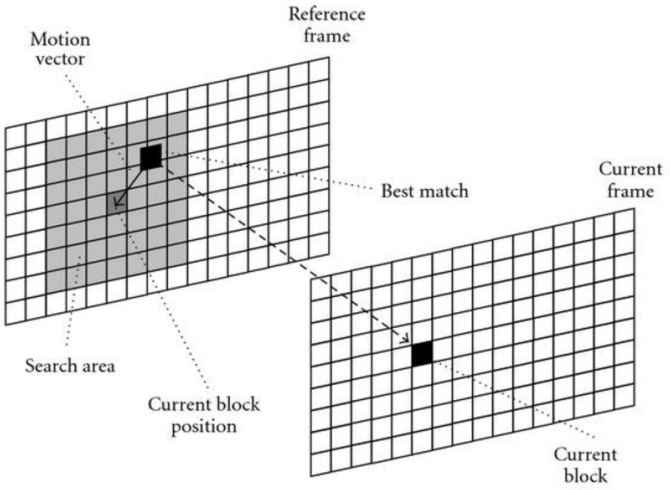


Figure 2.1: Process of block-based motion estimation

block-based motion estimation. Adaptive rood pattern search (ARPS) is the fast block-based motion estimation [52]. The ARPS has overall performance better than three steps search [53], four steps search [54], simple and efficient search [55], diamond search [56], and cross diamond search [57–59] because ARPS uses neighboring blocks as a clue to find the best match block instead of finding entire video frame. Using neighboring blocks makes the MV estimation faster because it can skip unnecessary searching points. From raster-scan order, there are four available neighboring blocks as shown in Figure 2.2. Note that the block with the ∘ symbol represents the estimating block while the gray color blocks represent the neighboring blocks. To estimate accurate MVs, the four neighboring blocks

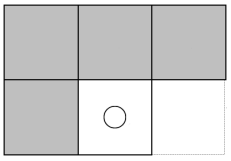
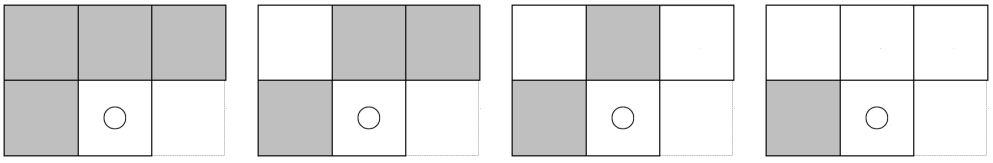


Figure 2.2: Four positions of reliable neighboring blocks

are used as support regions. There are four types of support regions following in Figure 2.3. Note that the block with the ∘ symbol represents the estimating block, the gray blocks represent the selected neighboring blocks for the support region, and the white blocks represent non-selected neighboring blocks for the support region. The support region type A uses all neighboring blocks for the support region (Figure 2.3(a)). It spends the longest time processing comparing with the other types. The neighboring block at the left, upper, and upper-right positions are used for the support region type B (Figure 2.3(b)). This support region type

is used in motion estimation of video coding standard H.263 [60]. The support region type C is simpler than both type A and type B (Figure 2.3(c)). It includes the only two neighboring blocks in horizontal and vertical positions. Finally, the support region type D processes the fastest in motion estimation comparing with the other types (Figure 2.3(d)). It uses the previous block of the raster-scan order (i.e. the left neighboring block) as the support region. The ARPS uses the support type D for the fast block based motion estimation. The ARPS has two steps of



(a) Type A (b) Type B (c) Type C (d) Type D

Figure 2.3: Four types of support regions

motion estimation. Figure 2.4 briefly explains the procedure of the ARPS. First, the ARPS tries to locate the starting point of searching inside the searching area following below.

- The center position of the searching area.
- The two pixels left off-center position.
- The two pixels right off-center position.
- The two pixels upper off-center position.
- The two pixels below off-center position.
- The same position as the estimated MV in the support region.

Note that the default range of off-center position is set by two pixels for estimating the first MV. When the ARPS estimates the next MV the range of off-center position is adjusted by refer to the longest displacement MV in horizontal and vertical directions. In this step, the ARPS can skip the unnecessary searching points. Since the sixth position can be one of the five first position, there are at least five positions for the ARPS to find the starting point of searching. Then, the ARPS computes the sum of absolute differences (SAD) in all starting points of searching. The searching point with the lowest SAD values is set for the new center position of searching.

Second, the ARPS refines the searching by repeating the first step with new four off-center positions. The new off-center positions are located at one pixel of left, right, upper, and lower off-center positions. Then the ARPS computes all SAD values and compares to each other. If the lowest SAD value is located at the center of the searching, the MV is found. Otherwise, the ARPS sets the new center position of searching at the lowest SAD value and repeat this step.

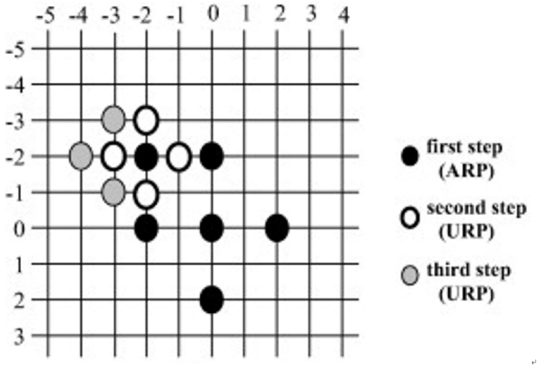


Figure 2.4: Procedure of the adaptive rood pattern search

2.2.2 Camera Motion Histogram Descriptor (CAMHID)

The camera motion histogram descriptor (CAMHID), a histogram based approach for video shot characterization, is accomplished by Hasan et al. [1]. Figure 2.5 shows the framework of CAMHID. There are four steps of the procedure. First, the MVs are estimated from the video. Second, The estimated MVs are filtered for the MVs of interest. They directly refer to the MVs in the background image. To find the MVs of interest, two directional gradients of MV displacement, horizontal and vertical directions, are computed following Eqs. 2.1 and 2.2.

$$\nabla u_{(x,y)} = \{\nabla u^1_{(x,y)}, \nabla u^2_{(x,y)}, \ldots, \nabla u^n_{(x,y)}\} \text{ where } \nabla u^i_{(x,y)} = u^{i+1}_{(x,y)} - u^i_{(x,y)} \tag{2.1}$$

$$\nabla v_{(x,y)} = \{\nabla v^1_{(x,y)}, \nabla v^2_{(x,y)}, \ldots, \nabla v^n_{(x,y)}\} \text{ where } \nabla v^i_{(x,y)} = v^{i+1}_{(x,y)} - v^i_{(x,y)} \tag{2.2}$$

where $\nabla u_{(x,y)}$ and $\nabla v_{(x,y)}$ are the sets of horizontal and vertical gradients of MV displacement in each block position (x, y), respectively. Both $\nabla u^i_{(x,y)}$ and $\nabla v^i_{(x,y)}$ are the horizontal and vertical gradients of MV displacement in i-th video frame of each block position (x, y), respectively.

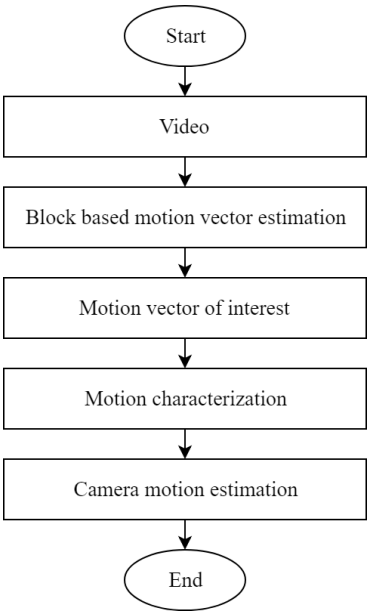


Figure 2.5: Framework of the camera motion histogram descriptor

To identify the MV of interest, a statistics-based traditional measure of distance is used. The consistency of the MV of each block position is investigated. If the MV in the block position (x, y) has the low values of $\nabla u^i_{(x,y)}$ and $\nabla v^i_{(x,y)}$, then the motion is declared as the MV of interest (i.e. MVs in the background image). Otherwise, it is not an MV of interest (i.e. MVs in the foreground image).

Third, the CAMHID computes the average magnitude of the interested MVs for n consecutive video frames in order to handle the unnoticeable small motions. By the average magnitude, the CAMHID can recognize no motion accurately. For the directions, the CAMHID divides the angular range into 12 ranges equally (Figure 2.6). Fourth, the CAMHID divides the MV field on the video frame into

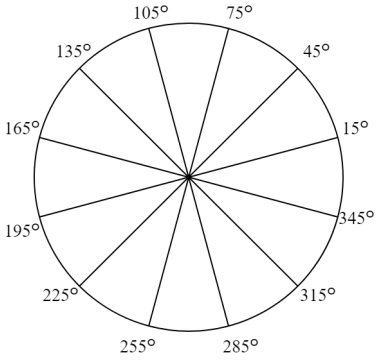
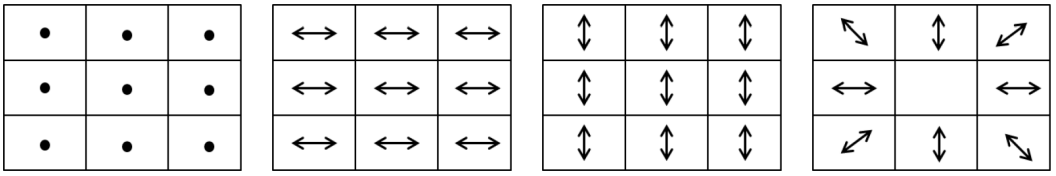


Figure 2.6: Angular ranges for motion vectors estimation in CAMHID [1]

3×3 sub-field equally. In each sub-field, the CAMHID finds the dominant MV by

using two histograms of MV magnitude and MV orientation. Figure 2.7 shows four dominant MV templates of the camera motions. For each template, each sub-field contain the dominant MV which represents the camera motion pattern. For stationary camera motion (Figure 2.7(a)), all sub-fields contain zero dominant MVs. Panning camera motions have the horizontal dominant MVs (Figure 2.7(b)) while tilting camera motions have the vertical dominant MVs (Figure 2.7(c)). Finally, zooming camera motions have diagonal dominant MV at the corners of sub-fields, horizontal dominant MVs at the left and right sub-fields and vertical dominant MVs at the upper and lower sub-fields (Figure 2.7(d)). The CAMHID considers the camera motion in video frames by choosing the best matching to the templates.



(a) Stationary (b) Panning left and right (c) Tilting up and down (d) Zoom-in and out

Figure 2.7: Four templates of camera motion patterns

2.2.3 Learning Based Camera Motion Characterization Scheme

Okade et al. recognized the camera motions by using two 1D MV histograms [2]. After MV estimation, the magnitude and orientation of MVs are calculated following in Eqs. (2.3) and (2.4).

$$m = \sqrt{u^2 + v^2} \tag{2.3}$$

$$\theta = \arctan(\frac{v}{u}) \tag{2.4}$$

where m is the MV magnitude, θ is the MV orientation, u is the displacement of the horizontal MV, and v is the displacement of the vertical MV.

After computing both MV magnitude m and MV orientation θ, they are distributed by using 1D MV histograms. In MV magnitude histogram, it has zero motion and non-zero motion. This histogram mainly uses for indicating stationary and non-stationary camera motions. In MV orientation histogram,

it has eight angular ranges which are described in Figure 2.8. From Figure 2.8, the instructions at below describe each camera motion property following Okade's method [2].

- Stationary: The dominant MV orientation is located in the 1st range and the dominant MV magnitude is located at zero motion.
- Panning left: The dominant MV orientation is located in the 4th and 5th angular ranges.
- Panning right: The dominant MV orientation is located in the 1st and 8th angular ranges.
- Tilting up: The dominant MV orientation is located in the 2nd and 3rd angular ranges.
- Tilting down: The dominant MV orientation is located in the 6th and 7th angular ranges.
- Zooming: All MV orientations are spread on all angular ranges.

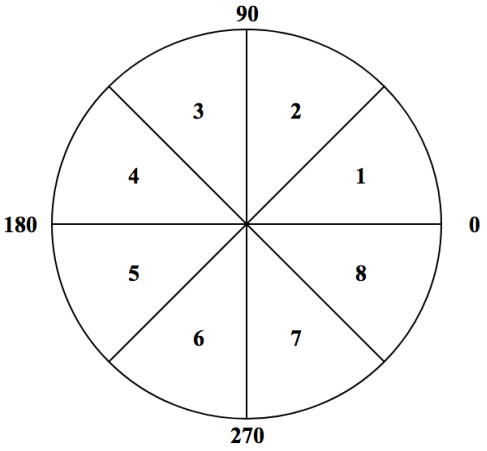


Figure 2.8: Angular ranges for motion vectors estimation in Okade's method [2]

Then they designed the three levels of procedure in order to classify the camera motions following Figure 2.9. At the first level, the coefficient of variation in Eq. (2.5) is computed. This coefficient of variation is used for estimating zooming camera motions.

$$cv = \frac{\sigma}{\mu} \tag{2.5}$$

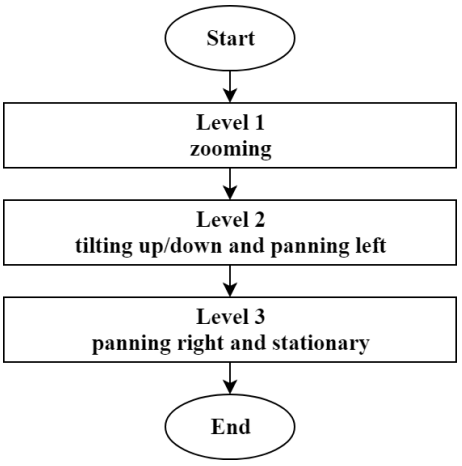


Figure 2.9: Three levels of Okade's procedure [2]

where cv is the coefficient of variation between the orientation ranges, and σ and μ are the standard deviations and the average of all orientation range counts respectively.

If cv has a low value, the MV orientation are separate evenly in all orientation ranges as the zooming camera motion property. Otherwise, the MV orientations are accumulated in one of the eight ranges which need to be investigated at the next level.

At the second level, the remaining camera motions can be classified by using the dominant in the 1D MV orientation histogram. From Figure 2.9, it is obviously to classify panning left, tilting up, and tilting down camera motions. Because of their design, the stationary and panning right camera motions are considered in the final level.

In the third level, the 1D MV magnitude histogram is included to distinguish between the stationary and panning right camera motions. If the dominant MV magnitude is located at the zero motion, it is stationary camera motion. Otherwise, it is panning right camera motion.

Table 2.1 summarizes the characteristics of all camera motions based on the three-level classification.

Table 2.1: Characteristics of all camera motions in Okade's method [2]

Camera motions	cv value	The location of dominant MVs	
		MV orientation	MV magnitude
Station	high	1st angular range	zero motion
Panning left	high	4th and 5th angular ranges	non-zero motion
Panning right	high	1st and 8th angular ranges	non-zero motion
Tilting up	high	2nd and 3rd angular ranges	non-zero motion
Tilting down	high	6th and 7th angular ranges	non-zero motion
Zooming	low	all angular ranges	non-zero motion

2.3 Attractive Moment Identification Using Motions in Picture

According to the usage of motions to identify the attractive moments in pictures, there are several existing works from several study areas. In this section, we show how the motions can be used for retrieving the attractive moments in pictures and videos.

2.3.1 Affective Impact of Motions in Picture

In the past decade year, several efforts to extract and to estimate the human moods from videos are proposed. Motion features in the videos are useful to indicate how viewers feel when they are watching videos. The motion features can be accessed easily and can be used in the psychophysiological study area. The motions in picture were used to study an affective impact of viewers. Note that the affective impact is represented by human emotional space "arousal-valence" [5, 6].

Detenber et al. designed an experiment to explore the relationship between picture with motion and its emotional responses [5]. They were expected that the motions are leading the viewer's excitement. In their experiments, they asked subjective people of average 20 years old to watch still images and moving images. Then, they directly rated the emotional response via heart rate and skin conductance in two relationships: 1) motion and arousal (i.e. strength of emotion), and 2) motion and valence (i.e. a sign of emotion). This experiment was also extended by including the facial electromyography at Corrugator supercilii muscle (i.e. the muscle near the eyes) and Zygomaticus major muscle (i.e. the muscle near cheek) [6]. The purpose of this experiment is to confirm the significant effect of motions in picture

on the arousal and to investigate the relationship between motion in pictures and the valance. Based on their experimental results, they concluded that:

- The motion directly affects to the arousal level.
- The motion does not affect to the valance level.
- It was suggested that motion may capture and sustain attention as well as influence certain aspects of emotional responses.

From their conclusions, they lend us to know that the motion can determine the strengthness of emotion. For example, high motion activity in the picture strengthen the emotional impact.

2.3.2 A Paradigm in Human Emotional Space Using Video Components

Hanjalic had started a new direction in video content analysis by synthesizing a human emotional space "arousal-valance" using motion and sound features [18]. This work was inspired by the study area of psychophysiology [16, 17]. The arousal model indicates the intensity or level of emotion while the valence model indicates positive and negative emotions. Hanjalic used both motion and sound features to synthesize the arousal model and used only sound feature to synthesize the valence model. Since we had known from the previous section "The motion directly affects to the arousal level", thus we focus on how Hanjalic synthesizes the arousal model.

In the arousal model, Hanjalic considers the function $G_i(k)$ that can change the arousal value over time k by each i-th feature component. The function $G_i(k)$ can be interpreted as one component of the arousal model which is described in Eq. 2.6

$$A(k) = G_1(k) + G_2(k) + \cdots + G_i(k) \tag{2.6}$$

where $A(k)$ is the arousal value at time k and $G_i(k)$ is an i-th feature component at time k.

Since Hanjalic used motion and sound as the feature components of the arousal model, i in Eq. 2.6 is equal to 2. The first component, motion features are computed in form of activity in MV magnitude of each video frame. Hanjalic

expected that the image with high motion activity has high excitement. Eq. 2.7 describes the motion activity in mathematical expression. Note that this component has value in a range between 0% and 100%.

$$m(k) = \frac{100}{max(mag)} \times (\sum_{i=1}^{B} mag_i(k))\% \tag{2.7}$$

where $m(k)$ is the motion activity at time k, $max(mag)$ is the longest MV magnitude, $mag_i(k)$ is the i-th MV magnitude in at time k.

The second component, loudness of sound is included in the arousal model. Hanjalic also expected that loud sound has high excitement. Eq. 2.8 describes the sound energy in the mathematical expression. Note that this component also has value in the range between 0% and 100%.

$$s(k) = 100 \times e(k)(1 - \frac{1}{W}\sum e(k))\% \tag{2.8}$$

where $s(k)$ is the sound component at time k, $e(k)$ is the normalized sound energy in scale between 0 and 1, and W is the constant parameter.

From Eq. 2.6, the feature component $G_i(k)$ were defined as the weight of feature components.

$$G_1(k) = w_1 m(k) \tag{2.9}$$

$$G_2(k) = w_2 s(k) \tag{2.10}$$

where $G_1(k)$ is the weight of motion component and $G_2(k)$ is the weight of sound component.

Therefore, Eq. 2.6 is rewritten in Eq. 2.11.

$$A(k) = w_1 m(k) + w_2 s(k) \tag{2.11}$$

where $A(k)$ is the arousal at time k, $m(k)$ is the motion activity at time k with weight parameter w_1, and $s(k)$ is the sound component at time k with weight parameter w_2. Note that the summation of the two weights w_1 and w_2 must equal to 1.

Hanjalic uses his arousal model in order to find the most attractive moments in videos [4]. In his conclusion, using the motion and sound features for finding the attractive moments in videos is appropriate. However, the performance of the

approach needs to be increased with better features inside the video. By human perspective, some attractive moments can occur when less motion and less sound energy. Thus, they may lack meaning. Note that all Hanjalic's ideas of this direction are summarized in [61].

2.3.3 Video Summarization of Attractive Moments Using Camera Motions

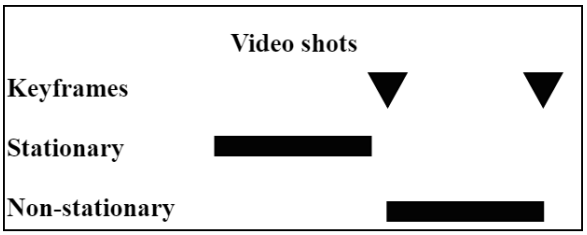
In order to make viewers access videos easily, video indexing technique using visual features such as motions have been developed. From these techniques, it provides new applications such as video summarization and video browsing.

The video summarization is a method that shortens the original video. It composes of important video frame's contents which are called keyframes. The keyframes selection aims for representing the important contents inside videos. Since the method in the previous section expects that the attractive moments are in high motions and loud sound, it may lack some cases in human perspective. To retrieve the attractive moments in videos, camera motions can be included. A study on the effect of panning, tilting, and zooming camera motions was presented [32]. In this study, the viewer's eyes are tracked to monitor how much effort do viewer pay attention while the viewer is watching a video in several situations. From the study, the zooming camera motions make the viewer pay attention more than the panning and tilting camera motions. Guironnet et al. mentioned that "We think that camera motion carries important information on video content. For example, a zooming camera motion makes spectator attention to focus on a particular event" [3]. Then, they introduced a rule based method to make a video summarization by using the camera motions.

