# Extracting Highlights from a Badminton Video Combine Transfer Learning with Players' Velocity

Shu Tao[1], Jiankun Luo[1], Jing Shang[1], and Meili Wang[1,2,3(✉)]

[1] College of Information Engineering,
Northwest A & F University, Xianyang, China
wml@nwsuaf.edu.cn
[2] Key Laboratory of Agricultural Internet of Things,
Ministry of Agriculture and Rural Affairs, Yangling 712100, Shaanxi, China
[3] Shaanxi Key Laboratory of Agricultural Information Perception
and Intelligent Service, Yangling 712100, China

**Abstract.** We present a novel method for extracting highlights from a badminton video. Firstly, we classify the different views of badminton videos for video segmentation through building classification model based on transfer learning, and achieve high-precision with real-time segmentation. Secondly, based on object detection by the object detecting model YOLOv3, we locate players in a video segment and calculate the players' average velocity to extract highlights from a badminton video. Video segments with higher players' average velocity reflect the intense scenes of a badminton game, so we can regard them as highlights in a way. We extract highlights by sorting badminton video segments with higher players' average velocity, which make users save their time to enjoy the highlights of an entire video. We laterally evaluate the proposed method through verifying whether a segment has admitted objective details such as exciting response from audiences and positive evaluation from narrators.

**Keywords:** Badminton video · Views classification · Video segmentation · Average velocity · Highlights extraction · Key audio emotional

## 1 Introduction

Badminton videos are usually time-consuming, so it is necessary to select highlights to save video viewers time. In addition, professional users, such as players or coaches, have a great demand for statistics reasons of those video highlights. If a badminton video can be quickly analyzed and segmented, then special moment detections can be carried out instead of manual scrutinization of the whole games, and it can assist coaches to make better programs on players' training plans and game strategies [1]. To avoid expensive data streaming or excessively


F. Tian et al. (Eds.): CASA 2020, CCIS 1300, pp. 82–91, 2020.
https://doi.org/10.1007/978-3-030-63426-1_9

spending our time on watching games, users' preference for sports video shift from watching entire videos to highlights of videos. For example, watching the top-ten exciting or crucial of a badminton game directly corresponds to the video highlight extraction and abnormal scenario detection in sports videos.

Facing with the increasing amount of badminton videos and demanding user needs, badminton videos need to be effectively organized, expressed, managed and retrieved so that users can quickly obtain the useful information accordingly. To solve the problem, this paper proposes a method of badminton video analysis based on machine learning. The major contributions of this work lie in two aspects: (1) Segment a badminton video based on the classification of observation view; (2) Automatically extract the highlights from the video segments. Our work on badminton videos also provides a prototype of video analysis and can be easily extended to other types of sports videos in the future work.

## 2 Related Work

Many scholars have conducted interesting studies on sports video analysis and most of them concentrates on keyframe extraction [2,3]. However, the key frames extracted from the video cannot show the motion direction and trajectory of objects, which belong to the static critical messages in the video. Conversely, the highlights belong to the dynamic critical messages in the video and consist of several crucial clips of an entire video, which can satisfy users' demand of watching the dynamic critical demands of the video without manual processing.