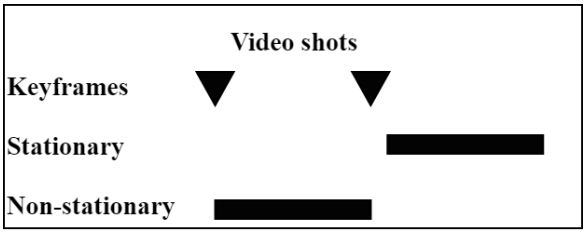
First, they split the video into several video shots which each shot contains only one camera motion. Then, heuristic rules are defined (Figure 2.10):

- If two consecutive video shots are represented in stationary and non-stationary (i.e. panning, tilting, and zooming) camera motions, the two video frame at the beginning and the end of the non-station video shots are selected as keyframes (Figures 2.10(a) and 2.10(b)).

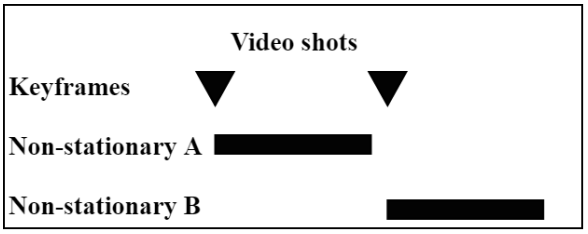
- If two consecutive video shots are represented in non-stationary camera motions, a video frame at the beginning of each video shot is selected as keyframes (Figure 2.10(c)).



(a) Video shots from stationary to non-stationary



(b) Video shots from non-stationary to stationary



(c) Video shots from non-stationary A to non-stationary B

Figure 2.10: Keyframes selection in Guironnet's heuristic rules [3]

This heuristic rules process two consecutive video shot at an iteration. Figure 2.11 shows an example of keyframes selection using these heuristic rules. The video contains three sequences of video shots: 1st) stationary, 2nd) panning, 3rd) stationary. By applying these heuristic rules, the video frames at the beginning and at the end of the second video shot are selected as keyframes. Finally, the two selected keyframes are used as video frame indexes in order to create a short version video that has attractive moment. From Figure 2.11, the original video is shortened to the second video shot as the attractive video.

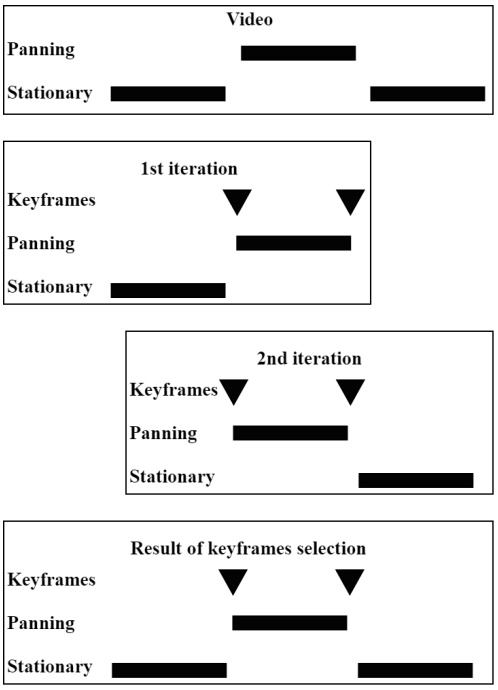


Figure 2.11: Example of keyframes selection using Guironnet's heuristic rules [3]

2.4 Thrill in Attractiveness of Games

In this section, we discuss "what is the attractive moments in other study areas?". Three masters model was designed to study an inner meaning of games which focuses on games solving to know its true properties, feeling senses, and uncertainty [62]. The three masters model reveals that the attractiveness of game relates to a harmony in fairness, judges, and thrill. The three masters model corresponds to the three important characteristics: competitiveness (i.e. the master of winning), entertainment (i.e. the master of playing), and communication (i.e. the master of understanding). This model relates to the game theory [63], game refinement theory [64, 65], game information dynamic [66], etc. From the three masters model, the master of playing is the closest to this research because the master of playing mainly focuses on the game progress while the master of understanding and the master of winning mainly focus on an algorithm to solve the game and player's skills respectively. From the three masters model, the game's uncertainty elaborates on the entertainment which is described by subsections below.

2.4.1 Entertainment and Uncertainty

Uncertainty is the key-factor to measure the interesting rate of a game because people are not attracted by the game itself but the unpredictable or the uncertainty in the game [10]. The relationship between entertainment and uncertainty had been discussed and analyzed in psychology. In each game-playing, there are three kinds of uncertainty which can be characterized by

- Uncertainty in winning strategy: it strictly concerns about a strategy for playing a game. It relates to the difficulty of the best strategy in order to win the game.
- Uncertainty in information: people are attracted by the games because of unpredictable game results. The predictions which are made by the viewers are more neutral than the predictions which are made by players, since the viewers are not involved in the game directly and there are no psychological affects in judgment.
- Uncertainty in game theoretical value: it represents the game outcome of simulated games. For example, it assumes that all players have the same skill level. This uncertainty directly relates to game's rules.

These three kinds of uncertainty are similar to the three masters model [62]. From the human perspective, the uncertainty in information is important to measure the attractiveness in games because it was shown by a report that changing in information has an impact on entertainment during game-playing [67]. In the report, non-attractive games usually are predicted easily because all actions in the games never change the game result. In contrast, fascinating games are attractive when an action in the games changes the game result. Therefore, people are attracted by these games because they are hard to predict the game results until the final action.

2.4.2 Changing in Information and Uncertainty

There are several games and entertaining videos around the world. Some of them are less attractive and some of them are highly attractive. By observing roughly,

they are similar to each other. Because of some different parts of the games, it may make them become attractive games.

In the past decade, entertaining game properties are examined by Majek and Iida [8]. They developed a framework "the uncertainty in game outcome" to exam two players zero-sum games that are enjoyable. The uncertainty in game outcome is based on the concepts of information theory. There are several reasons that can describe "why people are attracted by the games". One of them is the winning trend: if the games are hard to predict "which team will win?" until the games are end, they have a good balance of winning tread and can be considered as the attractive games. From the uncertainty in game outcome framework, it considers the popular games (e.g. chess and soccer) as testbeds. They are played by novice players whose have similar skill levels. It is important that if the games are not attracted by the novice players, the games are not surviving until the present day. For chess, the uncertainty in game outcome can be analyzed by the distance of pieces' positions from the current position to the end position. For soccer, it is difficult to analyze the distance of the game position because of large data size. Therefore, the uncertainty in game outcome is analyzed by using the statistical data during the game-playing.

2.4.3 Changing in Game Outcome and Entertainment

From the uncertainty in game outcome framework, a mathematical model of game refinement was presented [65, 68]. This model explores a game progress information to measure the entertainment. The realistic game progress information is created as a non-linear function to represent that the game information is unpredictable. The second derivative of the game progress information is derived to find the acceleration in the sense of information dynamic. This acceleration is similar to the acceleration in the second Newton's law [69]. The acceleration on the game progress information, which is called "game refinement value", relates to emotional impacts in human minds. The game refinement value is mainly measured by using the statistical game data. For example, a number of score attempting and a number of game score in soccer games. By these statistical data, the uncertainty in outcome can be measured. If the game refinement value is low (i.e. slow in game progress information speed), the game outcome is hardly changed and the games become less interesting because of predictable result.

In contrast, if the game refinement value is high (i.e. fast in game progress information speed), the game outcome is easily changed and is hard to predict the game result. This mathematical model is also similar to the probabilistic in excitement which is presented by Vecer [12].

2.5 Conclusions

In this chapter, we review related works start from how to extract the camera motions from the videos, and how to use them for retrieving the attractive moments. In step of extracting the camera motions, both CAMHID and Okade's method used MV magnitude and MV orientation histograms to analyze the camera motions in each video frame. CAMHID spends a long time for filtering the MVs in background before identify the camera motions. It may not necessary to use template matching to identify stationary, panning, and tilting camera motions because they are not complex as same as zooming camera motions. Okade's method is a fast camera motion estimation method. It can identify each camera motion via its coefficient of variation. However, it cannot distinguish especially zoom-in and zoom-out camera motions. For this part, we see another way to extract the camera motions from the videos which is described in Chapter 3.

In the past, there is an evidence about the relationship between motions in picture and human emotional impacts. It shows that the motions in picture have an effect in strength of human emotional impact. By inspiration of the study in psychophysiology, Hanjalic had synthesized the human emotional space "arousal-valence" and used it for finding the most attractive moments in videos. However, it may lack some cases in human perspective. Thus, the performance needs to be increased with better features inside the videos. A sentence which is mentioned by Guironnet et al. "We think that camera motion carries important information on video content". Then, they introduced a simple rule based method to find the attractive moments in video and to summarize as a short video version. Their rules especially select video frames with non-stationary camera motions as keyframes that represent the attractive moments. From the study on the effect of several camera motions, zooming camera motions make the viewers pay attention more than the other camera motions. Thus, the idea of using camera motions to find the attractive moments is investigated.

Chapter 3

Camera Motions Extraction Using 2D Motion Vector Histogram

This chapter is an updated and abridged version of the following publication.

- P. Prasertsakul, T. Kondo, and H. Iida. Video shot classification using 2D motion histogram. In *2017 14th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*, pages 202-205, June 2017.
- P. Prasertsakul, T. Kondo, H. Iida, and T. Phatrapornnant. Camera operation estimation from video shot using 2D motion vector histogram. *The Visual Computer* (Submitted April 2018).

3.1 Introduction

The camera motions have an important role in several applications. For example, retrieving a particular video shot, editing an entire video sequence, and encoding a sequence depending on the significance of each scene. In this chapter, we explain how to extract camera motions from videos. The target camera motions which are extracted by our proposed method comprise of stationary (S), panning left (PL), panning right (PR), tilting up (TU), tilting down (TD), diagonal panning

(D), object tracking (T), zoom-in (ZI), zoom-out (ZO), scene change (SC), and combination of zooming (e.g. ZI+PL, ZI+PR, ZO+PL, ZO+PR, etc.) camera motions.

In the past, several researchers used parametric models to classify camera motions in videos [33–38]. Computing projective transformation parameters in consecutive video frames is a traditional approach. In MPEG video domain works [39–41], MVs in predicted frames (P-frames) and interpolated bi-directional frames (B-frames) are used to classify the camera motions by analyzing projective transformation parameters between two video frames.

For non-parametric models, image features can be used to detect camera motions, such as edge features [42]. In work [43], sequential video frames are divided into sub-images, equally. The dominant MVs in each sub-image lead to the camera motion patterns. In works [44, 45] present template matching in optical flow to classify the camera motions.

1D MV histograms are non-parametric models that are useful tools to classify the camera motions from the videos [1, 2, 46–51]. Shot characterization using difference in 1D MV histogram of consequent video frames are presented in [49, 50]. However, they use the 1D MV histogram in order to find only SC camera motions. In the MPEG video domain, three 1D MV histograms, including of image intensities, horizontal MVs, and vertical MVs, are extracted from intra-frame (I-frame) in order to find SC camera motions, while 1D MV histograms of MV orientations in predicted frame (P-frame) and bi-directional predicted frame (B-frame) are used to find PL, PR, TU, TD, ZI, and ZO camera motions [46]. Two 1D MV histograms of Cartesian coordinates (i.e. horizontal MVs histogram and vertical MVs histogram) are used to detect panning (i.e. PL and PR), tilting (i.e. TU and TD), and zooming (i.e. ZI and ZO) camera motions [47]. From MV fields, the MVs in Cartesian coordinates are converted into polar coordinates because a 1D MV histogram of polar coordinates can classify more directional movement of camera [1, 2, 48, 51]. However, it is difficult for the 1D MV histogram to classify complex camera motions, such as T and a combination of two camera motions (e.g. ZI+PL camera motions) since there is no link between the two 1D MV histograms.

An existing work [70] presents a 2D histogram based approach that utilizes both MV magnitudes and MV orientations synchronously, to detect a sequence of fighting or violent scenes. However, this approach mainly classifies S camera

motion. Therefore, it means that this approach cannot work well when the camera is moving.

3.2 Methodology

Figure 3.1 shows the framework of the proposed method. Before the classification start, the MVs are extracted from the videos using adaptive rood pattern search (ARPS) [52]. The ARPS is a powerful and fast block-based motion estimation approach as mentioned in the previous chapter. Recently, there is an implementation of the ARPS in [71]. It shows that the ARPS can process a video with resolution 1920×1080 pixels up to 30 frames per second. It means that the ARPS has fast computational times for estimating MVs in a large scale of images. For experiments, the proposed method is conducted by using MATLAB [72]. Note that there is a MATLAB source code of the ARPS technique on the website [73]. The proposed method comprises of 4 steps A, B, C, and D as shown in Figure 3.1. In Step A., a 2D MV histogram is generated directly in the polar coordinates system. The 2D MV histogram is utilized to classify the most dominant single camera motion in Step B. Unclassified camera motions in Step B are then further analyzed to obtain multiple camera motions in Step C. Finally, Step D. focuses on the analysis of ZI and ZO camera motions using MV field that is estimated by the ARPS.

3.2.1 2D Motion Vector Histogram Generation

The 2D MV histogram is generated directly in the polar coordinates. The MV magnitudes are stored as the radius of the polar coordinates while the MV orientations are stored as the orientation of the polar coordinates. In the ARPS, the default block sizes is set to 16×16 pixels. If the input videos have the resolutions at least HD720 (1280×720 pixels), the block size is changed to 40×40 pixels. However, the default block size cannot be used to some resolutions (e.g. 640×360 pixels in widescreen VGA). To solve this problem, the block size is slightly adjusted automatically until the block size is suitable for the video resolutions. Since the ARPS searches MVs from the center of the search area to the border of it, the search area size can be set freely. Therefore, there is no effect in computational time if we set the search area size for an extremely large. Finally, the size of the

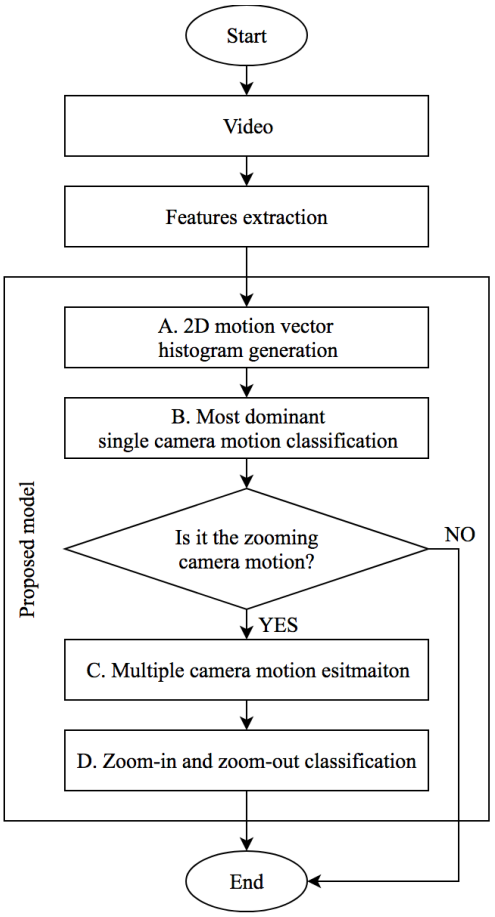


Figure 3.1: Framework of the proposed method

2D MV histogram is set by a square shape which contains the maximum size of horizontal and vertical MVs, $max(|u|, |v|)$. Note that each cell in the 2D histogram contains MV information (m, θ) in Eqs. (3.1)-(3.2).

$$m = \sqrt{u^2 + v^2} \tag{3.1}$$

$$\theta = \arctan \frac{v}{u} \tag{3.2}$$

where m denotes the MV magnitude, θ is the MV orientation, u and v are displacements of horizontal and vertical MVs respectively.

Figure 3.2 shows a cropped 2D MV histogram of size 7×7 array cells containing MV magnitudes m and MV orientations θ information. The center of the 2D MV histogram represents the no motion MV, where m is zero and θ is any value. The extracted MVs are accumulated at the array cell where it has the closest values m and θ. Finally, the 2D MV histogram is normalized in a range between 0 and 1. In the normalized 2D MV histogram, we made iterative experiments to filter the

2D MV histogram bins or the reliable MVs from all MVs. Finally, we accept the histogram peaks higher than or equal to the threshold value 0.14. The thresholded 2D MV histogram is visualized in black and white colors. The white cells indicate histogram bins that are higher than or equal to the threshold level, while the black cells correspond to the histogram bins that are lower than the threshold level.

(4.24, 135.00)	(3.61, 123.69)	(3.16, 108.44)	(3.00, 90.00)	(3.16, 71.57)	(3.61, 56.31)	(4.24, 45.00)
(3.61, 146.31)	(2.83, 135.00)	(2.24, 116.57)	(2.00, 90.00)	(2.24, 63.43)	(2.83, 45.00)	(3.61, 33.69)
(3.16, 161.57)	(2.24, 153.44)	(1.41, 135.00)	(1.00, 90.00)	(1.41, 45.00)	(2.24, 26.57)	(3.16, 18.43)
(3.00, 180.00)	(2.00, 180.00)	(1.00, 180.00)	(0.00,*)	(1.00, 0.00)	(2.00, 0.00)	(3.00, 0.00)
(3.16, 198.44)	(2.24, 206.57)	(1.41, 225.00)	(1.00, 270.00)	(1.41, 315.00)	(2.24, 333.43)	(3.16, 341.57)
(3.61, 213.69)	(2.83, 225.00)	(2.24, 243.44)	(2.00, 270.00)	(2.24, 296.57)	(2.83, 315.00)	(3.61, 326.31)
(4.24, 225.00)	(3.61, 236.31)	(3.16, 251.57)	(3.00, 270.00)	(3.16, 288.43)	(3.61, 303.69)	(4.24, 315.00)

Figure 3.2: 2D motion vector histogram in polar coordinates (m, θ)

3.2.2 Most Dominant Single Camera Motion Classification

The proposed method divides the degree orientation into eight ranges equally in order to define the eight directional camera movements (Figure 3.3).

From the divided degree orientation, if the the dominant MV magnitude is zero ($m = 0$), the scene is considered as S. If $m > 0$, we decide the dominant direction of MVs. Four directional camera movements, PL, PR, TU, and TD, are defined as follows:

- PL: The dominant MVs are in degree orientation ranges from 0.0° to 22.5° and from 337.5° to 360.0°.

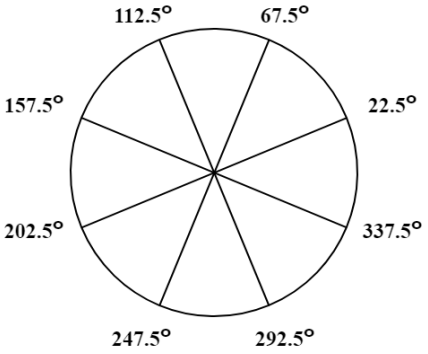


Figure 3.3: Angular ranges for estimating the directional camera movements

- PR: The dominant MVs are in degree orientation range from 157.5° to 202.5°.
- TU: The dominant MVs are in degree orientation range from 247.5° to 292.5°.
- TD: The dominant MVs are in degree orientation range from 67.5° to 112.5°.

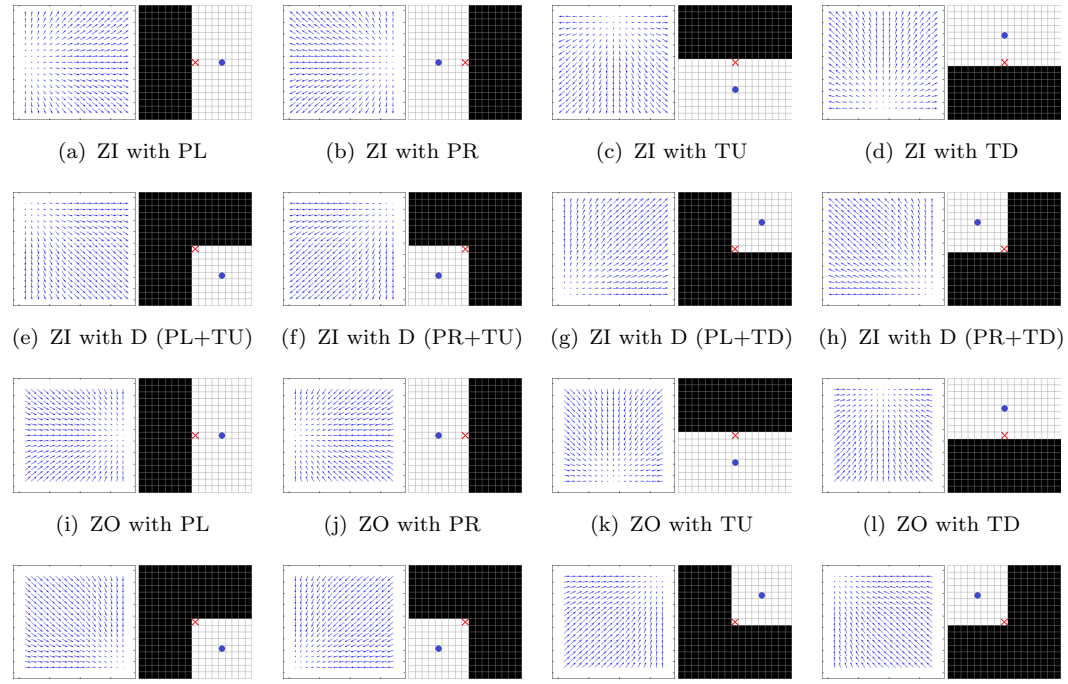
The remaining four ranges can be a combination of panning and tilting as described in more details in the next section.

3.2.3 Multiple Camera Motion Estimation

Both panning and tilting camera motions can occur simultaneously in practical videos. We consider the combination of panning and tilting camera motions as D. There are 4 more directions after the horizontal and vertical directions are discussed earlier (Figure 3.3). They are described below:

- D (PL+TU): The dominant MVs are in degree orientation range from 292.5° to 337.5°.
- D (PR+TU): The dominant MVs are in degree orientation range from 202.5° to 247.5°.
- D (PL+TD): The dominant MVs are in degree orientation range from 22.5° to 67.5°.
- D (PR+TD): The dominant MVs are in degree orientation range from 112.5° to 157.5°.

A challenge for detecting multiple camera motion patterns had been mentioned in the previous work [2]. By using the 2D MV histogram, these patterns can be estimated. Figure 3.4 illustrates multiple camera motions between zooming and panning camera motions, and between zooming and tilting camera motions. For both ZI and ZO, the centroids of converging and diverging MV fields are at the center of the MV field. The centroids of converging and diverging MV fields are shifted away from the center when there is an integrating camera motion (i.e. panning and tilting camera motions). For ZI, the centroid of diverging MV field is shifted-off from the center to the same direction of the integrating camera motion, while the centroid of converging MV field are shifted-off from the center to the opposite direction of the integrating camera motion for ZO. To estimate the



(a) ZI with PL (b) ZI with PR (c) ZI with TU (d) ZI with TD

(e) ZI with D (PL+TU) (f) ZI with D (PR+TU) (g) ZI with D (PL+TD) (h) ZI with D (PR+TD)

(i) ZO with PL (j) ZO with PR (k) ZO with TU (l) ZO with TD

(m) ZO with D (PL+TU) (n) ZO with D (PR+TU) (o) ZO with D (PL+TD) (p) ZO with D (PR+TD)

Figure 3.4: Simulation of multiple camera motions

multiple camera motions of zooming, the largest white region, which represents the zooming camera motion, is selected. Then its center of gravity (COG) is localized by Eqs. (3.3) to (3.6) which are described in Gonzalezs textbook [74].

$$M_{00} = \sum_{x} \sum_{y} C(x, y) \tag{3.3}$$

$$M_{10} = \sum_{x} \sum_{y} x \cdot C(x, y) \tag{3.4}$$

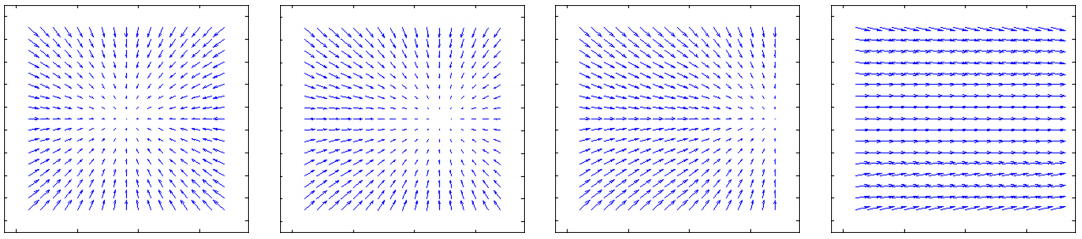
$$M_{01} = \sum_{x} \sum_{y} y \cdot C(x, y) \tag{3.5}$$

$$x_c = \frac{M_{10}}{M_{00}}, y_c = \frac{M_{01}}{M_{00}} \tag{3.6}$$

where $C(x, y)$ is a value of the cells on coordinates (x, y) and (x_c, y_c) is the coordinates of the center of gravity. For the cell values, the white cells value is $C(x, y) = 1$ while the black cells value is $C(x, y) = 0$.

- ZI/ZO: The COG is at the center of the 2D MV histogram.
- ZI/ZO with PL: The COG is at the right of the 2D MV histogram.
- ZI/ZO with PR: The COG is at the left of the 2D MV histogram.
- ZI/ZO with TU: The COG is at the lower half of the 2D MV histogram.
- ZI/ZO with TD: The COG is at the upper half of the 2D MV histogram.
- ZI/ZO with D (PL+TU): The COG is at the lower right of the 2D MV histogram.
- ZI/ZO with D (PR+TU): The COG is at the lower left of the 2D MV histogram.
- ZI/ZO with D (PL+TD): The COG is at the upper right of the 2D MV histogram.
- ZI/ZO with D (PR+TD): The COG is at the upper left of the 2D MV histogram.

These multiple camera motions can be affected by speeds. Figure 3.5 illustrates the combination of ZO with PL camera motions. It is considered as regular ZO camera motion, if the PL camera motion is not existing (Figure 3.5(a)). The centroid of the converging MV field is shifted away from the center of MV field depending on the speed of PL. If it shifts for a short distance, it is ZO with slow PL camera motion (Figure 3.5(b)). If it shifts for a long distance, it is ZO with fast PL camera motion (Figure 3.5(c)). If the speed of PL camera motion is extremely fast, it is considered as PL rather than ZO camera motion (Figure 3.5(d)).



(a) Zoom-out without panning left (b) Zoom-out with slow panning left (c) Zoom-out with fast panning left (d) Zoom-out with extremely fast panning left

Figure 3.5: Simulations of zoom-out with the difference in panning left speed

3.2.4 Camera motion classification

To classify the camera motions, we separate the property of the 2D MV histogram into two cases:

- Case I: There is one white cell in the 2D MV histogram.
- Case II: There are at least two white cells in the 2D MV histogram.

Case I is the simplest case to classify S, PL, PR, TU, TD, and D camera motions because there is only one white cell in the 2D MV histogram. Referring to the previous sections, we label the 2D MV histogram in Figure 3.2 with the camera motions as shown in Figure 3.6. The array cell at the center of the 2D MV histogram represents S, while the other array cells at the off-center of the 2D MV histogram represent PL, PR, TU, TD, and D. The camera motions in Case I can be clasified by following the position of the array cell in white color.

Case II is more complex than the Case I because the 2D MV histogram contains several patterns of white cells. To classify the camera motions in Case II, we group all connected white cells as a white region. Then, we look at the scenarios below.

- If there is one white region of array cells with at least shape size 3×3 array cells, we classify as ZI and ZO camera motions. Note that the shape size 3×3 array cells represent the slowest zooming speed which has one pixel of motion in all orientations.
- If there is one white region with shape size smaller than 3×3 array cells, we consider as S, PL, PR, TU, TD, and D camera motions depending its dominant MVs.

D (PR+TD)	D (PR+TD)	TD	TD	TD	D (PL+TD)	D (PL+TD)
D (PR+TD)	D (PR+TD)	D (PR+TD)	TD	D (PL+TD)	D (PL+TD)	D (PL+TD)
PR	D (PR+TD)	D (PR+TD)	TD	D (PL+TD)	D (PL+TD)	PL
PR	PR	PR	S	PL	PL	PL
PR	D (PR+TU)	D (PR+TU)	TU	D (PL+TU)	D (PL+TU)	PL
D (PR+TU)	D (PR+TU)	D (PR+TU)	TU	D (PL+TU)	D (PL+TU)	D (PL+TU)
D (PR+TU)	D (PR+TU)	TU	TU	TU	D (PL+TU)	D (PL+TU)

Figure 3.6: 2D motion vector histogram with camera motion labels

- If there are two white regions and the larger white region has the size smaller than 3×3 array cells, we classify as T camera motion. We can explain that one white cell represents background motion while another white cell represents foreground motion.

- If there are two white regions and the larger white region has size at least 3×3 array cells, we classify as ZI and ZO camera motion.

- If there are more than two white regions, we consider as ZI, ZO, and SC camera motions. First, we calculate the average MV magnitude to separate the white cells into two groups: 1) the white cells with MV magnitude less than or equal to the average MV magnitude and 2) the white cells with MV magnitudes greater than the average MV magnitude. Finally, we compare the number of white cells from these two groups. If the first group has white cells more than or equal to the second group, it is considered as a ZI and ZO camera motions. Otherwise, it is SC camera motion.

The classification at above is summarized by the decision tree, as shown in Figure 3.7.

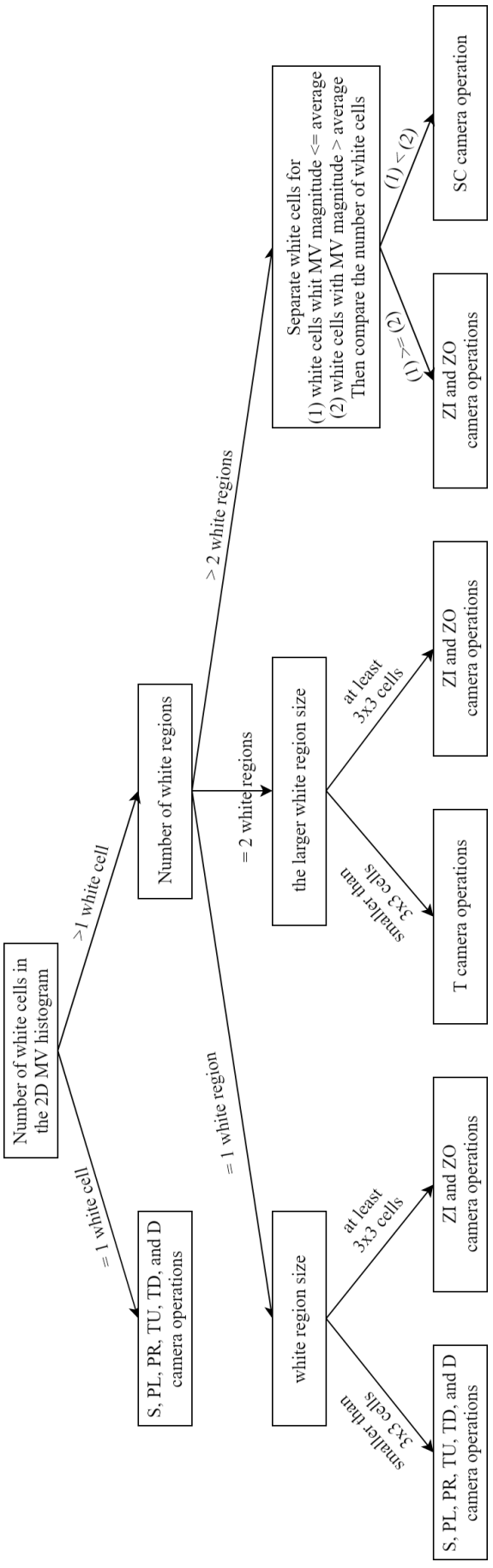
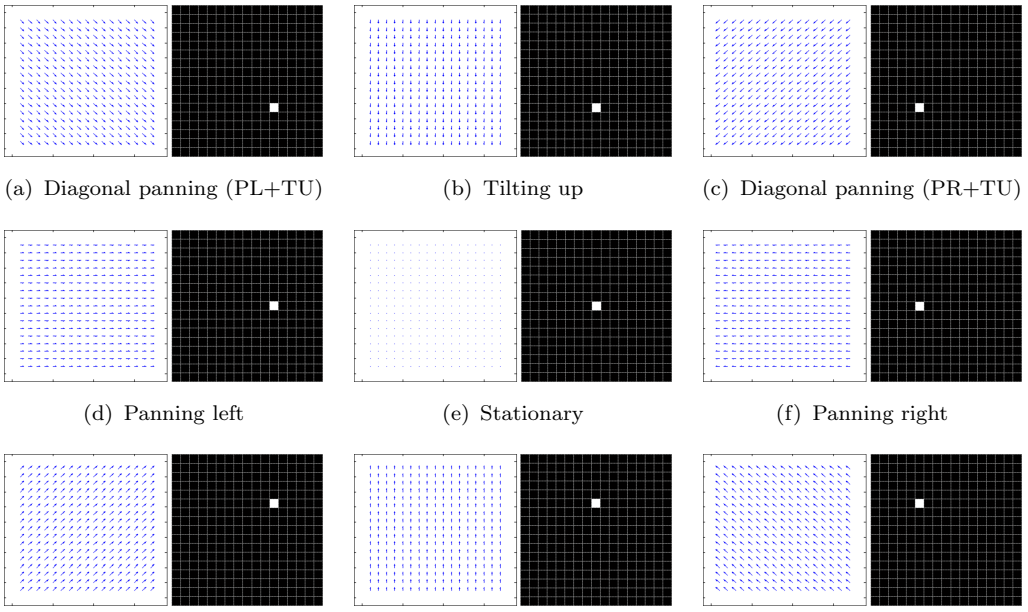


Figure 3.7: Decision tree of most dominant single camera motion classification

From Figures 3.8 to 3.11 give an explanation of all camera motions in the 2D MV histogram via MV field simulations. Figures 3.8 and 3.9 show the simulation of panning and tilting camera motions at different speed. By using the 2D MV histogram, the operational speed can be known directly. If there is no operation in camera, the white cell is always at the center of the 2D MV histogram as S camera motion (Figures 3.8(e) and 3.9(e)). For panning and tilting camera motions, their white cells are located close to the center of the 2D MV histogram for slow operation (Figures 3.8(a), 3.8(b), 3.8(c), 3.8(d), 3.8(f), 3.8(g), 3.8(h), and 3.8(i)), while their white cell are located away from the center of the 2D MV histogram for fast operation (Figures 3.9(a), 3.9(b), 3.9(c), 3.9(d), 3.9(f), 3.9(g), 3.9(h), and 3.9(i)). Figure 3.10 shows the simulation of zooming camera motions



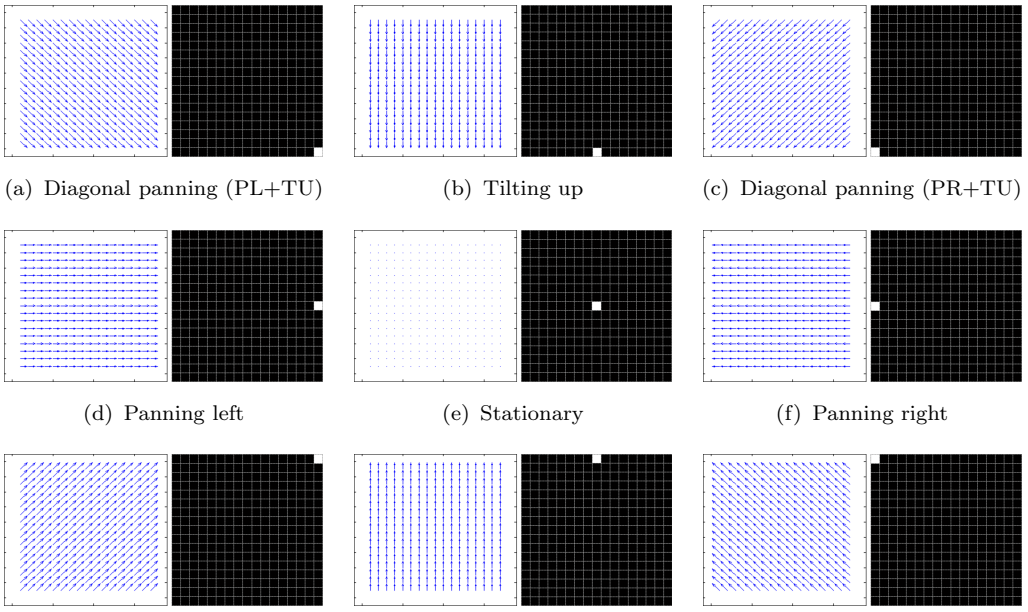
(a) Diagonal panning (PL+TU) (b) Tilting up (c) Diagonal panning (PR+TU)

(d) Panning left (e) Stationary (f) Panning right

(g) Diagonal panning (PL+TD) (h) Tilting down (i) Diagonal panning (PR+TD)

Figure 3.8: Simulation of slow camera motions

at different speed. From no operation in camera (i.e. S camera motion), the white cell is stretched linearly by the operational speed of the zooming camera motions. The MV magnitudes from the center to the border of the MV field slightly increase for slow operation (Figures 3.10(a) and 3.10(d)), while they constantly increase for fast operation (Figures 3.10(b) and 3.10(e)). Therefore, the 2D MV histogram contains a white square with different sizes depending on the zooming speed. However, when the camera operates zooming extremely fast, it makes the white cells spread away from each other (Figures 3.10(c) and 3.10(f)). Figure 3.11 shows the simulation of two special camera motion patterns. When the camera tries to

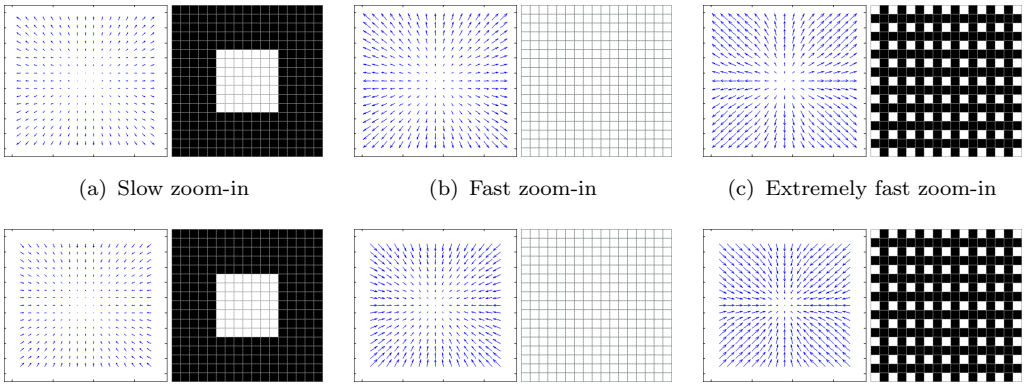


(a) Diagonal panning (PL+TU) (b) Tilting up (c) Diagonal panning (PR+TU)

(d) Panning left (e) Stationary (f) Panning right

(g) Diagonal panning (PL+TD) (h) Tilting down (i) Diagonal panning (PR+TD)

Figure 3.9: Simulation of fast camera motions

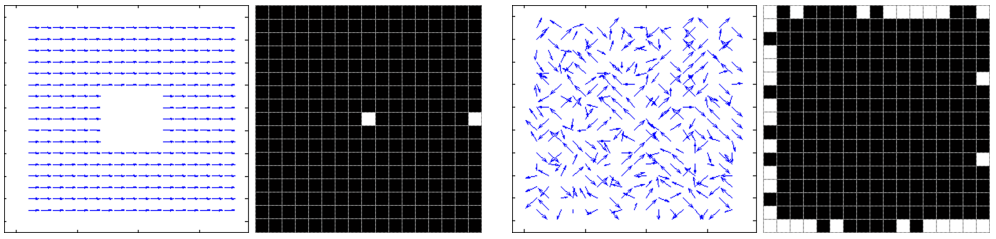


(a) Slow zoom-in (b) Fast zoom-in (c) Extremely fast zoom-in

(d) Slow zoom-out (e) Fast zoom-out (f) Extremely fast zoom-out

Figure 3.10: Simulation of zooming camera motions

follow an object, the MVs at the center of the MVs field usually have less motion than the other MVs (Figure 3.11(a)). Thus, there are at least two white regions on the 2D MV histogram which the one region represents the object and another one represent the background. SC is the special camera motion which is about changing a viewpoint from a camera to the other camera. In MV estimation, it estimates random MVs since it selects the most approximate matching from two different video frames (Figure 3.11(b)).