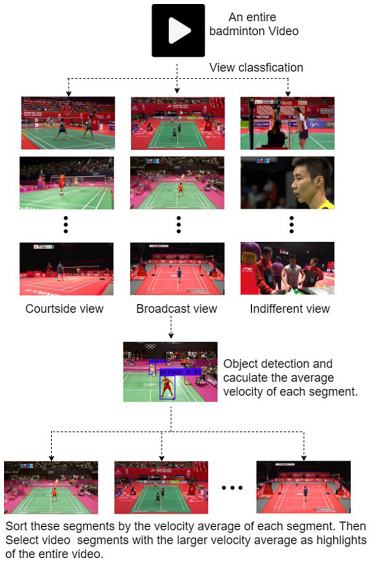



**Fig. 1.** The overall pipeline of our method.

This paper is different from those for keyframe extraction, and focuses on extracting highlights from an entire badminton video. Careelmont [4] proposed a method for the segmentation of badminton video and obtained the players' hitting strategies by extracting the moving objects from each frame in combination with the time dimension. In view of the highlights extraction, some methods extracted the highlights according to the user's different concerns and interest [5–8]. Sports video highlights can be interested on the basis of players' or referees' gestures and postures [9], and can be detected by the referee whistle sound [10], but these factors can not really reflect the popularity degree of a video. Fan et al. [11] employed real-time text stream, e.g.. opinion comments and posts, from social media to detect important events in live sports videos, but this method has some limitations because in many cases there are no comments or posts in sports videos. Yu et al. [12] proposed a method of perceiving audio emotional semantics and extracting video highlights extraction according to audio emotional semantics. However, the irrelevance of video scene leads to the low accuracy. In this paper, we extract highlights by calculating and sorting the overall velocity of both players in a badminton video, and through detecting key audio element [13] to evaluate the effectiveness of the proposed method.

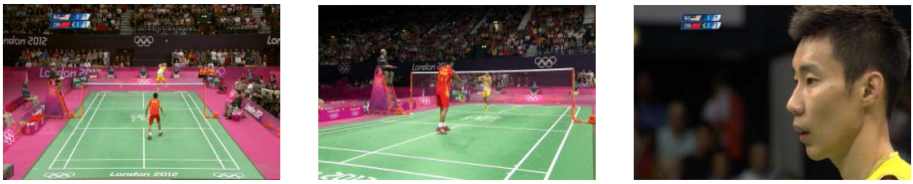
## 3 The Proposed Method

Our proposed method consists of two parts. Firstly, the classification feasibility of three views in badminton videos is verified by the method of clustering, which establishes the classification model of views, so as to achieve the segmentation of badminton video. Secondly, we extract highlights of a video through calculating and sorting the overall velocity of both players. The overall pipeline of our method is illustrated in Fig. 1. And the predicted highlights are evaluated by checking whether the key audio element exist on them.

### 3.1 Badminton Video Segmentation

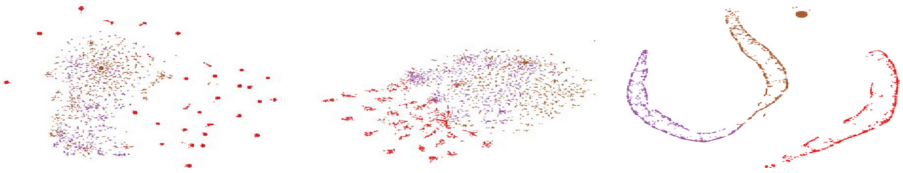
The views of soccer video include panoramic, medium, close-up and other views [13]. By observing and comparing, we can categorize the views of badminton videos as broadcast, courtside and indifferent view. See Fig. 2 for more details. The classification of views about broadcast and courtside are not affected by the background color or the specific view. But for the computer, it is difficult to understand the content of the image. Therefore, it is necessary to carry out pre-experiment to verify whether the classification of these views is reasonable. We verify the rationality of the classification based on the method of K-Means clustering [19]. Then we use the algorithm of t-SNE [20] to visualize the results of the status about pre-processing, processing by pre-training model and by the classification model MobileNet respectively, as shown in Fig. 3. And we can observe three views are better separated in Fig. 3(c).

Transfer learning can effectively solve the problem of insufficient data. In the multi-category tasks, the ImageNet dataset ISLVRC 2012 is often used, and it



(a) Broadcast view (b) Courtside view (c) Indifferent view

**Fig. 2.** View categories of badminton videos. Broadcast view and courtside view are the views of the camera overlooking the badminton court and looking at the front horizontal respectively, and indifferent view refers to others such as advertisements, game break, etc.