(a) Object tracking (b) Scene change

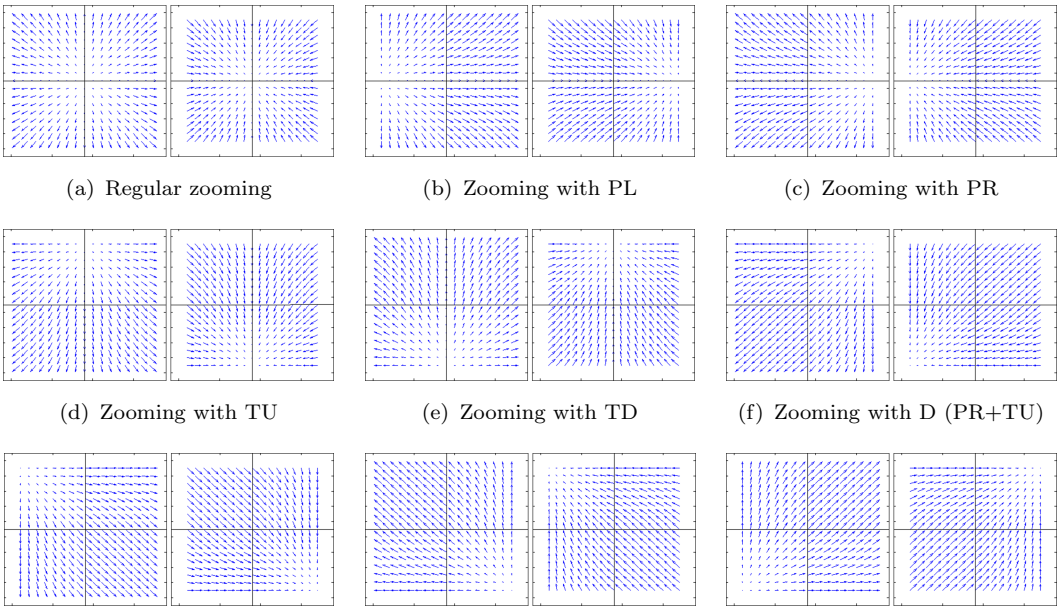
Figure 3.11: Simulation of two special camera motions

3.2.5 Zoom-in and Zoom-out Classification

It is a limitation for the 2D MV histogram to recognize both ZI and ZO camera motions because these two camera motions produce a similar pattern in the 2D MV histogram. In order to solve their limitations, we use the MV field which is estimated by the ARPS in the first step. The MV field is divided into four quadrants and compared with 9 templates of zooming camera motions, as shown in Figure 3.12.

To distinguish between ZI and ZO camera motions, we select one pair of the templates by referring to the center of gravity (x_c, y_c). Because of Step B. and Step C., the proposed method already recognized converging and diverging MV fields as zooming operations. Then, we simply distinguish between ZI and ZO camera motions by investigating the number of signs of MV components, horizontal MV u and vertical MV v. Table 3.1 shows the criteria for distinguishing between ZI and ZO camera motions for each MV field quadrant position. If the MV field quadrant position contains MVs of more than a half of the total MVs, the camera motion (i.e. ZI or ZO) is voted to the MV field quadrant. Note that ZI and ZO with D operations have the same sign of MV components. We involve zero-MV components which represent the centroids of converging and diverging MV fields. Finally, both voting scores are compared to each other following the instructions below.

- If the ZI score is greater than the ZO score, it is a ZI camera motion.
- If the ZO score is greater than the ZI score, it is a ZO camera motion.
- If both of them have the same score, it is considered SC in order to treat misclassifications from the previous steps.



(a) Regular zooming (b) Zooming with PL (c) Zooming with PR

(d) Zooming with TU (e) Zooming with TD (f) Zooming with D (PR+TU)

(g) Zooming with D (PL+TU) (h) Zooming with D (PR+TD) (i) Zooming with D (PL+TD)

Figure 3.12: Divided zoom-in (left) and zoom-out (right) templates

Table 3.1: Criteria for distinguishing between zoom-in and zoom-out camera motions

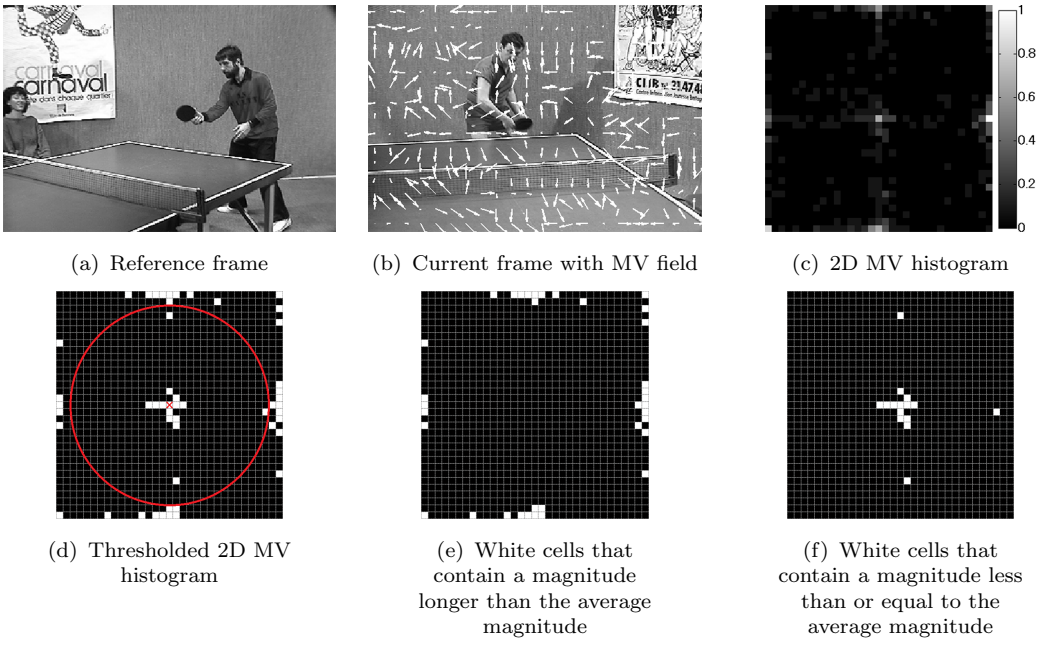
Camera motions		MV field quadrant positions			
Main	Sub	Upper left	Upper right	Lower left	Lower right
	-	$(-u, v)$	(u, v)	$(-u, -v)$	$(u, -v)$
	PL	(u, v)	(u, v)	$(u, -v)$	$(u, -v)$
	PR	$(-u, v)$	$(-u, v)$	$(-u, -v)$	$(-u, -v)$
	TU	$(-u, -v)$	$(u, -v)$	$(-u, -v)$	$(u, -v)$
ZI	TD	$(-u, v)$	(u, v)	$(-u, v)$	(u, v)
	D (PR+TU)	$(-u, -v)$	$(-u, -v)$ and $(0, 0)$	$(-u, -v)$	$(-u, -v)$
	D (PL+TU)	$(u, -v)$ and $(0, 0)$	$(u, -v)$	$(u, -v)$	$(u, -v)$
	D (PR+TD)	$(-u, v)$	$(-u, v)$	$(-u, v)$	$(-u, v)$ and $(0, 0)$
	D (PL+TD)	(u, v)	(u, v)	(u, v) and $(0, 0)$	(u, v)
	-	$(u, -v)$	$(-u, -v)$	(u, v)	$(-u, v)$
	PL	$(u, -v)$	$(u, -v)$	(u, v)	(u, v)
	PR	$(-u, -v)$	$(-u, -v)$	$(-u, v)$	$(-u, v)$
	TU	$(u, -v)$	$(-u, -v)$	$(u, -v)$	$(-u, -v)$
ZO	TD	(u, v)	$(-u, v)$	(u, v)	$(-u, v)$
	D (PR+TU)	$(-u, -v)$	$(-u, -v)$	$(-u, -v)$ and $(0, 0)$	$(-u, -v)$
	D (PL+TU)	$(u, -v)$	$(u, -v)$	$(u, -v)$	$(u, -v)$ and $(0, 0)$
	D (PR+TD)	$(-u, v)$ and $(0, 0)$	$(-u, v)$	$(-u, v)$	$(-u, v)$
	D (PL+TD)	(u, v)	(u, v) and $(0, 0)$	(u, v)	(u, v)

3.3 Results and Discussion

In this section, four examples of camera motion estimation, which are done by the proposed method, are discussed.

3.3.1 A Real Video in Scene Change Camera Motion

Figure 3.13 shows an example of SC camera motion where Figure 3.13(a) and 3.13(b) exhibit a sudden changing in a video sequence. Figure 3.13(b) indicates an irregular MV field of the image, compared with the previous frame in Figure 3.13(a), because the two time-sequential frames show different views. Figure 3.13(c) demonstrates a corresponding 2D MV histogram before thresholding. After thresholding, we have a 2D binary histogram, as shown in Figure 3.13(d). This is Case II with more than two white regions in Figure 3.7. The $\times$ symbol and circle in Figure 3.13(d) correspond to the centroid of the 2D MV histogram and the average MV magnitude, respectively. The white cells in Figure 3.13(d) are then separated outside of the red circle in Figure 3.13(e) and inside of the circle in Figure 3.13(f). Since the number of cells within the circle (i.e. 15 cells) is less than that outside of the circle (i.e. 36 cells), the proposed method considers the camera motion as an SC camera motion.



(a) Reference frame (b) Current frame with MV field (c) 2D MV histogram

(d) Thresholded 2D MV histogram

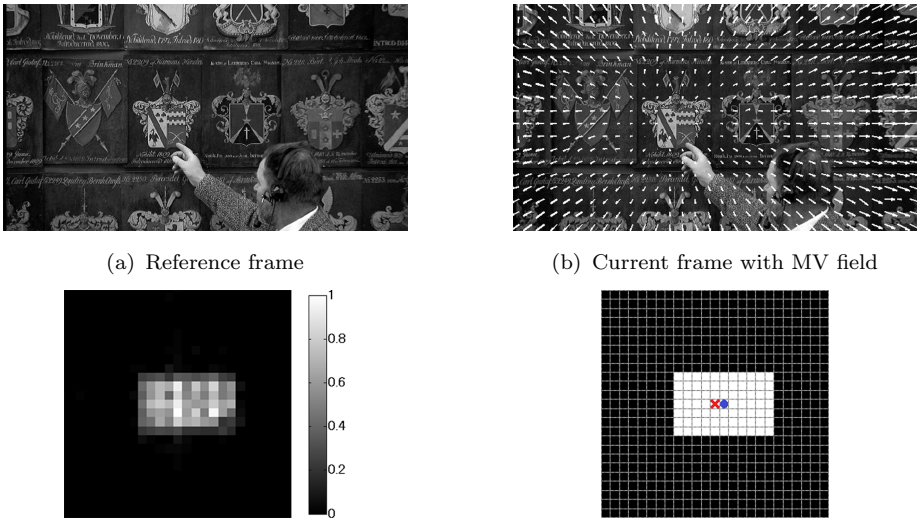
(e) White cells that contain a magnitude longer than the average magnitude

(f) White cells that contain a magnitude less than or equal to the average magnitude

Figure 3.13: An example of scene change camera motion estimation by the proposed method

3.3.2 A Plain Background in Zooming Video

Figure 3.14 shows an example of a ZI camera motion. Since the camera slowly extends the focal length to perform ZI from Figure 3.14(a) to 3.14(b), the MV field shows diverging MVs on the image in Figure 3.14(b). The 2D MV histogram of the MV field shows a square shape of histogram bins (Figure 3.14(c)). After thresholding, Figure 3.14(d) shows that we have a rectangular white region that has the size larger than 3×3 array cells. This is Case II, especially the left path in Figure 3.7. Since the MV field shows the diverging MV, the proposed method considers as ZI camera motion. The proposed method then notices that the centroid of the detected rectangle (• symbol) does not coincide with the centroid of the 2D MV histogram (× symbol). The displacement between two symbols is $(x_c, y_c) = (1, 0)$, which means that the integrating camera motion is PL camera motion. Thus, the proposed method detects the combination of ZI and PL camera motions.



(a) Reference frame

(b) Current frame with MV field

(c) 2D MV histogram

(d) Thresholded 2D MV histogram

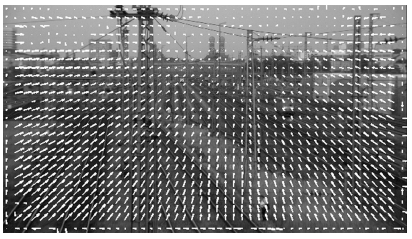
Figure 3.14: An example of zoom-in camera motion estimation by the proposed method

3.3.3 A Complex Background in Zooming Video

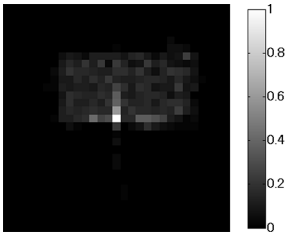
Figure 3.15 demonstrates an example of ZO camera motion. The video frame in Figure 3.15(a) is transformed to Figure 3.15(b) by ZO with TD camera motions. The converging MV field in Figure 3.15(b) presents its corresponding 2D MV

histogram in Figure 3.15(c). It does not have an arrangement of histogram peaks as well as Figure 3.14(c) because of complexness in picture. After thresholding, we obtain a thresholded 2D MV histogram with an unwell white rectangle as shown in Figure 3.15(d). The thresholded 2D MV histogram shows that it is the right path of Case II in Figure 3.7. Subsequently, the white cells outside and inside of the circle (i.e. the average MV magnitude) are separately plotted in Figure 3.15(e) and 3.15(f), respectively. Since the number of white cells in Figure 3.15(e) is less than the number of white cells in Figure 3.15(f), the proposed method decides that is zooming camera motion. From the MV field analysis, it considers that is ZO camera motion. In addition, it is found that the $\times$ symbol is slightly displaced, higher than the $\bullet$ symbol of the centroid. the proposed method considers that the camera motion is a combination of ZO and TD camera motions.

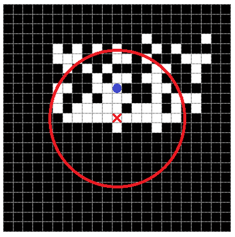
(a) Reference frame



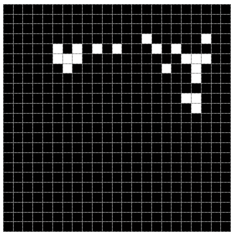
(b) Current frame with MV field



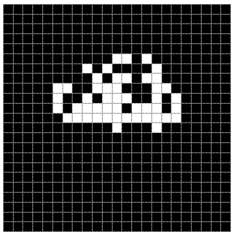
(c) 2D MV histogram



(d) Thresholded 2D MV histogram



(e) White cells that contain a magnitude longer than the average magnitude

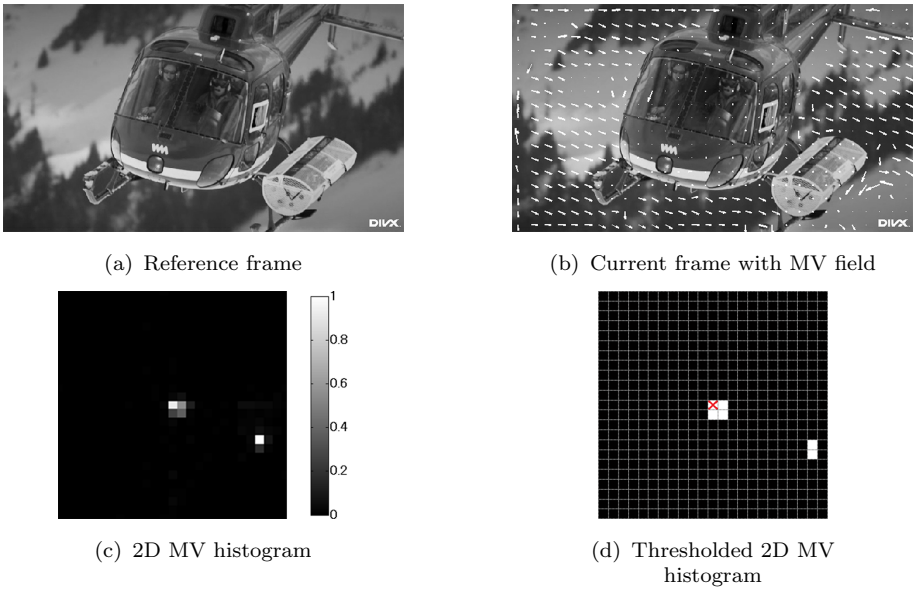


(f) white cells that contain a magnitude less than or equal to the average magnitude

Figure 3.15: An example of zoom-out camera motion estimation by the proposed method

3.3.4 A Sequence of Object Tracking Camera Motion

Figure 3.16 shows an example of T camera motion. Figure 3.16(a) and 3.16(b) indicate that a helicopter is tracked in the upper left direction. Two bright spots in the 2D MV histogram in Figure 3.16(c) correspond to MVs in the background and foreground. After thresholding, there are two white regions, which correspond to the middle path of Case II in Figure 3.7. One region is located at the center of the 2D MV histogram, while the other region is located off-center of the 2D MV histogram. The white region at the center represents the foreground that is the helicopter being tracked. On the other hand, the off-center white region represents uniform MVs in the background. Since those two white regions are smaller than 3×3 array cells, the proposed method considers that the camera motion is T.



(a) Reference frame

(b) Current frame with MV field

(c) 2D MV histogram

(d) Thresholded 2D MV histogram

Figure 3.16: An example of object tracking camera motion estimation by the proposed method

3.4 Evaluations and Comparisons

For evaluation, the proposed method is run on the MATLAB with an Intel Core i7-4750HQ Processor at 2 GHz. We use various video sequences from websites [75–78]. Table 3.2 shows a list of video sequences and their sources that are used for evaluating the proposed method. There are 25 video sequences from several categories (e.g. sport, video games, movie, and general videos). To make

ground truths, we manually annotate the camera motions at the middle of the two consecutive video frames.

Table 3.2: Video sequences and their source for evaluations

Sources	Video sequences
[75]	big_buck_bunny, coastguard, DOTA2, elephants_dream, EuroTruckSimulator2, park_joy, parkrun, shields, soccer, station2, STARCRAFT, stockholm, tennis, tractor, vidyo1, vidyo3, vidyo4, and washdc
[76]	Helicoppter_Flight
[77]	Bosphorus, Jockey, ReadySetGo, ShakeNDry, and YachtRide
[78]	stefan and desert

3.4.1 Evaluation on Camera Motion Extraction

We create 2,000 frames of testing videos for each camera motion using the videos in Table 3.2. They are stationary, panning, tilting, diagonal panning, tracking, and scene change video sequences. Therefore, there are totally 14,000 frames.

From experiments, we found that the proposed method never classify actual panning, tilting, diagonal panning, and zooming camera motions as non-actual panning, tilting, diagonal panning, and zooming camera motions. For example, PL as PR, TU as TD, ZI as ZO, and D (PL+TU) as D (PR+TU). Thus, we build the confusion matrix as shown in Table 3.3 which The rows of the confusion matrix correspond to the ground truth while the columns of the confusion matrix correspond to the estimation results. From the confusion matrix, we compute accuracy (ACC), sensitivity (SEN), precision (PRE), specificity (SPE), and F1 score (F1), which are defined as:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.7}$$

$$SEN = \frac{TP}{TP + FN} \tag{3.8}$$

$$PRE = \frac{TP}{TP + FP} \tag{3.9}$$

$$SPE = \frac{TN}{TN + FP} \tag{3.10}$$

$$F1 = \frac{2TP}{2TP + FP + FN} \tag{3.11}$$

where TP, that is, true positive, means the number of correct classifications of the true camera motions, TN, true negative, indicates the number of correct classifications of the non-true camera motions, FP, false positive, denotes the number of incorrect classifications of the true camera motions, and finally FN, false negative, means the number of incorrect classifications of the non-true camera motions.

Table 3.3: Classification results of all camera motions by the proposed methods

Ground truth	Proposed method						
	S	PL&PR	TU&TD	D	ZI&ZO	T	SC
S	1995	3	2	0	0	0	0
PL&PR	59	1848	0	79	10	3	1
TU&TD	60	0	1881	15	29	15	0
D	4	74	26	1732	144	2	18
ZI&ZO	24	170	53	198	1414	13	128
T	0	215	2	69	25	1687	2
SC	0	2	2	6	0	8	1982
ACC	98.9	95.6	98.5	95.5	94.3	97.5	98.8
SEN	99.8	92.4	94.1	86.6	70.7	84.4	99.1
PRE	93.1	79.9	95.7	82.5	87.2	97.6	93.0
SPE	98.8	96.1	99.3	96.9	98.3	99.7	98.8
F1	96.3	85.7	94.9	84.5	78.1	90.5	96.0

From Table 3.3, the proposed method has ACC and SPE performance scores more than 94.3% for all camera motions. It means "In overall, all camera motions can be classified by the proposed method correctly". The proposed method has the PRE score in PL&PR lower than the PRE score in TU&TD because of T camera motions. Since most of tracking video sequences are performed by PL and PR camera motions, the misclassification of T camera motions are PL and PR camera motions. The proposed method also has low PRE scores for diagonal (82.5%) and zooming (87.2%) camera motions. Since both D and zooming (i.e. ZI and ZO) with D camera motions are similar to each other, they usually are misunderstood easily. In F1 score, the proposed method has the score 78.1% for zooming camera motions. Following the confusion matrix, the proposed method misunderstands the zooming camera motions as panning, tilting, diagonal panning, and scene change camera motions. All the misclassifications are caused by the involvement of zooming operations such as ZI+PL and ZI+D.

We evaluate the performance in zooming camera motion by generating the confusion matrix of all zooming camera motions (Table 3.4). They are regular zooming camera motions and the combination of zooming camera motions. We also merge all zooming with panning, tilting, and diagonal panning into Z+PL/PR, Z+TU/TD, and Z+D respectively because there is no misclassification between the two of them in the experiments. As mentioned in Figure 3.5, if the speed of integrating camera motions (i.e. panning, tilting, and diagonal) are extremely fast, the proposed method considers it as the integrating camera motions instead of zooming. Moreover, the zooming camera motions are wrongly classified as SC. There is a sequence from video "tractor" that is ZI with PL and T camera motions. Since the proposed method cannot recognize the zooming pattern in Step D, this sequence is considered as SC.

Table 3.4: Classification results of all zooming camera motions by the proposed method

Ground truth	Proposed method				
	Z	Z+PL/PR	Z+TU/TD	Z+D	Other
Z	438	10	12	3	37
Z+PL/PR	12	167	6	17	298
Z+TU/TD	11	4	394	38	53
Z+D	4	29	22	247	198
ACC	95.6	81.2	92.7	84.5	-
SEN	87.6	33.4	78.8	49.4	-
PRE	94.2	79.5	90.8	81.0	-
SPE	98.2	97.1	97.3	96.1	-
F1	90.8	47.0	84.4	61.4	-

3.4.2 Comparisons of Camera Motion Extraction in Several Methods

We compare the performances with parametric based approach [41, 79, 80] and non-parametric based approaches [1, 2, 47, 81]. Tables 3.5 and 3.6 show lists of all camera motions that can be detected by each method. There are two categories of the camera motions: 1) single camera motions and 2) multiple camera motions. For the single camera motions, S, PL, PR, TU, TD, ZI, ZO, T, and SC camera motions are in Table 3.5. For the multiple camera motions, D, all ZI and ZO camera motions with PL, PR, TU, TD, and D camera motions are in Table 3.6. Note that we group both ZI and ZO camera motions as the same category, which

is named "zooming (Z)" camera motion because all methods never consider ZI as ZO camera motion, and never consider ZO as ZI camera motion.

Table 3.5: Lists of single camera motions that can be detected by each method

Methods	Single camera motion								
	S	PL	PR	TU	TD	ZI	ZO	T	SC
Abdollahian [79]	✓					✓[2]	✓[2]		
Weng [41]		✓	✓	✓	✓	✓[2]	✓[2]		
Narayanan [80]	✓	✓	✓	✓	✓				
Duan [47]	✓	✓	✓	✓	✓	✓	✓		✓[1]
Hasan [1]	✓	✓	✓	✓	✓	✓	✓	✓	✓
Okade [2]	✓	✓	✓	✓	✓	✓[2]	✓[2]		
Derue [81]	✓	✓	✓	✓	✓	✓	✓		
Proposed method	✓	✓	✓	✓	✓	✓	✓	✓	✓

Table 3.6: Lists of multiple camera motions that can be detected by each method

Methods	Multiple camera motions					
	D	Z+PL	Z+PR	Z+TU	Z+TD	Z+D
Abdollahian [79]						
Weng [41]						
Narayanan [80]						
Duan [47]						
Hasan [1]						
Okade [2]						
Derue [81]						
Proposed method	✓	✓	✓	✓	✓	✓

From the features tables, most of all methods can classify S, PL, PR, TU, TD, ZI, and ZO camera motionss. The method [79] is an exceptional case because it uses horizontal and vertical motion parameters to detect such blur and shaky video frames instead of panning and tilting camera motions. Since the standard deviation in MV orientation is used by the methods [2, 41, 79], ZI and ZO camera motions can be detected but ZI and ZO cannot be distinguished. Only the proposed method and Hasan's method can detect T and SC camera motions. Although Duan's method can detect SC camera motions, they are considered as errors in motion estimation rather than the camera motions. Both Hasan's method [1] and Derue's method [81] have the similar approach the MV field is divided into 3×3 sub-region equally. Since the proposed method uses 2D MV histogram instead

[1] [47] considers as errors in MV field rather than the camera motions.
[2] [2, 41, 79] consider both ZI and ZO as the same class, zooming camera motion.

of 1D MV histogram, the combination of two camera motions (e.g. D, PL+PR, ZI+PL) can be detected.

For comparisons, we re-implement three recent existing methods [1, 2, 81] by following their instructions since they did not provide MATLAB source code files. They also use the MV magnitude and MV orientation histograms to classify the camera motions as same as the proposed method. Therefore, we use the same test videos in the previous section. However, we exclude the SC video sequences because there is no exact instruction to describe the SC camera motion in the existing methods.

Table 3.7 shows the comparative performance of F1 score in six camera motions. From the table, all methods succeed to estimate the S camera motions from the test videos. The proposed method has F1 scores in PL&PR and TU&TD higher than the existing methods because they cannot detect D camera motions. Since all existing methods cannot detect the combination of zooming with panning and zooming with tilting, all existing methods have low F1 score in ZI&ZO. However, the existing method [2] has the F1 score in ZI&ZO higher than the existing methods [1, 81] because some combination of zooming camera motions are detected as zooming camera motion correctly by using standard derivation. For T camera motions, the proposed method performs the classification better than the existing method [1].

Table 3.7: Comparison in F1 scores by using stationary, panning, tilting, zooming, diagonal, and tracking video sequences

Methods	F1 scores for each class					
	S	PL&PR	TU&TD	ZI&ZO	D	T
Hasan [1]	95.4	61.1	65.9	34.2	0.0	83.1
Okade [2]	97.4	56.9	62.7	59.9	0.0	0.0
Derue [81]	95.8	52.0	61.6	33.2	0.0	0.0
Proposed	96.3	85.8	94.9	78.1	84.6	90.7

3.4.3 Computational Evaluation

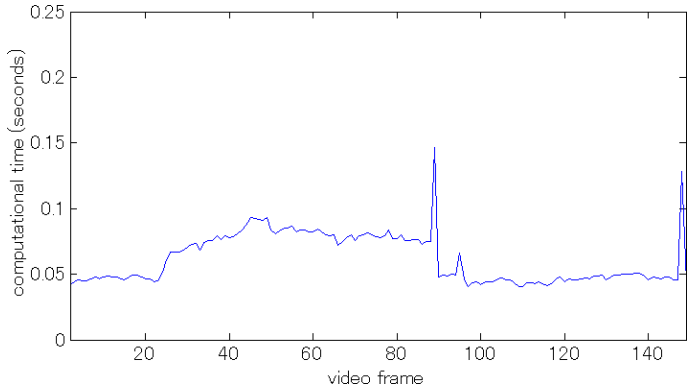
Figure 3.17 shows the computational time of the video sequence "Table tennis" (Figure 3.13(a)). This video has a resolution of 352×240 pixels. There are totally 330 MVs using ARPS block of size 16×16 pixels. The video contains a series of camera motions: S camera motion from frame numbers 1 to 25, ZO camera motion

from frame numbers 26 to 88, SC camera motion at frame number 89, S camera motion from frame numbers 90 to 147, and SC camera motion at frame number 148. By following the steps of the proposed method, the ARPS motion estimation uses times from 0.05 to 0.15 seconds to obtain one MV field (Figure 3.17(a)). From the MV field, the proposed method consumes 0.005 seconds in order to detect the S and SC camera motions, while it consumes 0.010 seconds in order to detect the ZO camera motion (Figure 3.17(b)). It takes longer 0.005 seconds compared with S, SC camera motions for distinguishing between ZI and ZO camera motions from four sub-regions of the MV field. In summary, the whole system uses between 0.05 and 0.10 seconds per frame to identify the camera motions (Figure 3.17(c)).

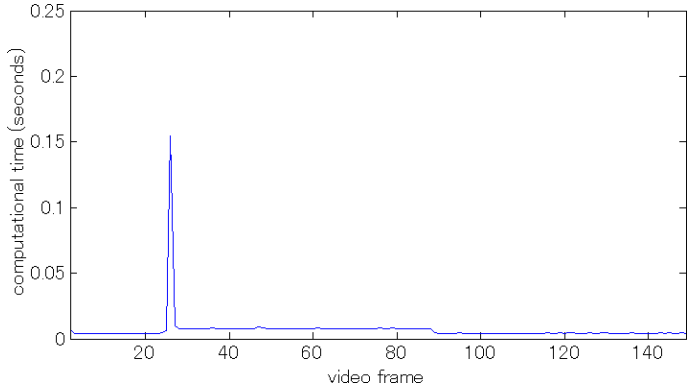
Figure 3.18 shows the computational time of the high definition video sequence, 1920×1080 pixels, namely "Station2" (Figure 3.15(a)). There are totally 1,296 MVs using ARPS block of size 40×40 pixels. The video sequences are ZO+TD camera motion from frame numbers 1 to 243, TD camera motion from frame numbers 244 to 267, and S camera motion from frame numbers 268 to 312. By increasing the video resolution, the ARPS motion estimation uses a longer time from 0.05-0.15 seconds to 0.4-0.8 seconds to obtain an MV field (Figure 3.18(a)). The proposed method identifies the camera motions with between 0.04 and 0.08 seconds per frame (Figure 3.18(b)). Totally, the whole system uses within 0.9 seconds per frame, in order to describe the camera motion (Figure 3.18(c)).

3.5 Conclusions

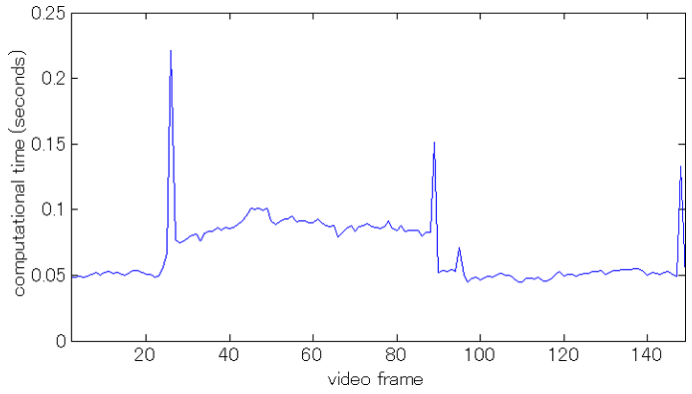
In this chapter, we propose an original technique for classifying several camera motions in several videos. First of all, we obtain series of MV fields by applying the existing block based motion estimation, which is called "ARPS", to an input video. Then, we generate the 2D MV histogram in the polar coordinates system that each histogram peak refers to MV magnitude m and MV orientation θ simultaneously. By analyzing the 2D MV histogram, we can detect a variety of camera motions that include S, PL, PR, TU, TD, ZI, ZO, T, and SC camera motions. Moreover, the proposed method can also detect a combination of these various camera motions, such as ZI+PL camera motions and PL+TU camera motions. We also utilize MV fields to separate between diverging MV fields (i.e. ZI camera motion) and converging MV fields (i.e. ZO camera motion). In this manner, the proposed method uses both MV field and 2D MV histogram to recognize all mentioned



(a) ARPS process time



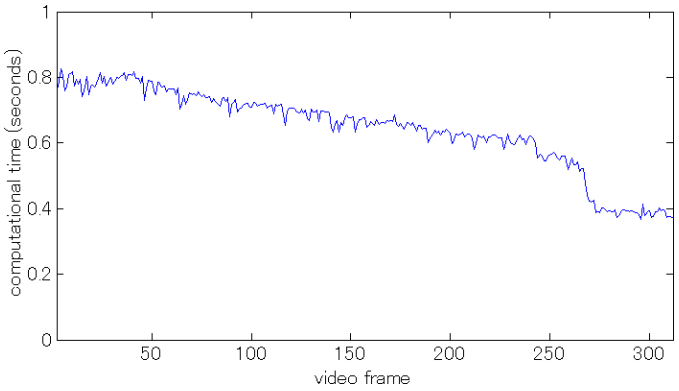
(b) Proposed method process time



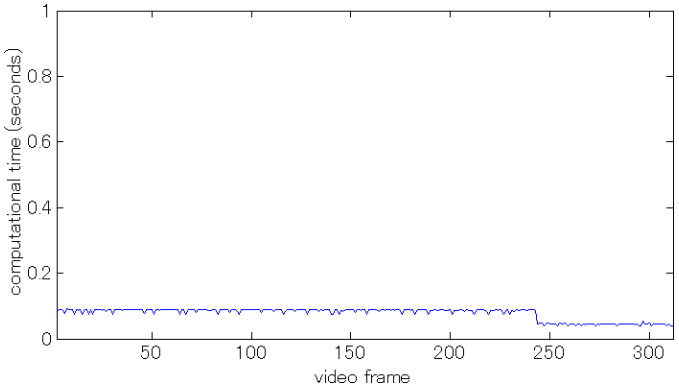
(c) Total process time

Figure 3.17: The computational time in video "Table tennis"

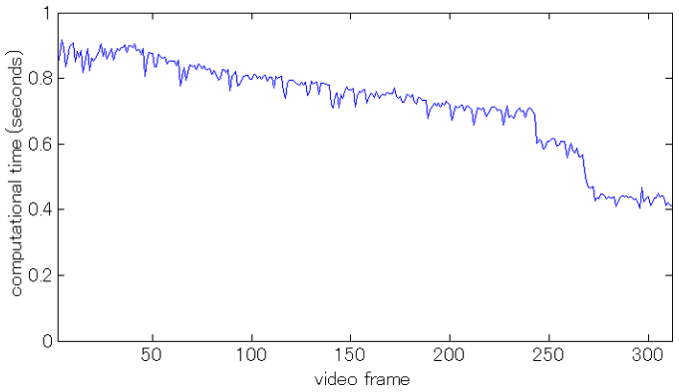
camera motions. Finally, our experiments show that the proposed method is computationally efficient. The proposed method requires only 1/10 of the time necessary for the MV estimation by ARPS. The proposed method can achieve a processing time of 5-10 millisecond per frame for a low-resolution video sequence (e.g. 352×240 pixels), and 40-80 millisecond for a high-resolution video sequence (e.g. 1920×1080 pixels). In the evaluation, we found that 3 camera motions combination cannot be recognized by the proposed method. As a future work, we



(a) ARPS process time



(b) Proposed method process time



(c) Total process time

Figure 3.18: The computational time in video "Station2"

are interested in classifying more than two camera motions.

Chapter 4

Automatic Retrieval of Attractive Moments in Sport Videos

4.1 Introduction

We are interested in searching for attractive scenes in sports videos. From several studies, the motion is one of the visual features that respond to an amount of human attention in videos [82]. The attractiveness in videos is simply measured by using MV magnitudes as motion activity. It is similar to existing works [18, 20, 83] that the motion activity is used for synthesizing human emotional space "arousal-valence" in order to find highlight videos.

In work [29], a framework of human attention model is presented to extract the attractive scenes using visual and audio features. Keyframes are extracted from entire sport videos using visual and contextual features, presented by Shih et al. [30]. Shih includes a game score in order to improve the determination of attractive moment, when the game score is changed. However, we realize that the attractive moment should happen when the game score is going to be changed rather than after the game score is changed. Therefore, we are motivated to propose a new video analysis method, which is more human perception.

In the direction of game theory, there are several models that examine the attractiveness in game progress [64, 65, 67, 68]. In past decade, Iida et al. had presented a logistic model that examines the attractiveness in discontinuous games such as Chess and Go [64, 67]. They concerned on the game information by

using game positions in a search tree. Recently, a mathematical model of game progress, which is inspired by the logistic model, was presented by Sutiono et al. [65, 68] based on the concept of uncertainty in game outcome. Sutiono's model not only examines the attractiveness in the discontinuous games but also examine the attractiveness in continuous games such as soccer, basketball, and volleyball. They construct a realistic game progress information in non-linear function that consists of two statistical information: 1) the number of score attempting or time to achieve the score, and 2) the number of scores. Then, they apply the mathematical operation "second derivative" to the game progress information to measure its acceleration or speed. The acceleration of the game progress information is similar to the acceleration in the second Newton's law [69]. They call the acceleration of the game progress information "game refinement value". This value relates to emotional impacts in human minds. Sutiono et al. also mentioned that

- The game progress information in linear functions is less attractiveness in game because the game outcome can be predicted easily.
- The game progress information in non-linear functions is more attractiveness in game because it is difficult to predict the game outcome.

It can be described by the mathematical operation "second derivative". The game progress information in linear function always has the acceleration equal to zero, while the game progress information in non-linear function always has the acceleration greater than zero. It means that slow speed in game progress information (i.e. low game refinement value) means the game information is hardly changed and its game outcome can be predicted easily. In contrast, high speed in game progress information (i.e. high game refinement value) means the game information is changed easily and its game outcome cannot be predicted easily. Therefore, They construct the realistic game progress information in non-linear function.

A mathematical model using probabilistic to find the attractiveness in soccer games was presented by Vecer [12]. This model is similar to [65, 68] that use statistical information. However, Vecer's model uses only game scores and time. Then, the winning, losing, and draw probabilities are computed. To find the attractive moments, changing in each probabilistic value is computed. If the probabilistic value is largely changed, it expects to be the attractive moments.

In contrast, it expects to be non-attractive moments if the probabilistic value has no changing or less changing.

In this chapter, we extend idea in Chapter 3 to find the attractive moments from the entire videos automatically using the camera motions. A model of the response of attractiveness was designed by using the camera motions in order to estimate the response curve of attractiveness. From the curve, the attractive moments can be extracted from the input video.

4.2 Methodology

Figure 4.1 presents the extended framework of the proposed method in Chapter 3. The proposed method consists of three main parts. First, we prepare the input data that are generated by the estimated camera motions in Chapter 3. We use the mathematical operation "convolution" [84] to compute response of attractiveness. Second, the response of attractiveness is used to select keyframes, which represent potentially attractive moments, from the video. Third, we select keyframes from the response curve and use them to generate a short video as an application.

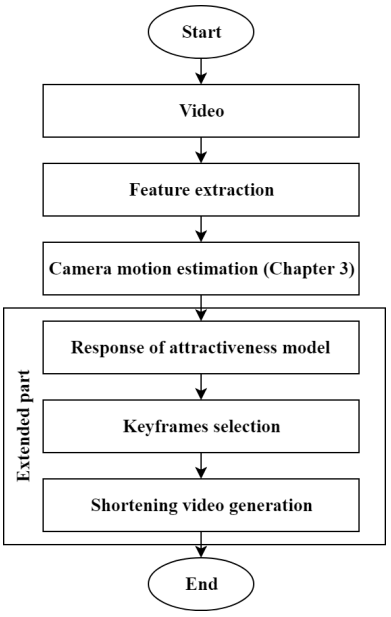


Figure 4.1: Extended framework of the proposed method

Refer to three kinds of uncertainty [10] and the three masters model [62], the master of playing and the uncertainty in information focus on the game progress.

They investigate the game progress to find the attractive moments, which the game information is uncertainty (i.e. unpredictable). By using the statistical information, the attractive moments are expected by changing in information (e.g. game scores and game results) during gameplay.

In video domains, information in sport games is recorded in the videos by the video makers whose have expert skills, especially in large official games. Changing in video information is more complex than changing in the statistical game information. Each video information contains several states or moments of game information. If we mention the uncertainty in information, the attractive moments in the video are expected by unusual moments during gameplay. The unusual moments are represented in a situation that cannot be occurred or be predicted easily. For examples, 1) viewers predict that a soccer player has an opportunity to make a score but they cannot predict that the referee comes to break the game due to a foul, and 2) no one cannot predict that stealing the ball from opponent players becomes a foul. Thus, we give the definition of the attractive moments in sport videos "The unusual behavioral moments that make a changing in game states and it is hard to predict their occurrence times". The camera motions, which are performed by the video makers, can guide the attractive moments potentially.

4.2.1 Response of Attractiveness Model

To generate the response of attractiveness, we use the mathematical operation "convolution" [84]. Convolution is the mathematical operation which uses two input functions to create the third function as the output. The convolution mathematical expression is expressed in Eq. (4.1).

$$(f * g)(t) = \int_0^t f(\tau)g(t-\tau)d\tau \tag{4.1}$$

where $f * g$ denotes the convolution result between functions f and g.

For the two input functions, the first function $F(n)$ is the estimated camera motion in Chapter 3. We simply create the binary function which has values between 0 and 1. Note that $F(n) = 1$ represents the n-th video frame that has the interested camera motion, while $F(n) = 0$ represents the n-th video frame that has the non-interested camera motions. As mentioned in the previous chapter, "Guironnet think that camera motion carries important information on video content" and "zooming camera motions make the viewers pay attention more than the other camera motions". We hypothesize that zooming camera motions have the most potential to retrieve the attractive moments from the video compare with the other camera motions. Thus, we describe the first function $F(n)$ by Eq. (4.2).

$$F(n) = \begin{cases} 1 & \text{if } C(n-1,n) \text{ is zoom-in or zoom-out camera motion} \\ 0 & \text{otherwise} \end{cases} \tag{4.2}$$

where $F(n)$ is the first function value of the camera motion at n-th video frame and $C(n-1,n)$ is the camera motion estimation result from Chapter 3, which is estimated by using current frame n and its reference video frame $n-1$ as described in [85].