(a) Pre-processing (b) Processed by pre-training model (c) Processed by MobileNet

**Fig. 3.** The result comparing of the view classification. (a) The result before processing; (b) The result by the pre-training model; (c) The result by the classification model after pre-training. The red points, purple points and brown points represent the broadcast view, the courtside view, and the indifferent view respectively. (Color figure online)

includes a total of 1000 categories [21]. Before extracting the badminton highlights, the entire badminton video needs to be segmented into some useful segments relevant to the badminton game. The main view of useful video segments is broadcast view, which shows the action of badminton players in a game clearly and keeps the view for a long period. These video segments are the valid parts of the badminton videos, which laying the foundation for the highlight extraction. See Fig. 4 for the diagram of Segmenting a badminton video. The model of view classfication can effectively segment useful badminton video by transfer learning, which takes the model trained by ImageNet as the pre-training model. The main procedure of the badminton video segmentation is organized as follows. The following processing shall be done when each frame is input to the memory: If the view of the current frame is predicted to be broadcast and the previous frame is not broadcast, a storage queue is created for storing consecutive frames of the video segment which stores the current frame as the first frame. If both the current frame and its previous frame are predicted to be broadcast, the current frame is input to the latest created storage queue. Other cases are not dealt with. Finally, the storage queues amount is also the segments amount of an entire badminton video, and each storage queue corresponds to a storage space of each video segment.

**Table 1.** Training results of each network model.

| Pretraining network | Model size (Mb) | Accuracy of verification set | Loss of verification set | FPS |
|---|---|---|---|---|
| VGG16 [14] | 113 | 0.9836 | 0.0317 | 45.37 |
| ResNet50 [15] | 183 | 0.9770 | 0.0824 | 37.42 |
| InceptionV3 [16] | 168 | 0.9704 | 0.0729 | 36.58 |
| Xception [17] | 162 | 0.9770 | 0.0672 | 47.59 |
| MobileNet [18] | 26 | 0.9737 | 0.1071 | **52.31** |

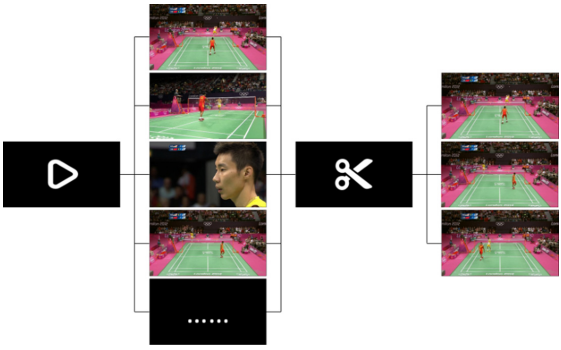


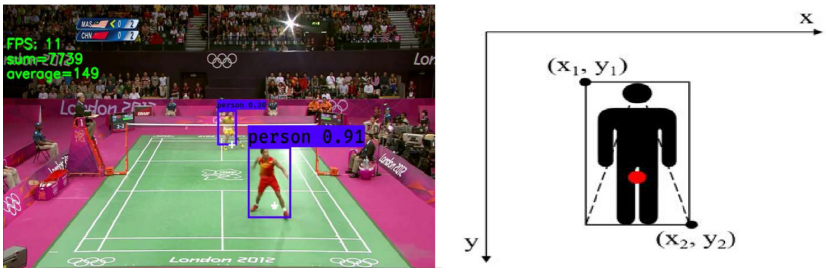
**Fig. 4.** Diagram of segmenting a badminton video.

### 3.2 Highlight Extraction

On the premise of maintaining the superiority of running speed, YOLOv3 [22] improve the detecting accuracy, especially strengthen the performance of recognize small objects. Thus we adopt YOLOv3 to detect players in a badminton video. In YOLOv3, the residual model Darknet-53 is used as the feature extractor, and realize multi-scale detection by FPN [23]. The detection result as shown in Fig. 5 (a). And Fig. 5 (b) shows the barycenter of players' object box, the top-left corner of a player object box is defined as $(x_1, y_1)$, and the bottom-right corner of the object box is defined as $(x_2, y_2)$, the barycenter of a player can be defined as Eq. 1. The target most likely to be a player may jump back and forth, thus it is necessary to consider all the players as a whole in each frame. The barycenter and velocity of the last player detected in two adjacent frames approximates the overall barycenter and overall velocity of all players. If we define the barycenter coordinates between $i$th frame and $(i+1)$th frame are $(p_i, q_i)$, $(p_{i+1}, q_{i+1})$ respectively, then the overall velocity $v_{i,i+1}$ of the players between these two adjacent frames can be defined as Eq. 2.