For the second function, we use the mean filter functions. Two famous adjustable mean filter functions "Kaiser filter" [86] and "Gaussian filter" [87] are included for this experiments. Eqs. (4.3) and (4.4) show the mathematical expression of both Kaiser filter and Gaussian filter respectively.

$$K(n) = \begin{cases} \dfrac{I_0\left(\beta\sqrt{1-\left(\frac{n-(N/2)}{N/2}\right)^2}\right)}{I_0(\beta)} & \text{if } 0 \le n \le N \\ 0 & \text{otherwise} \end{cases} \tag{4.3}$$

where $K(n)$ is a value in the Kaiser filter, I_0 is the zero-th order modified Bassel function of the first kind, N is the length of the filter, and β is the shape parameter.

$$G(n) = \begin{cases} e^{-\frac{1}{2}\left(\alpha \frac{n}{(N-1)/2}\right)^2} & \text{if } 0 \leq n \leq N \\ 0 & \text{otherwise} \end{cases} \tag{4.4}$$

where $G(n)$ is a value in the Gaussian filter, N is the length of the filter, and α is the shape parameter.

From the mathematical expression, the two mean filters are illustrated in Figure 4.2 as an example. We use the convolution because the input data may have

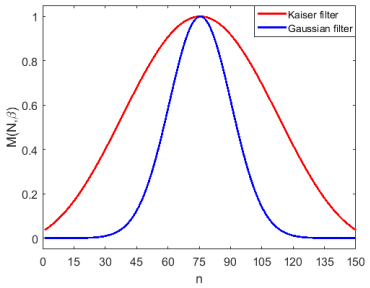


Figure 4.2: Mean filters with length $N = 150$ and shape parameter $\alpha = \beta = 5$

fluctuating values which are caused by video compression and video noises. When we generate the input function from the estimated camera motions, there are isolate '1' value in among of '0' values and isolate '0' value in among of '1' values. In mathematical idea, doing the convolution with the mean filters is the good choice to tackle these errors. Figure 4.3 shows a visual explanation of the convolution operation. We can simply explain that the convolution operation is computed by finding the overlapping area between the two input functions (i.e. camera motions and mean filter). When computing the convolution at video frame t, it finds the overlapping area the first function values from video frames $t-(N/2)$ to $t+(N/2)$. However, there is a problem in computing the convolution. Refer to our framework

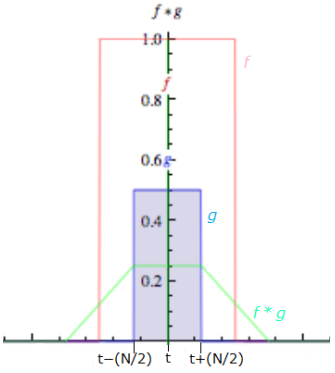


Figure 4.3: Visual explanation of the convolution operation

in Figure 4.1, we already know the camera motion until video frame t. Therefore, we cannot find the overlapping area from video frames $t+1$ to $t+(N/2)$. In order to solve the problem, we decide to cut the mean filter into a half to consider the overlapping area between video frames $t+1$ to $t+(N/2)$ equal to zero. Eqs. (4.5) and (4.6) show the modified of the Kaiser filter and Gaussian filter respectively.

$$K'(n) = \begin{cases} \frac{I_0\left(\beta\sqrt{1-\left(\frac{n-(N/2)}{N/2}\right)^2}\right)}{I_0(\beta)} & \text{if } 0 \leq n \leq N/2 \\ 0 & \text{otherwise} \end{cases} \tag{4.5}$$

where $K(n)$ is a value in the Kaiser filter, I_0 is the zero-th order modified Bassel function of the first kind, N is the length of the filter, and β is the shape parameter.

$$G'(n) = \begin{cases} e^{-\frac{1}{2}\left(\alpha\frac{n}{(N-1)/2}\right)^2} & \text{if } 0 \leq n \leq N/2 \\ 0 & \text{otherwise} \end{cases} \tag{4.6}$$

where $G(n)$ is a value in the Gaussian filter, N is the length of the filter, and α is the shape parameter.

After we compute the convolution between the function value of the camera motions and the mean filter, we get the third function which is the smoothen version of the first function.

$$\widetilde{F}(n) = F(n) * M(N, \beta) \tag{4.7}$$

where $\widetilde{F}(n)$ is the convolution result between the video signal $F(n)$ and the mean filter $M(N, \beta)$ (i.e. Kaiser filter $K'(n)$ and Gaussian filter $G'(n)$) with length N and parameter shape β.

Next, we normalize the convolution result following Eq. (4.8) in order to scale the value range from 0 to 1 as same as the original input data. We call the normalized convolution result "response of attractiveness". In each video frame, the response value is related to the attractive moments via the zooming camera motions.

$$F'(n) = \frac{\widetilde{F}(n)}{max(\widetilde{F}(n))} \tag{4.8}$$

where $F'(n)$ is the normalized convolution result, and $\widetilde{F}(n)$ is the convolution result.

4.2.2 Keyframes Selection and Shortening Video Generation

To extract keyframes from the response curve of attractiveness, we find the local maxima points in the third function (i.e. the response of attractiveness). Since our experiments are done by using MATLAB, we find the local maxima points by using MATLAB command "findpeaks" [88]. Each local maxima point contains the video frame index that is used to retrieve the video frame (i.e. keyframe), which has the potentially attractive moment.

From keyframes, we generate a short video version of the original video. First, we check the video frame index of keyframes and its next keyframe. If they have the close distance, we group them together to represent as a video clip. After we group all keyframes, we check the length of video clip. If they have the length shorter than two seconds, we exclude this video clip from the short video because it is difficult to watch in the real situation.

4.3 Results and Discussion

We use video sources of soccer matches from the Internet [89]. They are from UEFA Champion League 2015, UEFA Champion League 2016, and FIFA World Cup 2014. We use only video frames that are parts of the in playing game.

4.3.1 Response Curves of Attractiveness Over Video Frames

Figure 4.4 shows two response curves which are generated by using a video match in the first half of UEFA Champion League 2015. The two curves are: 1) the response curve that is generated by Kaiser filter and 2) the response curve that is generated by Gaussian filter. In the graph, x-axis represents the n-th minutes of video while the y-axis represents the response values which is Eq. (4.8). In this test video, the match starts before the 1st minute of the video and ends at before 46th minute of the video. We also show timestamps of in-game moments including of score attempting, goal, foul, card, corner kicks, and free kick. Note that these

moments are recorded by the official website [90]. In this video, there are 11 score attempting, 1 goal, 16 fouls, 2 cards, 6 corner kicks, and 1 free kick moments.

From Figure 4.4, we see that there are response of attractiveness not only at the timestamps but also out of the timestamps. It means that not only the moments at the timestamps, which are recorded by the official league, are potentially attractive but also there are some moments that are attractive. Comparing all timestamps of in-game moments, all score attempting moments have the response of attractiveness. The goal moment also has the response of attractiveness because it is one of the score attempting moments. There are 4 out of 16 fouls that do not have the response of attractiveness because of how serious in foul effect. For example, the first foul at the 1st minute is made by a player who tries to steal the ball from the opponent player, while the second foul at the 6th minute is made by a player accidentally slides to the opponent player's leg. All 2 card moments at 10th and 41st minutes have the response of attractiveness since they are serious in foul effect. For corner kick, 2 out of 6 moments at 29th and 35th minutes do not have the response of attractiveness because players have no opportunity to make a score from the corner kicks. Since players have a chance to make score attempting from the corner kick at the 16th minute, the response value of attractiveness is high. Comparing the two response curves, both curves have similar peak shape in the response of attractiveness. Refer to Figure 4.2, we use the mean filters with the same length and the same shape parameter. We have a smooth shape of mean filter (i.e. Kaiser filter) and a sharp shape of mean filter (i.e. Gaussian filter). In results, the response curves that are generated by Gaussian filter are steeper than the response curves that are generated by Kaiser filter.

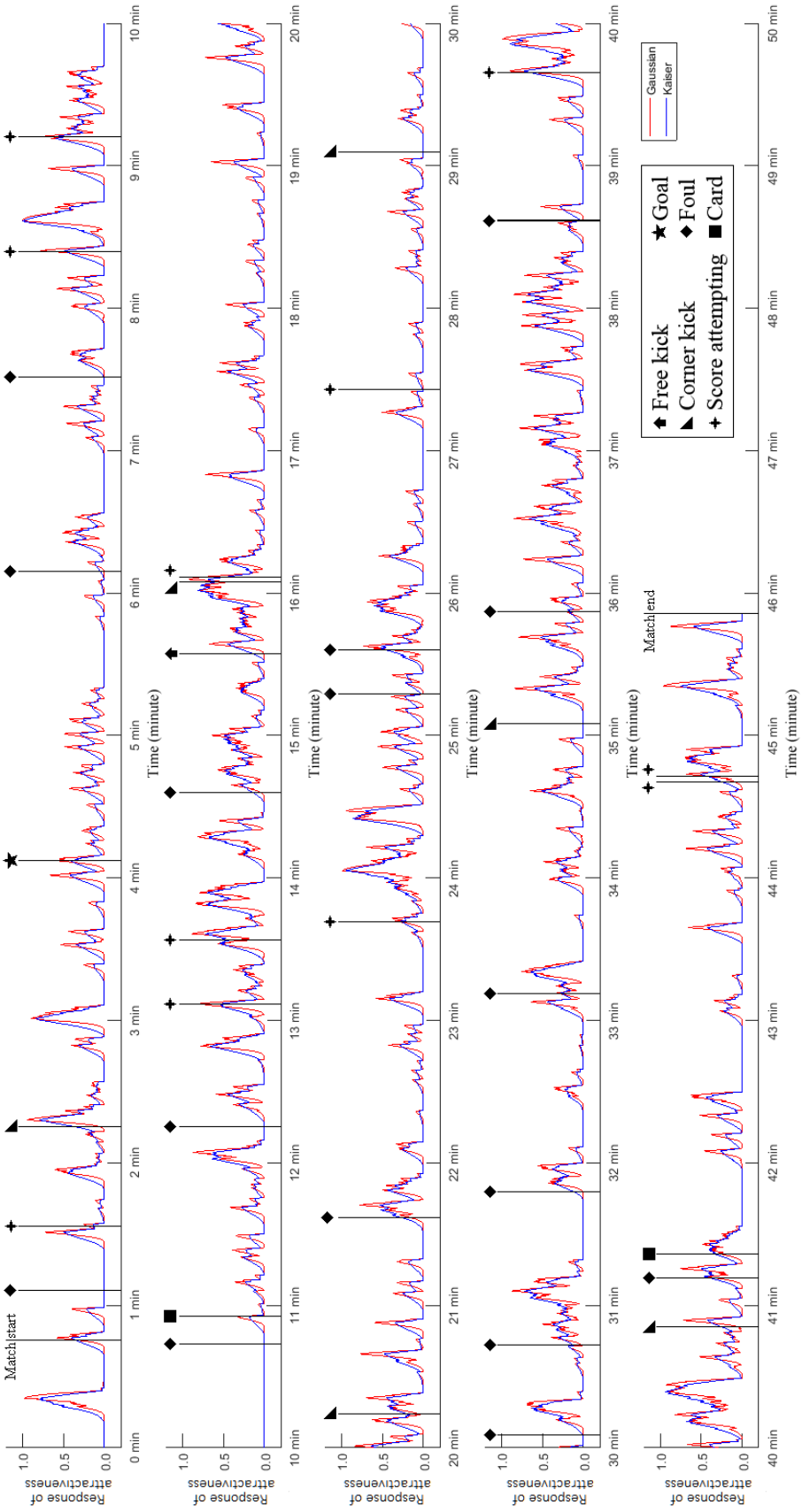


Figure 4.4: Response of attractiveness progress over video frames where the mean filter parameters are set $N = 150$ and $\alpha = \beta = 5$

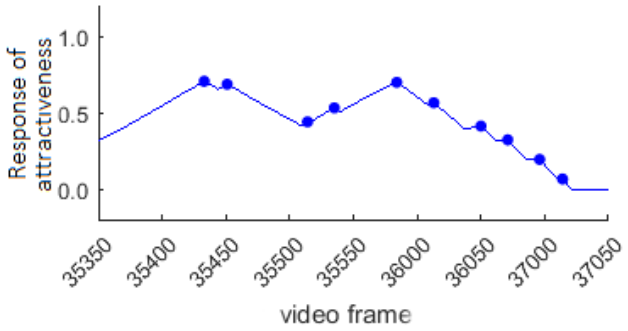
4.3.2 Keyframes Selection and Local Maxima Points

In this section, we discuss the two response curves in Figure 4.4. From the two curves, we found that both curves have the same indexes of local maxima points and the same total number of local maxima points. From the video test, we show three examples of attractive moments including of score attempting, foul, and a player claim to the referee's judgment moments as shown in Figures 4.5 to 4.7.

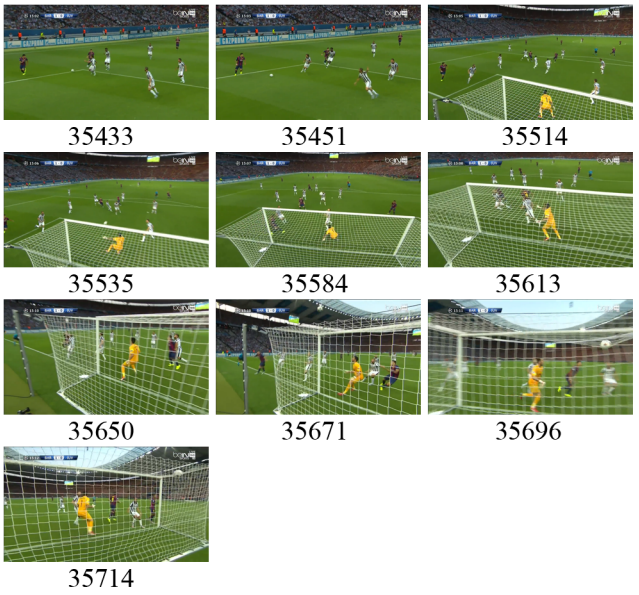
The video sequence from frame number 35,350 to 37,050 is a sequence of score attempting (Figure 4.5). In this sequence, the curves have 10 local maxima points (Figure 4.5(a)). From the curves, 10 keyframes are extracted from the video (Figure 4.5(b)). From each keyframe to the next keyframe, they are proceeded by zooming camera motions. At frame number 35,433, the zoom-out camera motion is used until the goal mouth is appeared at the frame number 35,514. Then the zoom-in camera motion is used to track the score attempting start from frame number until this moment is ending. Figure 4.6 shows a foul video sequence from frame number 65,000 to 65,200. There are six local maxima points in the response curve (Figure 4.6(a)). The keyframes in Figure 4.6(b) show that the foul moment starts at frame number 65,045. The camera utilizes slightly zooming with panning at the same time in order to follow the two players until one of the players fell down because of sliding as shown in frame number 65,106. From frame number 65,116, the camera operates zoom quickly in order to focus on the player who did the sliding. Figure 4.7 shows an argument between the player and referee from frame number 66,000 to 66,300. From Figure 4.7(a), there are six local maxima points in the response curve. Thus, the proposed method extracts six keyframes from the video as shown on Figure 4.7(b). At this moment, a player walks toward the referee because the player does not satisfy with the referee's judgment. The camera slightly operates zooming to the player starting from frame number 66,062 until the player can discuss with the referee at frame number 66,213.

4.3.3 Application in Video Shortening

We use keyframes in order to shorten the original video as an application. As mentioned in the previous section, the keyframes are used as video frame index in order to generate a video clip. If the two keyframes are close to each other,



(a) Response of attractiveness progress over video frames

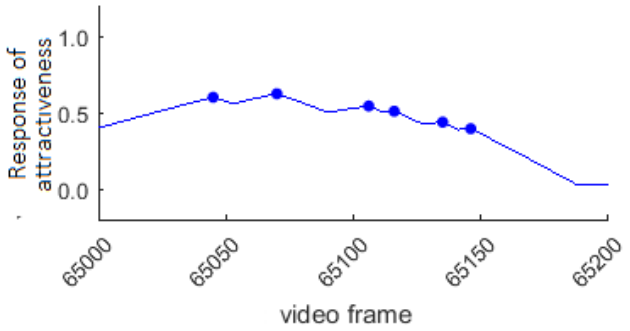


(b) Corresponding keyframes

Figure 4.5: An example of score attempting moment by the proposed method

we group them in order to represent as a video clip. The first keyframe and the lastest keyframe are set as the starting frame and ending frames of the video clip.

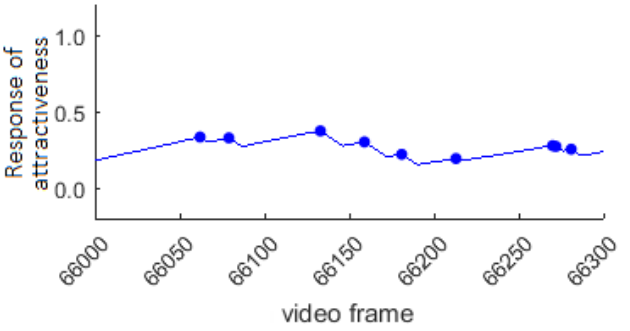
We use three full matches, totally six videos, from UEFA Champion League and FIFA World Cup. Each video contains the length of 45 minutes or a half game. After we observe the videos, we said that the toughness of the game is sorted by: 1st) UEFA Champion League 2016, 2nd) UEFA Champion League 2015, and 3rd) FIFA World Cup 2014. After we shorten these videos, Table 4.1 shows comparisons between the original video and shorten video in soccer videos. From all short videos, we got different video length depending on the toughness of the game. As we expect that UEFA Champion League 2015 have the longest shorten video length because it has the highest toughness of game. In the other word, there are several attractive moments if the matches have tough games.



(a) Response of attractiveness progress over video frames

(b) Corresponding keyframes

Figure 4.6: An example of foul moment by the proposed method



(a) Response of attractiveness progress over video frames

(b) Corresponding keyframes

Figure 4.7: An example of the player claim to the referee's judgment moment by the proposed method

Table 4.1: Comparisons between original and shorten videos in soccer matches

League titles	n-th half	Original (mins:sec)	Shorten (mins:sec)
UEFA 2015	1st half	45:51	11:52
	2nd half	52:51	12:41
UEFA 2016	1st half	46:27	15:04
	2nd half	48:47	17:04
FIFA World cup 2014	1st half	47:10	6:50
	2nd half	48:03	8:04
Total		289:09	71:35

4.4 Comparisons and Evaluations

4.4.1 Comparisons of Shorten Videos in Several Camera Motions

We show that why zooming camera motions are suitable to retrieve the attractive moments. First, we modify the first function in Eq. (4.2) for generating two additional shorten videos. First, the shorten video that contains stationary camera motions. Second, the shorten video that contains both panning and tilting camera motions. We use the video of the first half of the UEFA Champion League 2015 for this comparison. For each shorten video, the shorten video in stationary camera motions contains the length of 9 and a half minutes, the shorten video in panning and tilting camera motions contain contains the length of 27 minutes, and the shorten video in zooming camera motions contains the length of 11 minutes and 52 seconds. Figures 4.8 to 4.10 show thumbnail previews of three shorten videos of which are generated by Thumbnail me 3.0 [91]. In the thumbnail previews, there are 52 video frames which are selected in the same among of time.

In Figure 4.8, several far distances of viewpoints are in this video. Normally, the video makers did not move the cameras for capturing overall information. In some situations, the video makers hold the cameras after operating panning or zooming camera motions to broadcast results of each action. For example, goal keeper's face expression at time 0:50, a player is injured by sliding at time 1:00, score attempting at time 2:30, showing audiences' face expression at time 5:50, and players discuss with the referee at time 7:30.

The shorten video of panning and tilting camera motions (Figure 4.9) mainly shows about ball passing and ball dribbling because they are common behaviors in soccer that players have to bring the ball to the opponent's goal mouth for making a score. Thus, this video has the longest length compared with the other shorten videos. The score attempting moments (e.g. at times 5:02, 8:00, 15:30, and 23:32) are retrieved easily by using panning and tilting camera motions since all players have an opportunity to make a score after passing and dribbling the ball to the goal mouth.

Figure 4.10 shows several soccer moments in both far and close distances of viewpoints. Several kinds of soccer moments are retrieved by zooming camera motions. For example, ball is out of the field (e.g. at time 0:13), score attempting (e.g. at times 0:52, 1:44, 3:02, 10:11), fouls (e.g. at times 3:54, 7:09, and 9:06), and players discuss with the referee at time 8:53.

Table 4.2 summarizes both original and shorten videos in the UEFA Champion League 2015. In the original video, there are 214 video shots including of 24 score attempting, 4 goal, 12 corner kick, 32 foul, 3 card, 5 free kick, 6 player switching, and 128 nothing moments. In overall, the shortened video, which is made by panning and tilting camera motions, has the longest length compared with the other shorten videos (i.e. stationary and zooming camera motions).

In score attempting, the shorten videos of panning and tilting, and zooming camera motions can retrieve these moments more than the shortened video of stationary camera motions. Since the shortened video of stationary camera motion mainly shows results of each action as mentioned at above, several score attempting moments are not included in this shorten video.

Zooming camera motion has more potential to retrieve the goal moments than the other camera motions because video makers have to confirm that "is it a success in score attempting?". while the stationary camera motion may wait for results of the score attempting after operating zooming and panning camera motions. Therefore, the zooming camera motions have more opportunity to find the goal moments.

Figure 4.8: Thumbnail previews of shorten video which is generated by stationary camera motions

Figure 4.9: Thumbnail previews of shorten video which is generated by panning and tilting camera motions

Figure 4.10: Thumbnail previews of shorten video which is generated by zooming camera motions

From corner kick moments, all players have opportunities to make a score. They can make a score from the corner or make passing the ball to the players whose are close to the goal mouth in order to make a score. In shorten videos, corner kick moments are included by the zooming camera motion as well as the score attempting moments. Several corner kick moments are not available in stationary and panning camera motion because the camera is already in the good position for capturing the score attempting moments. The video makers may not move the camera for waiting the score attempting moments. Thus, zooming camera motions can retrieve the corner kick moments as well as score attempting moments.

Foul and card moments are included in the shorten video depending on how serious of foul effects as mentioned at the previous section. The video makers operate several camera operations in order to show how serious in fouls. Therefore, all shorten videos have the number of fouls video shots. Moreover, there is no card moment in the panning and tilting camera motion since the camera is at the good position to capture the referee.

The free kick is different from the corner kick that player is allowed to kick off the ball at foul positions, which can be occurred anywhere. Only shorten video of stationary camera motion has one free kick but there is no such opportunity to make a score. At this moment, it is different from the other moments that any situation can happen after the free kick. If players have an opportunity to make a score from the free kicks, the free kicks may be included in the shortened video.

Player switchings are available in the two shortened videos which made by panning and tilting, and zooming camera motions. Since the video makers operate zooming camera motion to the panel, they can confirm that which player is switched by whom. Panning and tilting camera motions also use to follow the player who is going to be switched. Therefore, these two kinds of camera motions are good to retrieve player switching moments.

Finally, the remaining 128 video shots that have no moments such score attempting, goal, corner, etc. They usually contain ball passing and ball dribbling. With these video shots, we can decide that which camera motion has the potential to retrieve the attractive moments in soccer videos. From the Table 4.2, we decide that zooming camera motion has potential to retrieve the attractive moments because the shorten videos in stationary, and panning and tilting camera motions contain nothing moments more than a half of the video length.

Table 4.2: Comparison between the original video and three shortened videos which are made by each camera motion

Moments	Number of moments in the original video	Number of moments in shorten videos which are generated by		
		Stationary	Panning & tilting	Zooming
Score attempting	24	7	20	18
Goal	4	2	2	3
Corner kick	12	3	1	7
Foul	32	10	8	8
Card	3	1	0	1
Free kick	5	1	0	0
Player switching	6	3	5	6
Nothing	128	86	125	49
Total	214	113	161	92

4.4.2 Comparisons of Attractive Moments Retrieval in Several Methods

Table 4.3 shows a comparison in the retrieval of attractive moments between the proposed method and existing methods. There are two kinds of models: 1) statistical models (i.e. models in works [12, 68]) and 2) video analysis models (i.e. models in works [4, 92, 93] and the proposed method).

Table 4.3: Comparison in retrieval of attractive moments

Methods	Input	Pre processing	Attractive moments retrieval approaches
[12, 68]	Statistical game information	No	Mathematical approach
[4]	Sport videos by professional video makers	Motion and sound measurement	Mathematical approach
[92]	Home videos by non-professional video makers	Parametic camera motion estimation	Counting the number of video frames
[93]	Home videos by non-professional video makers	Parametic camera motion esitmation	Rule based
Proposed method	Sport videos by professional video makers	Non-parametic camera motion estimation	Mathematical approach

In the two mathematical models [12, 68], they have the similar idea with the proposed but we use the camera motion instead of the game score. The model in [68] uses the number of game scores and the number of score attempting while the model in [12] uses the number of game scores and time to find the attractive moments. Both models retrieve the attractive moments in sport games or sport videos where the game score is changed because they expect that changing in game score also changes the game outcome. The speed of the game progress information [68] represents the attractiveness in game while changing in probabilistic values [12] represents the attractiveness in game. The proposed method not only expects the changing in game score to be the attractive moments but also expects the other situations to be the attractive moments (e.g. score attempting) via the camera motions. Since the videos are recorded by professional video makers, they know that what kind of camera motion should be operated for each situation.

Figure 4.11 shows a timeline in soccer moments when a game score is successfully made. The sequence of this moment usually proceeds with dribbling the ball to the goal mouth, kicking the ball to make a score, the ball is in the goal mouth, and updating the game score. In the two mathematical models, the attractive moment can be recognized after the 4th time period. The proposed method recognizes the attractive moments at between the 1st and 2nd periods, and between the 2nd and 3rd periods. Since the video makers realize the score attempting moment, they operate the camera motions near the 2nd period. In the real situation, the moments at between the 1st and 2nd periods, and between the 2nd and 3rd periods are more potentially attractive than the moment after the 4th period because the result at the 3rd period is difficult to predict. Assume

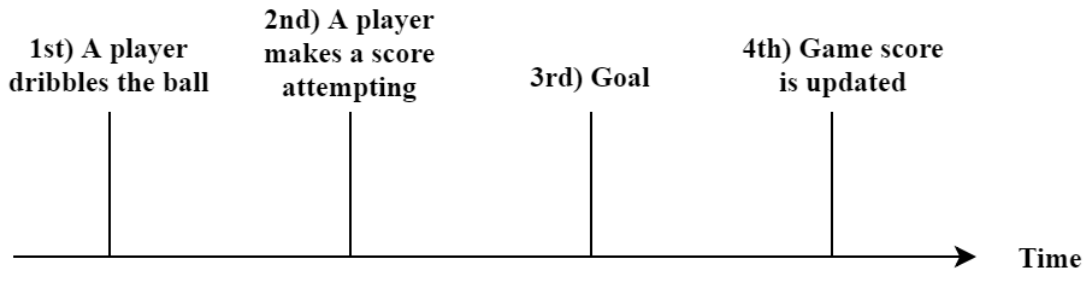


Figure 4.11: Timeline in soccer moments when a game score is successfully made

that if the score is unsuccessfully made (i.e. the 3rd and 4th periods are not in the timeline), the model in [12] misses this moment because the game score is not changed. The model in [68] may notice this moment because of the number

of score attempting but it may not considers as the attractive moments because the speed of game progress information is decreased. The proposed method still recognizes the attractive moments near the 2nd period as mentioned above.

Hanjalic's method [4] is a video analysis to find the attractive moments in sport videos using motion and sound. We use the same soccer video "UEFA Champion League 2015" for comparison. To retrieve the attractive moments, the local maxima points in each response are localized. Figure 4.12 shows the responses of attractiveness progress in the first half of the UEFA Champion League 2015 for each feature. They consist of zooming camera motion (i.e. the proposed method), motion, sound, and the combination of motion and sound. From Figure, most of the response in the motion has peaks at the middle of two peaks of the response in the zooming camera motion. In other words, there are high responses of attractiveness in the motion when the camera operates panning and tilting camera motions. Note that there are no responses of attractiveness in the motion when it is stationary camera motion. Figure 4.13 shows the attractive moments that are extracted from the response of attractiveness in the motion. It has the attractive moments as same as the attractive moments in the panning and tilting camera motions. However, the attractive moments in the motion mostly contain close-up viewpoints. Although the motion feature can extract the attractive moments in close-up viewpoints, it consists of nothing moments as same as the attractive moments in the panning and tilting camera motions. For sound feature, high sound energy in the human perspective is more attractive. However, it is difficult to find the attractive moments using sound feature because the input video mainly contain the similar energy levels of sound from audiences and there is no sound from commentators. The response of attractiveness in sound has curves similar to the straight line. Therefore, the attractive moments cannot be found by using sound. For the combination of motion and sound, both responses of attractiveness in motion and sound are combined. However, it has the responses as same as the motion because of the sound. Therefore, the attractive moments can be extracted from the motion only.

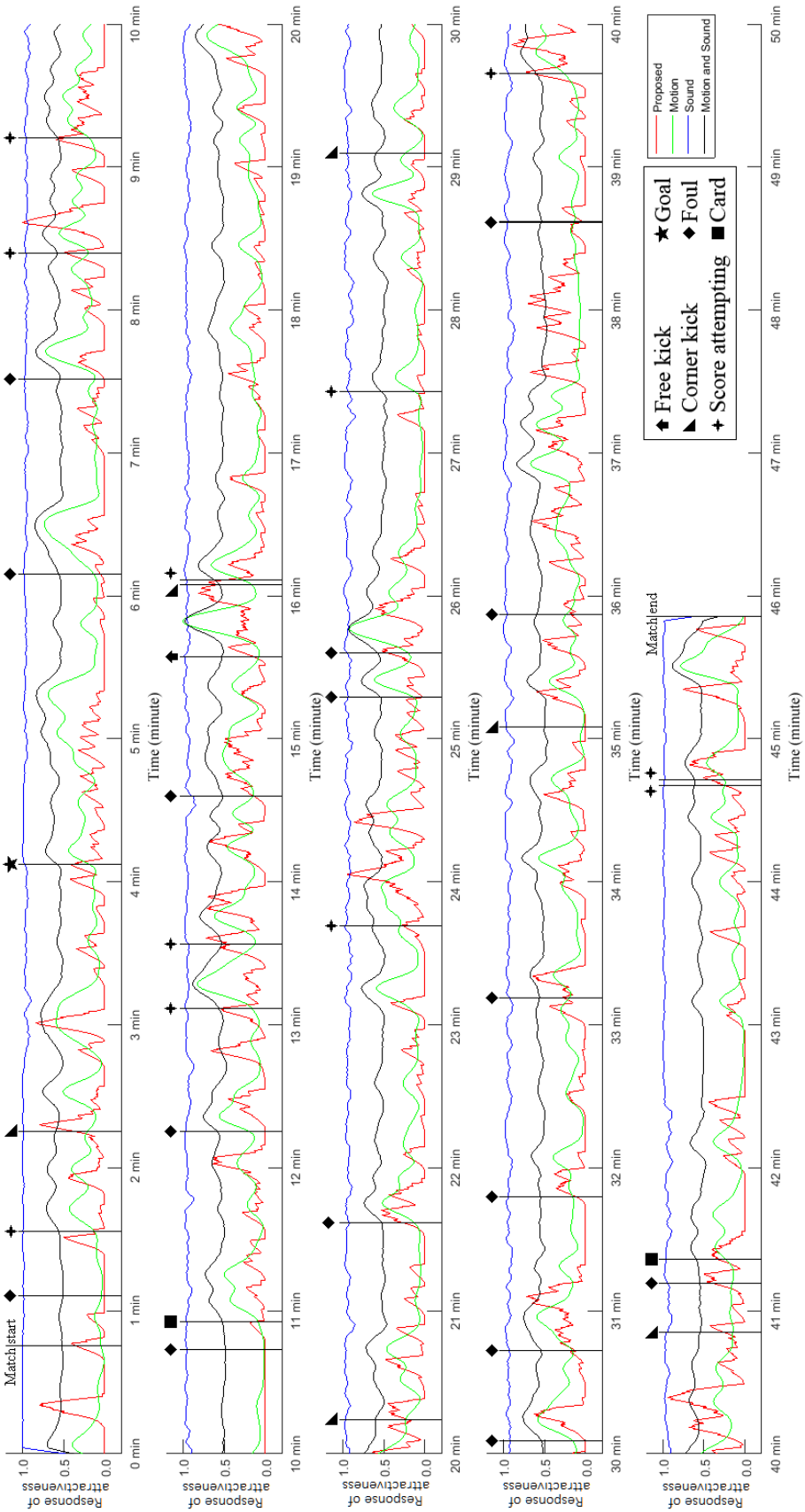


Figure 4.12: Responses of attractiveness progress over video frames for each feature

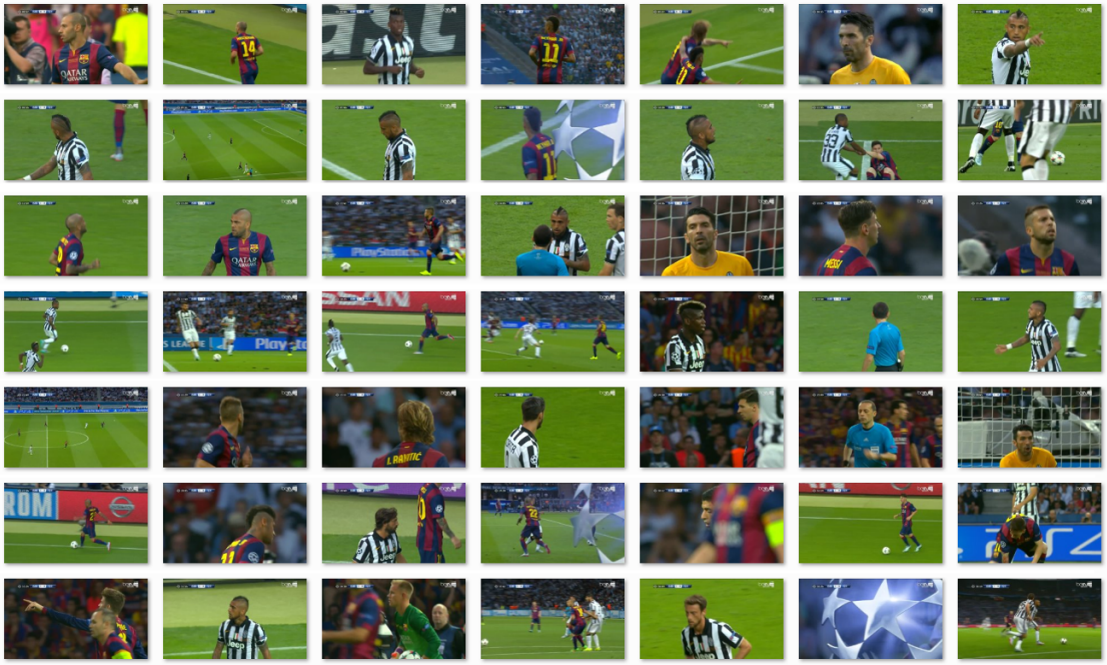


Figure 4.13: Attractive moments which are extracted by the response of attractiveness in the motion

In the other video analysis models [92, 93], they put efforts to find the attractive moments in home videos, which are made by non-professional video makers, using zooming camera motions as same as the proposed method. Both existing methods retrieve the attractive moments in video frames while the proposed method retrieve the attractive moments in both video frames and a shortened video. The model in [92] finds the video frame of attractive moments which has patterns of one second of zoom-in then followed by two seconds of stationary camera motions, while the model in [93] finds the video frame of attractive moments after operating zooming camera operations with two conditions. First, a video frame after zoom-in camera motion must be considered as the attractive moments. Second, a video frame after zoom-out camera motion with appropriate operational speed is considered as the attractive moments. For comparison, the proposed method considers both zoom-in and zoom-out camera motions as same as the model in [93] while the model in [92] considers only zoom-in camera motions. Moreover, during zooming camera motions are considered as the attractive moments by the proposed method which is different from the two existing models that during zooming camera motion are not considered as the attractive moments. Because of difference in output, the two existing models select the video frames of the attractive moments after the camera motions which have more clarity in images while the proposed method retrieves

the shorten videos of the attractive moments during the camera motions which contain their story.

4.4.3 Subjective Evaluation

In this subjective evaluation, six volunteer subjects are invited to evaluate the effectiveness of the proposed method. This evaluation aims to evaluate the attractiveness of the test videos for the future study. We let the subjects watch the test videos that all subjects never see the test video before. We use the shorten videos from UEFA Champion League 2015 as the test videos for this evaluation. They are the 1st half and 2nd half of the game, totally 22 minutes of video length. Therefore, this evaluation takes time at least 22 minutes for each volunteer subject.

In steps of the evaluation, first, the subjects watch each video shot in the test videos as shown in Figure 4.14. Then, they are asked by the questions "Is the n-th

Figure 4.14: A video shot in the test videos with n-th video shot at upper left

video shot attractive or exciting to them or not?" Finally, the subjects will select video shots inside the test videos that they feel attracting or exciting.

From the evaluation, in the 1st half, the subjects had selected averagely 25 video shots that they feel attractive and exciting. Most of them are in the score attempting, goal, and corner kick moments. Note that the subjects feel attractive to the corner kick moments when there is a score attempting. In the 2nd half, the subjects had selected averagely 24 video shots that can attract their mind. It is also similar to the 1st half that most of the selected video shots are in the moments of score attempting.

4.5 Conclusions

In this chapter, an extension of work in Chapter 3 is presented as an automatic retrieval of attractive moments in sport video using camera motions. At first, we define the attractive moments "The unusual behavioral moments that make a changing in game states and it is hard to predict their occurrence times". We design the mathematical model that the operator "convolution" is used for analyzing the response of attractiveness via the camera motion. In the comparison of the three camera motions, zooming camera motion is suitable to retrieve the attractive moments. Not only changing in game score (i.e. goal) is included in the attractive moments but the other situations (e.g. score attempting, foul, and claiming to the referee's judgment) are also included in the attractive moments. Selected keyframes, which have the same story or close to each other, are used for making a shorten video as an application. We had compared the proposed method with the existing works. From the comparison, our proposed method can recognize the attractive moments more realistic because of cues in camera motions that are given by the video makers. We also apply subjective evaluation to the proposed method in order to evaluate the attractiveness in generated shorten video for future study. From evaluation results, even the proposed method generates the shorten video consisting of several kinds of the attractive moments, the subjects are mainly attracted by the score attempting, goal, and corner kick with score attempting moments. In future work, we will improve the proposed method to be able to filter the shortened video and make the better shorten video.

Chapter 5

Conclusions and Future Works

In this chapter, we give the conclusions in this dissertation and answers the research questions.

5.1 Summary

- **Chapter 2: Literature review**
  In this chapter, we discuss the two research questions that are given in the Chapter 1. The first question discusses how to extract the camera motions in the videos. The second question discusses how the attractive moments can be retrieved by the camera motion. In the way of extracting the camera motions, non-parametric models (i.e. 1D MV histograms) are popular techniques to extract the camera motions in the videos. In general, the video frame is divided into several sub-images equally. Then, the dominant histogram peak of each sub-image can lead to stationary, panning, tilting, and zooming camera motion patterns. However, we found that only zooming camera motion need to analyze by using sub-images because stationary, panning, and tilting camera motions can be fast classified by the dominant MV histogram of the entire image.

  The human emotional space "arousal-valence" in psychophysiological study area shows that the motions in picture have a relationship with the affective impact of viewers. It shows that the arousal is affected by the motions in picture directly. Then, a paradigm of the arousal-valence space in the videos is presented by using motion and sound features. By this emotional space,

the attractive moments are retrieved when the video contains the moments with high motion activity and high sound energy. However, some attractive moments are not involved because of lack of meaning.

To improve the performance of the retrieval of attractive moments, we found that camera motion carries important information on video contents. Especially, zooming camera motions make the viewers pay more attention when they are watching videos. Since video makers realize an attractiveness in unusual behavior moments, they operate zooming camera motions in order to emphasize the information clearly. There is an existing model that find the attractive moments using the camera motions. Basically, the start frame non-stationary camera motions

In other study areas, uncertainty in information is the key-factor to measure the attractiveness in the games. People or the viewers are attracted by the games because of unpredictable in game results. Because of an action in the games, the game states are changed. It is similar to the situation that the video makers operate the camera motion because they realize the attractive moments. Thus, we hypothesized that the zooming camera motion has the relationship with the attractive moment.

- **Chapter 3: Extracting camera motion using 2D MV histogram**
  We present an original method to extract the camera motions from the videos using 2D MV histogram. The 2D MV histogram contains both MV magnitude and MV orientation information. The proposed method can extract both single camera motion and combination of two camera motions from the videos. For the single camera motion, they are stationary, panning, tilting diagonal panning, zooming, object tracking, and scene change camera motions. Combination of two camera motions (e.g. zooming with panning, and zooming with tilting camera motions) can be classified by the proposed method. Comparing with the existing methods, the proposed method can extract more types of the camera motions with better performance.

- **Chapter 4: Detecting attractive moments in soccer video using camera motions**
  The idea in Chapter 3 is extended for an automatic retrieval of attractive moments in sport videos. We use soccer video for experiments. We design the mathematical model to find the response of the attractiveness via the camera motion. Convolution is the operator that can compute the response of the

attractiveness. From the responding curve, we can retrieve keyframes of the attractive moments in soccer video. Comparing with the existing methods, they mainly focus that changing in game score has the attractiveness. In reality, the other moments (e.g. score attempting, foul, etc) can be considered as the attractive moments as the proposed method had done. We also compare the performance of the proposed method by using three kinds of the camera motion. It shows that zooming camera is suitable to retrieve the attractive moments in soccer video as the given hypothesis.

5.2 Answer to the Research Questions

- **Research question 1: How to design a model that can extract the camera motions from the video?**
  From Chapter 3, we develop an original technique to extract the camera motions from the video using 2D MV histogram instead of 1D MV histogram. Because of 2D MV histogram, complex patterns in single camera motion (e.g. object tracking and scene change camera motions) and a combination of two camera motions (i.e. zooming with panning and zooming with tilting camera motions) can be recognized. Although zoom-in and zoom-out camera motion have the same 2D MV histogram, we can solve this limitation by using the MV fields that are used as the input of the proposed method. Thus, the proposed method classifies the converging MV field as zoom-out camera motion and classifies the diverging MV field as zoom-in camera motion.