$$barycenter = \left( \frac{x_1 + x_2}{2}, \frac{y_1 + 2y_2}{3} \right) \tag{1}$$

$$v_{i,i+1} = \sqrt{(p_{i+1} - p_i)^2 + (q_{i+1} - q_i)^2} \tag{2}$$




(a) Players detection by YOLOv3 (b) Barycenter of players' object box

**Fig. 5.** Players detection and barycenter diagram.

$$V = \sum_{i=1}^{N-1} \frac{v_{i,i+1}}{T} \tag{3}$$

Highlight extraction allows users to observe the highlights of the entire video directly. The higher average velocity of the players in a badminton video segment could indicate the higher intensity of the game, thus highlights of each video can be regarded as the segments with the higher velocity of the players. Through calculating the average velocity $V$ of each segment of broadcast view, the average velocity of a video segment $V$ can be defined as Eq. 3, where $N$ and $T$ are the frame amount and the duration of a video segment respectively. Then these segments are sorted to obtain the highlights.

**Table 2.** Location of badminton highlights. For example, the data "2-1:1" means the first highlight of the video "Thailand 2019" is located in the second game when the score of both players is 1:1. And "*None*" indicates the segments amount of the corresponding entire video is less than 10.

| Thailand 2019 | Malaysia 2019 | Indonesia 2019 | Spain 2019 | German 2019 | England 2019 |
|---|---|---|---|---|---|
| **2−1:1** | 1−15:6 | 1−0:2 | 2−8:8 | 2−15:11 | 2−3:2 |
| 2−19:13 | 2−2:3 | 1−3:9 | 2−8:11 | 1−2:0 | 1−16:19 |
| 2−15:11 | 2−2:1 | 1−2:5 | 1−7:14 | 2−20:16 | 1−5:10 |
| 1−0:0 | 2−3:4 | 1−2:7 | 2−2:3 | 1−12:6 | 2−7:8 |
| 1−2:2 | 2−19:19 | 1−3:10 | 2−1:2 | 2−8:8 | 2−14:16 |
| 2−20:17 | 1−6:1 | 1−0:1 | 2−0: 2 | 2−13:11 | 1−14:18 |
| 2−12:9 | 2−2:2 | 1−0:3 | 1−1:8 | 2−7:7 | 2−12:16 |
| 2−13:11 | 2−11:11 | 1−2:9 | 1−0:0 | 2−13:9 | 1−5:11 |
| 1−3:3 | 1−14:6 | *None* | 1−1:6 | 1−15:9 | 2−16:17 |
| 2−17:13 | 2−7:9 | *None* | 1−8:14 | 1−3:0 | 2−12:15 |

# 4 Experiment

## 4.1 Badminton Video Segmentation

This paper adopts the pre-trained image classification model with ImageNet [21], which improves the classification results and accelerates the convergence speed of network training. The results obtained by the pre-trained models of each network are shown in Table 1. The test indexes include the model size, accuracy of the verification set, the loss of verification set and FPS. Since motion video analysis pursues real-time performance, a faster computation speed is needed, and the model with highest FPS is MobileNet, which implements real-time computing and uses deep convolution instead of traditional 3D CNN [24], thus reducing redundant representation of convolution kernels [18]. Therefore, we take MobileNet as the basic network structure for deep learning. The main parameters of the experimental model are as follows: (1) The size of the input image is 224 × 224; (2) The dimension of the fully connected layer is 3, which correspond to the three views; (3) The image pixel value is scaled within the interval [0, 1] to prevent the influence of affine transformation and improve the computational efficiency of neural network; (4) The optimization of the network training is Stochastic Gradient Descent (SGD). We set learning rate 0.0001 and momentum 0.9. (5) The model serve for multi-classification task, thus we use classification cross entropy as the loss function, can be derived as Eq. 4, where $C$ is number of view classes, $p_i$ the groundtruth label concerning view class $i$, and $q_i$ the predict result of class $i$; (6) The batch size of the training data is set to 32. And the training strategies are as follows: (1) If the loss of the verification set does not decrease after 5 iterations, the learning rate is reduced by half. (2) If the verification set loss does not decrease after 30 training iterations, stop the training. The method described above yielded us an averaged accuracy of 95%, which can be considered adequate for the further tasks.

$$L = -\sum_{i=1}^{C} p_i \log(q_i) \tag{4}$$

We test six entire badminton videos[1] from YouTube to validate the effectiveness of the segments of them. In order to verify the generalization ability of the classfication model, the six videos in Table 2 are completely different from the training set[1].