- **Research question 2: How can the attractive moments be retrieved by the camera motions?**
  From Chapter 4, we extend the method in Chapter 3 to retrieve the attractive moments in soccer video automatically. We decide to use stationary, panning, tilting, and zooming camera motions in order to investigate the relationship between the camera motions and the attractive moments in soccer videos. The outcome of this study is presented by: 1) keyframes of attractive moments in several situations and 2) a shorten video that contains only the attractive moments.

5.3 Future Works

- **Extracting Camera Motions Using 2D MV Histogram**
  As shown in the experiments, we found the limitation of the proposed method that combination of three camera motions (e.g. zooming with panning and object tracking camera motion) is not successfully classified by the proposed method. Therefore, this is the first future work of this study in order to classify the combination of three camera motions. We also consider the usage of the 2D MV histogram. Currently, we use a fixed threshold value that comes from iterative experiments. We might improve the proposed method in two ways: 1) using the 2D MV histogram without thresholding and 2) to make the proposed method adjust the threshold value automatically.

- **Detecting attractive moments in soccer video using camera motions**
  We had generated the shorten soccer video that contains the unusual behavior moments (i.e. the attractive moments). It consists of score attempting, goal, corner kick, foul, etc. After we evaluate the shorten video by subjective evaluation, we found that most subjects are attracted by the score attempting moments. For the shortening video application, we will improve the performance in order to make a better video based on the human perspective. Form the experiments, we have already known that zooming camera motions have the potential to find the attractive moments in the soccer video. It is very curious that between zoom-in and zoom-out camera motion which camera motion is more related to the attractive moments.

Bibliography

[1] M. A. Hasan, M. Xu, X. He, and C. Xu. Camhid: Camera motion histogram descriptor and its application to cinematographic shot classification. *IEEE Transactions on Circuits and Systems for Video Technology*, 24(10): 1682–1695, Oct 2014.

[2] M. Okade, G. Patel, and P. K. Biswas. Robust learning-based camera motion characterization scheme with applications to video stabilization. *IEEE Transactions on Circuits and Systems for Video Technology*, 26(3):453–466, Mar 2016.

[3] M. Guironnet, D. Pellerin, N. Guyader, and P. Ladret. Video summarization based on camera motion and a subjective evaluation method. *EURASIP Journal on Image and Video Processing*, 2007(1):060245, Jun 2007.

[4] A. Hanjalic. Adaptive extraction of highlights from a sport video based on excitement modeling. *IEEE Transactions on Multimedia*, 7(6):1114–1122, Dec 2005.

[5] B. H. Detenber, R. F. Simons, and G. G. Bennett Jr. Roll em!: the effects of picture motion on emotional responses. *Journal of Broadcasting & Electronic Media*, 42(1):113–127, 1998.

[6] R. F. Simons, B. H. Detenber, T. M. Roedema, and J. E. Reiss. Emotion processing in three systems: the medium and the message. *Psychophysiology*, 36(5):619–627, 2003.

[7] S. KnoblochWesterwick, P. David, M. S. Eastin, R. Tamborini, and D. Greenwood. Sports spectators' suspense: affect and uncertainty in sports entertainment. *Journal of Communication*, 59(4):750–767, Dec 2009.

[8] P. Majek and H. Iida. Uncertainty of game outcome. In *3rd International Conference on Global Research and Education in Intelligent Systems*, pages 171–180, 2004.

[9] G. Costikyan. *Uncertainty in games*. The MIT Press, 2013.

[10] A. Cincotti and H. Iida. Outcome uncertainty and interestedness in game-playing: a case study using synchronized hex. *New Mathematics and Natural Computation*, 2(2):173–181, Jul 2006.

[11] E.-J. Kim, G.-G. Lee, C. Jung, S.-K. Kim, J.-Y. Kim, and W.-Y. Kim. A video summarization method for basketball game. In *Advances in Multimedia Information Processing - PCM 2005*, pages 765–775, 2005.

[12] J. Vecer, T. Ichiba, and M. Laudanovic. On probabilistic excitement of sports games. *Journal of Quantitative Analysis in Sports*, 3(3):1–23, 2007.

[13] G. G. Lee, H. k. Kim, and W. Y. Kim. Highlight generation for basketball video using probabilistic excitement. In *2009 IEEE International Conference on Multimedia and Expo*, pages 318–321, Jun 2009.

[14] H. Iida, T. Nakagawa, K. Spoerer, and S. Sone. Three elemental game progress patterns. In *Intelligent Science and Intelligent Data Engineering*, pages 571–581, 2012.

[15] T. R. M. Nakagawa and H. Iida. Three game patterns. *International Journal of Research Studies in Computer Science and Engineering (IJRSCSE)*, 1(1): 1–12, May 2014.

[16] J. A. Russell. Affective space is bipolar. *Journal of personality and social psychology*, 37(3):345–356, 1979.

[17] J. A. Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161–1178, 1980.

[18] A. Hanjalic and Li-Qun Xu. Affective video content representation and modeling. *IEEE Transactions on Multimedia*, 7(1):143–154, Feb 2005.

[19] S. Zhang, Q. Tian, S. Jiang, Q. Huang, and W. Gao. Affective mtv analysis based on arousal and valence features. In *2008 IEEE International Conference on Multimedia and Expo*, pages 1369–1372, Jun 2008.

[20] K. Sun, J. Yu, Y. Huang, and X. Hu. An improved valence-arousal emotion space for video affective content representation and recognition. In *2009 IEEE International Conference on Multimedia and Expo*, pages 566–569, Jun 2009.

[21] D. S. Tan, S. See, and T. J. Tiam-Lee. Automatic rating of movies using an arousal curve extracted from video features. In *2014 International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment and Management (HNICEM)*, pages 1–6, Nov 2014.

[22] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259, Nov 1998.

[23] D. Walther and C. Koch. Modeling attention to salient proto-objects. *Neural Networks*, 19(9):1395–1407, 2006.

[24] F. Dirfaux. Key frame selection to represent a video. In *Proceedings 2000 International Conference on Image Processing (Cat. No.00CH37101)*, volume 2, pages 275–278, Sep 2000.

[25] A. A. Salah, E. Alpaydin, and L. Akarun. A selective attention-based method for visual pattern recognition with application to handwritten digit recognition and face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(3):420–425, Mar 2002.

[26] H. R. Tavakoli, E. Rahtu, and J. Heikkilä. Temporal saliency for fast motion detection. In *Computer Vision - ACCV 2012 Workshops*, pages 321–326, 2013.

[27] H. C. Shih, C. L. Huang, and J. N. Hwang. Video attention ranking using visual and contextual attention model for content-based sports videos mining. In *2007 IEEE 9th Workshop on Multimedia Signal Processing*, pages 414–417, Oct 2007.

[28] H. C. Shih, J. N. Hwang, and C. L. Huang. Content-based attention ranking using visual and contextual attention model for baseball videos. *IEEE Transactions on Multimedia*, 11(2):244–255, Feb 2009.

[29] Y.-F. Ma, L. Lu, H.-J. Zhang, and M. Li. A user attention model for video summarization. In *Proceedings of the Tenth ACM International Conference on Multimedia*, pages 533–542, 2002.

[30] H. C. Shih. A novel attention-based key-frame determination method. *IEEE Transactions on Broadcasting*, 59(3):556–562, Sep 2013.

[31] Y.-F. Ma, X.-S. Hua, L. Lu, and H.-J. Zhang. A generic framework of user attention model and its application in video summarization. *IEEE Transactions on Multimedia*, 7(5):907–919, Oct 2005.

[32] G. Abdollahian, Z. Pizlo, and E. J. Delp. A study on the effect of camera motion on human visual attention. In *2008 15th IEEE International Conference on Image Processing*, pages 693–696, Oct 2008.

[33] M.V. Srinivasan, S. Venkatesh, and R. Hosie. Qualitative estimation of camera motion parameters from video sequences. *Pattern Recognition*, 30(4):593–606, 1997.

[34] R. Wang and T. Huang. Fast camera motion analysis in mpeg domain. In *Proceedings 1999 International Conference on Image Processing (Cat. 99CH36348)*, volume 3, pages 691–694, 1999.

[35] J.-G. Kim, H. S. Chang, J. Kim, and H.-M. Kim. Efficient camera motion characterization for mpeg video indexing. In *2000 IEEE International Conference on Multimedia and Expo. ICME2000. Proceedings. Latest Advances in the Fast Changing World of Multimedia (Cat. No.00TH8532)*, volume 2, pages 1171–1174, 2000.

[36] R. Ewerth, M. Schwalb, P. Tessmann, and B. Freisleben. Estimation of arbitrary camera motion in mpeg videos. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, volume 1, pages 512–515, Aug 2004.

[37] S. Nikitidis, S. Zafeiriou, and I. Pitas. Camera motion estimation using a novel online vector field model in particle filters. *IEEE Transactions on Circuits and Systems for Video Technology*, 18(8):1028–1039, Aug 2008.

[38] J. Almeida, R. Minetto, T. A. Almeida, R. S. Torres, , and N. J. Leite. Estimation of camera parameters in video sequences with a large amount of scene motion. In *2010 17th International Conference on Systems, Signals and Image Processing (IWSSIP 2010)*, pages 348–358, Jun 2010.

[39] Y.-P. Tan, D. D. Saur, S. R. Kulkami, and P. J. Ramadge. Rapid estimation of camera motion from compressed video with application to video annotation.

IEEE Transactions on Circuits and Systems for Video Technology, 10(1): 133–146, Feb 2000.

[40] X. Zhu, A. K. Elmagarmid, X. Xue, L. Wu, and A. C. Catlin. Insightvideo: toward hierarchical video content organization for efficient browsing, summarization and retrieval. *IEEE Transactions on Multimedia*, 7 (4):648–666, Aug 2005.

[41] Y. Weng and J. Jiang. Fast camera motion estimation in mpeg compressed domain. *IEEE Transactions on Consumer Electronics*, 57(3):1329–1335, August 2011.

[42] A. Mahabalagiri, K. Ozcan, and S. Velipasalar. Camera motion detection for mobile smart cameras using segmented edge-based optical flow. In *2014 11th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 271–276, Aug 2014.

[43] S. Erturk. Digital image stabilization with sub-image phase correlation based global motion estimation. *IEEE Transactions on Consumer Electronics*, 49 (4):1320–1325, Nov 2003.

[44] S. Lee and M. H. Hayes. Real-time camera motion classification for content-based indexing and retrieval using templates. In *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 4, pages IV–3664–IV–3667, May 2002.

[45] N.-T. Nguyen, D. Laurendeau, and A. Branzan-Albu. A robust method for camera motion estimation in movies based on optical flow. *International Journal of Intelligent Systems Technologies and Applications*, 9(3/4):228–238, Nov 2010.

[46] N. V. Patel and I. K. Sethi. Video shot detection and characterization for video databases. *Pattern Recognition*, 30(4):583–592, 1997.

[47] L.-Y. Duan, J. S. Jin, Q. Tian, and C.-S. Xu. Nonparametric motion characterization for robust classification of camera motion patterns. *IEEE Transactions on Multimedia*, 8(2):323–340, Apr 2006.

[48] T. T. de Souza and R. Goularte. Video shot representation based on histograms. In *Proceedings of the 28th Annual ACM Symposium on Applied Computing*, pages 961–966, 2013.

[49] M. Tavassolipour, M. Karimian, and S. Kasaei. Event detection and summarization in soccer videos using bayesian network and copula. *IEEE Transactions on Circuits and Systems for Video Technology*, 24(2):291–304, Feb 2014.

[50] Y. Bendraou, F. Essannouni, D. Aboutajdine, and A. Salam. Video shot boundary detection method using histogram differences and local image descriptor. In *2014 Second World Conference on Complex Systems (WCCS)*, pages 665–670, Nov 2014.

[51] M. A. Hasan, M. Xu, X. He, and Y. Wang. A camera motion histogram descriptor for video shot classification. *Multimedia Tools and Applications*, 74(24):11073–11098, Dec 2015.

[52] Y. Nie and K.-K. Ma. Adaptive rood pattern search for fast block-matching motion estimation. *IEEE Transactions on Image Processing*, 11(12): 1442–1449, Dec 2002.

[53] R. Li, B. Zeng, and M. L. Liou. A new three-step search algorithm for block motion estimation. *IEEE Transactions on Circuits and Systems for Video Technology*, 4(4):438–442, Aug 1994.

[54] L.-M. Po and W.-C. Ma. A novel four-step search algorithm for fast block motion estimation. *IEEE Transactions on Circuits and Systems for Video Technology*, 6(3):313–317, Jun 1996.

[55] J. Lu and M. L. Liou. A simple and efficient search algorithm for block-matching motion estimation. *IEEE Transactions on Circuits and Systems for Video Technology*, 7(2):429–433, Apr 1997.

[56] S. Zhu and K.-K. Ma. A new diamond search algorithm for fast block-matching motion estimation. *IEEE Transactions on Image Processing*, 9(2):287–290, Feb 2000.

[57] C.-H. Cheung and L.-M. Po. A novel small-cross-diamond search algorithm for fast video coding and videoconferencing applications. In *Proceedings. International Conference on Image Processing*, volume 1, pages I–681–I–684, Sep 2002.

[58] C.-H. Cheung and L.-M. Po. A novel cross-diamond search algorithm for fast block motion estimation. *IEEE Transactions on Circuits and Systems for Video Technology*, 12(12):1168–1177, Dec 2002.

[59] C.-W. Lam, L.-M. Po, and C. H. Cheung. A new cross-diamond search algorithm for fast block matching motion estimation. In *International Conference on Neural Networks and Signal Processing, 2003. Proceedings of the 2003*, volume 2, pages 1262–1265, Dec 2003.

[60] M. Ghanbari. *Standard codecs: image compression to advanced video coding*. Institution Electrical Engineers, 2003.

[61] A. Hanjalic. Extracting moods from pictures and sounds: towards truly personalized tv. *IEEE Signal Processing Magazine*, 23(2):90–100, Mar 2006.

[62] H. Iida. Fairness, judges and thrill in games. *The Special Interest Group Technical Reports of Information Processing Society of Japan (IPSJ SIG Technical Reports)*, 2008(28):61–68, Mar 2008.

[63] J. von Neumann and O. Morgenstern. *Theory of Games and Economic Behavior*. Princeton University Press, 1944.

[64] H. Iida, K. Takahara, J. Nagashima, Y. Kajihara, and T. Hashimoto. An application of game-refinement theory to mah jong. In *Entertainment Computing – ICEC 2004*, pages 333–338, 2004.

[65] A. P. Sutiono, A. Purwarianti, and H. Iida. A mathematical model of game refinement. In *Intelligent Technologies for Interactive Entertainment*, pages 148–151, 2014.

[66] H. Iida, T. Nakagawa, and K. Spoerer. Game information dynamic models based on fluid mechanics. *Entertainment Computing*, 3(3):89–99, Aug 2012.

[67] H. Iida, N. Takeshita, and J. Yoshimura. A metric for entertainment of boardgames: its implication for evolution of chess variants. *Entertainment Computing: Technologies and Application*, pages 65–72, 2003.

[68] A. P. Sutiono, R. Ramadan, P. Jarukasetporn, J. Takeuchi, A. Purwarianti, and H. Iida. A mathematical model of game refinement and its applications to sports games. *EAI Endorsed Transactions on Creative Technologies*, 2(5): 1–7, Oct 2015.

[69] I. Newton, I. B. Cohen, and A. Whitman. *The principia: mathematical principles of natural philosophy*. University of California Press, 2003.

[70] Y. Chen, L. Zhang, B. Lin, Y. Xu, and X. Ren. Fighting detection based on optical flow context histogram. In *2011 Second International Conference on Innovations in Bio-inspired Computing and Applications*, pages 95–98, Dec 2011.

[71] S. P. Puthenpurayil, I. Chakrabarti, R. Virdi, and H. Kaushik. Very large scale integration architecture for block-matching motion estimation using adaptive rood pattern search algorithm. *IET Circuits, Devices Systems*, 10(4):309–316, Jul 2016.

[72] Matlab the language of technical computing, 1994. URL https://www.mathworks.com/. Accessed 04 April 2018.

[73] A. Barjatya. Block matching algorithms for motion estimation, 2011. URL https://www.mathworks.com/matlabcentral/fileexchange/8761-block-matching-algorithms-for-motion-estimation. Accessed 04 April 2018.

[74] R. C. Gonzalez and R. E. Woods. *Digital image processing (3rd Edition)*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 2006.

[75] Xiph.org video test media [derf's collection]. URL https://media.xiph.org/video/derf/. Accessed 04 April 2018.

[76] Video samples. URL http://www.divx.com/en/devices/profiles/video. Accessed 04 April 2018.

[77] Ultra video group. URL http://ultravideo.cs.tut.fi/#testsequences. Accessed 15 August 2018.

[78] Experimental data of duan's method and okade's method., 2012. URL http://www.facweb.iitkgp.ernet.in/~pkb/camera_classify.html. Accessed 04 April 2018.

[79] G. Abdollahian, C. M. Taskiran, Z. Pizlo, and E. J. Delp. Camera motion-based analysis of user generated video. *IEEE Transactions on Multimedia*, 12(1):28–41, Jan 2010.

[80] S. Narayanan and A. Makur. Camera motion estimation using circulant compressive sensing matrices. In *2013 9th International Conference on Information, Communications Signal Processing*, pages 1–5, 2013.

[81] François-Xavier Derue, Mohamed Dahmane, Marc Lalonde, and Samuel Foucher. Exploiting semantic segmentation for robust camera motion classification. In *Image Analysis and Recognition*, pages 173–181, 2017.

[82] S. Zhang and F. Stentiford. Motion detection using a model of visual attention. In *2007 IEEE International Conference on Image Processing*, volume 3, pages III–513–III–516, Sep 2007.

[83] V. F. Arguedas and J. M. Martnez Snchez. Towards methodological evaluation of affective video content annotation: First steps. In *2013 11th International Workshop on Content-Based Multimedia Indexing (CBMI)*, pages 237–242, Jun 2013.

[84] Convolution. URL http://mathworld.wolfram.com/Convolution.html. Accessed 23 March 2018.

[85] P. Prasertsakul, T. Kondo, and H. Iida. Video shot classification using 2d motion histogram. In *2017 14th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*, pages 202–205, Jun 2017.

[86] J. Kaiser and R. Schafer. On the use of the i0-sinh window for spectrum analysis. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28 (1):105–107, Feb 1980.

[87] F. J. Harris. On the use of windows for harmonic analysis with the discrete fourier transform. *Proceedings of the IEEE*, 66(1):51–83, Jan 1978.

[88] Find the local maxima by matlab, 2017. URL https://www.mathworks.com/help/signal/ref/findpeaks.html. Accessed 23 March 2018.

[89] Full match sports, 2014. URL http://fullmatchsports.com. Accessed 23 March 2018.

[90] The final round uefa champion league 2014/15 juventus vs barcelona, 2015. URL http://www.uefa.com/uefachampionsleague/history/seasons/#/iv/history/match/2015227. Accessed 23 March 2018.

[91] Thumbnail me 3.0. URL http://www.thumbnailme.com. Accessed 23 March 2018.

[92] J. R. Kender and B. L. Yeo. On the structure and analysis of home videos. In *Proceedings of Asian Conference on Computer Vision*, 2000.

[93] G. Abdollahian and E. J. Delp. Analysis of unstructured video based on camera motion. In *Proceedings of SPIE International Conference on Multimedia Content Access: Algorithms and Systems*, pages 1–12, 2007.

Publication lists

International Conference

- P. Prasertsakul, T. Kondo, T. Phatrapornnant, and T. Isshiki. A robust hand segmentation method based on color and background subtraction. In *5th International Conference on Information and Communication Technology for Embedded Systems (ICICTES 2014)*, Jan 2014.

- P. Prasertsakul and T. Kondo. A fingertip detection method based on the top-hat transform. In *2014 11th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*, pages 15, May 2014.

- J. Dulayatrakul, P. Prasertsakul, T. Kondo, and I. Nilkhamhang. Robust implementation of hand gesture recognition for remote human-machine interaction. In *2015 7th International Conference on Information Technology and Electrical Engineering (ICITEE)*, pages 247-252, Oct 2015.

- P. Prasertsakul, H. Iida, and T. Kondo. Boring game identification: case study using popular sports games. In *Annual Conference of the Society of Instrument and Control Engineers of Japan (SICE Annual Conference 2016)*, Sep 2016.

- P. Prasertsakul, T. Kondo, and H. Iida. Video shot classification using 2D motion histogram. In *2017 14th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*, pages 202-205, Jun 2017.

International Journal

- P. Prasertsakul and T. Kondo. A new fingertip detection method using the top-hat transform. *Thammasat International Journal of Science and Technology Asia*, 20(3):19-27, Jul-Sep 2015.
- P. Prasertsakul, J. Dulayatrakul, T. Kondo, and I. Nilkhamhang. A real-time hand segmentation method using background subtraction and color information. *Songklanakarin Journal of Science and Technology*, Dec 2017.