## 4.2 Evaluation of Extracted Highlights

As Phomsoupha and Laffaye [25] pointed out, a scored shot usually takes from 7 s to 15 s. In this paper, it is believed that the segment whose average velocity is in the top $\varepsilon$ can be regarded as a highlight, where $\varepsilon$ is the threshold of average velocity rank, such as 10. Next, evaluating whether the extracted highlights are representative of the entire video.

[1] https://github.com/taoshu1996/BUILDINGTAO.

Li et al. [13] extracted the highlights of football games by detecting the key audio corresponded video clip from the football game – the excited voice of the commentators or the whistle of the referees. To make statistical results as objective as possible, we adopt this method to evaluate whether the extracted highlights is satisfying by judging whether they have response from the commentary or the audience, such as boo, shouts and warm applause in the extracted highlights. The time threshold set to 15 s in the experiments, and 10 video segments with the largest velocities average are used for verifications. The criterion of verification is the existence of some specific sounds, such as boo, shouts and warm applause from the audience or the response from the commentators during the stroke. The highlights of tested badminton videos are located by recording the status of the original game video and the score status of all the players. The scores of the players corresponds to the highlight are shown in Table 2. The statistics about the reactions from audience and the comments from commentators in highlights are shown in Table 3, which is carefully done by several badminton players. There are 58 highlights in 6 badminton games were extracted. Those extracted highlights which are marked with symbol "O" or "$\Delta$" in Table 3 are considered as real highlights. In our experiments, 54 highlights[1] are evaluated to be true highlights, accounting for 93.10%, demonstrate the method in this paper is effective.

**Table 3.** Evaluation of badminton video highlights. The symbol "O" indicates that the highlight has a positive response from the audience, and "$\Delta$" indicates that the highlight has a positive evaluation from the commentators.

| Thailand 2019 | Malaysia 2019 | Indonesia 2019 | Spain 2019 | German 2019 | England 2019 |
|---|---|---|---|---|---|
| O$\Delta$ | O | O$\Delta$ | $\Delta$ | O$\Delta$ | O$\Delta$ |
| O$\Delta$ | O$\Delta$ | O$\Delta$ | $\Delta$ | | O$\Delta$ |
| O$\Delta$ | O$\Delta$ | O | O | O$\Delta$ | O |
| O | O$\Delta$ | O$\Delta$ | O | O | $\Delta$ |
| O$\Delta$ | O$\Delta$ | O$\Delta$ | O$\Delta$ | | O$\Delta$ |
| O$\Delta$ | O$\Delta$ | O$\Delta$ | $\Delta$ | $\Delta$ | O$\Delta$ |
| O | O$\Delta$ | O | $\Delta$ | O | O$\Delta$ |
| $\Delta$ | O$\Delta$ | O$\Delta$ | | O$\Delta$ | O |
| O$\Delta$ | O$\Delta$ | *None* | | O$\Delta$ | O$\Delta$ |
| O$\Delta$ | O$\Delta$ | *None* | O$\Delta$ | O$\Delta$ | $\Delta$ |

## 5 Conclusions

This paper establishes a scene classification model through transfer learning in order to segment a badminton video for highlight extraction. We locate badminton players of each segment based on the model YOLOv3, and calculate the velocities average of all the players. We select segments of the high velocities as highlights.

Our work provides a prototype of extending the current model to other types of sports video highlight extraction. And we intend to build a statistical model through learning existing extracts by machine in the future work.

**Acknowledgement.** This work is partially funded by Key Laboratory of Agricultural Internet of Things, Ministry of Agriculture and Rural Affairs, China (2018AIOT-09), Key Research and Development Program of Shaanxi Province (2018NY-127), and supported by the Shaanxi Key Industrial Innovation Chain Project in Agricultural Domain (Grant No. 2019ZDLNY02-05).

## References

1. Li, S., Yang, X.: The overview of video summary technology. Technol. Innov. Appl. (2018)
2. Xia, G., Sun, H., Niu, X., Zhang, G., Feng, L.: Keyframe extraction for human motion capture data based on joint kernel sparse representation. IEEE Trans. Ind. Electron. **64**(2), 1589–1599 (2016)
3. Roberts, R., Lewis, J.P., Anjyo, K., Seo, J., Seol, Y.: Optimal and interactive keyframe selection for motion capture. Comput. Vis. Media (2019)
4. Careelmont, S.: Badminton shot classification in compressed video with baseline angled camera. Master's thesis [Academic thesis] (2013)
5. Bu, Q., Hu, A.: An approach to user-oriented highlights extraction from a sport video, vol. 21. College of Information Science and Engineering (2008)
6. Huang, Q., Zheng, Y., Jiang, S., Gao, W.: User attention analysis based video summarization and highlight ranking. Chin. J. Comput. **31**, 1612–1621 (2008)
7. Chakraborty, P.R., Tjondronegoro, D., Zhang, L., Chandran, V.: Automatic identification of sports video highlights using viewer interest features. In: Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval (2016)
8. Wang, H., Huangyue, Yu., Hua, R., Zou, L.: Video highlight extraction based on the interests of users. J. Image Graph. **23**(5), 0748–0755 (2018)
9. Choroś, K.: Highlights extraction in sports videos based on automatic posture and gesture recognition. In: Nguyen, N.T., Tojo, S., Nguyen, L.M., Trawiński, B. (eds.) ACIIDS 2017. LNCS (LNAI), vol. 10191, pp. 619–628. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-54472-4_58
10. Kathirvel, P., Manikandan, M.S., Soman, K.P.: Automated referee whistle sound detection for extraction of highlights from sports video. Int. J. Comput. Appl. **12**(11), 16–21 (2011)
11. Fan, Y.-C., Chen, H., Chen, W.-A.: A framework for extracting sports video highlights using social media. In: Ho, Y.-S., Sang, J., Ro, Y.M., Kim, J., Wu, F. (eds.) PCM 2015. LNCS, vol. 9315, pp. 670–677. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24078-7_69
12. Yu, C., Weng, Z.: Audio emotion perception and video highlight extraction, vol. 27. College of Mathematics and Computer Science (2015)
13. Li, J., Wang, T., Hu, W., Sun, M., Zhang, Y.: Soccer highlight detection using two-dependence Bayesian network. In: IEEE International Conference on Multimedia & Expo (2006)
14. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition, vol. 09 (2014)

15. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778, June 2016
16. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: CVPR, pp. 2818–2826, June 2016
17. Chollet, F.: Xception: deep learning with depthwise separable convolutions. In: Computer Vision and Pattern Recognition (CVPR), pp. 1800–1807, July 2017
18. Howard, A.: MobileNets: efficient convolutional neural networks for mobile vision applications. In: Computer Vision and Pattern Recognition (CVPR), April 2017
19. Coates, A., Ng, A.Y.: Learning feature representations with k-means. In: Montavon, G., Orr, G.B., Müller, K.-R. (eds.) Neural Networks: Tricks of the Trade. LNCS, vol. 7700, pp. 561–580. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-35289-8_30
20. Laurens, V.D.M., Hinton, G.: Visualizing data using t-SNE. J. Mach. Learn. Res. **9**(2605), 2579–2605 (2008)
21. Russakovsky, O., et al.: ImageNet large scale visual recognition challenge. Int. J. Comput. Vis. **115**(3), 211–252 (2015). https://doi.org/10.1007/s11263-015-0816-y
22. Redmon, J., Farhadi, A.: YOLOv3: an incremental improvement. CoRR, abs/1804.02767 (2018)
23. Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B.: Feature pyramid networks for object detection (2016)
24. Tran, D., Bourdev, L.D., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3D convolutional networks. CoRR, abs/1412.0767 (2014)
25. Phomsoupha, M., Laffaye, G.: The science of badminton: game characteristics, anthropometry, physiology, visual fitness and biomechanics. Sports Med. **45**(4), 473–495 (2015). https://doi.org/10.1007/s40279-014-0287-2