What the Appearance Channel from Two-Stream Architectures for Activity Recognition Is Learning?

Reinier Oves García(✉) and L. Enrique Sucar

Computer Science Department, Instituto Nacional de Astrofísica Óptica y Electrónica, Sta. María Tonantzintla, 72840 Puebla, CP, Mexico
{ovesreinier,esucar}@inaoep.mx

Abstract. The automatic recognition of human activities from video data is being led by spatio-temporal Convolutional Neural Networks (3D CNNs), in particular two-stream architectures such as I3D that reports the best accuracy so far. Despite the high performance in accuracy of this kind of architectures, very little is known about what they are really learning from data, resulting therefore in a lack of robustness and explainability. In this work we select the appearance channel from the I3D architecture and create a set of experiments aimed at explaining what this model is learning. Throughout the proposed experiments we provide evidence that this particular model is learning the texture of the largest area (which can be the activity or the background, depending on the distance from the camera to the action performed). In addition, we state several considerations to take into account when selecting the training data to achieve a better generalization of the model for human activity recognition.

1 Introduction

Automatic decision making based on machine learning models has become in a tendency today, with applications in real-world scenarios such as medicine, robotics, and video surveillance among others. Sometimes, models based on deep learning cannot be well explained [13], resulting in that way in a lack of reliability, and therefore, they are rejected, especially in such areas where a mistake has severe implications (e.g. medical diagnosis and terrorism detection) [14]. Apart from ethical reasons, knowing what these models are learning can be crucial for areas where it is essential to prevent that recognition systems being deceived (e.g. security environments and autonomous navigation) [11].

Nowadays two-stream CNNs with three-dimensional convolutions [3] constitute the strongest tool to tackle video classification problems. These particular architectures can be seen as a combination of two independent models that learn independent features from two different representations: appearance, and motion. The appearance model learns from pure RGB frames while the motion model learns from the optical flow [19].

K. M. Figueroa Mora et al. (Eds.): MCPR 2020, LNCS 12088, pp. 251–260, 2020.
https://doi.org/10.1007/978-3-030-49076-8_24

Leading the state-of-the-art we can situate the I3D architecture [1] which is a two-stream CNN inflated from a 2D model pre-trained over the ImageNet dataset with spatio-temporal convolutions and pooling layers. Despite the high performance, this particular architecture constitutes a black box and little is know about what it is learning from the input data. Towards a better understanding and explainability of 3D CNNs for action recognition, we selected the appearance channel from the I3D architecture and the UCF101 dataset [17] in the conducted experiments. The motion channel will be omitted from the experiments, since the images produced by the optical flow, considerably lack texture.

This work, far from being a criticism, is presented to clarify some points that remain unclear when using two-stream architectures, especially the appearance model. Besides, it is intended to establish new guidelines for making a suitable selection of datasets that guarantee that the appearance model learns the appearance of the body language and nor the background steadiness. Additionally, we investigate the sensitivity of the appearance channel from the I3D architecture to various factors presented in the training data that could dramatically affect a posterior evaluation of the model: the texture of the background and the area covered by the action. To determine how sensitive the 3D CNNs can be to all these factors, we guide the conducted experiments towards answering the following questions:

- How good the performance of the appearance channel can be if the activity is always associated with the context (e.g. brushing your teeth is always done in a bathroom or boxing is always performed in a ring)?
- What happens if the context where the activity is performed changes abruptly?
- What is the impact of training only considering the region that occupies the activity (background-independent activities)?
- How much the camera target distance affects the generalization of the model?

Finally, a set of evidence that leads that 3D CNNs are learning the largest region in the video (the texture of the background or the texture of the activity) is presented. Furthermore, we provide a set of considerations that are useful for the suitable selection of training data to build models hard to beat.

The rest of the paper is organized as follows: Sect. 2 relates the main advances in the computer vision area for action recognition as well as the main drawbacks discovered so far. The experimental configurations and results are given in Sect. 3. In Sect. 4 we relate some guidelines and considerations useful for selecting a suitable dataset that guarantees the generalization of the models. Finally, in Sect. 5 the conclusions and future work are given.

2 Related Work

Human Action Recognition has presented a significant improvement in accuracy during the last years by applying CNNs [11] and transfer learning. The clever transformation of 2D models pre-trained over ImageNet [1,8] into models with

spatio-temporal convolutions has resulted in the more powerful tool to tackle this challenging task.

One of the most revolutionary approaches for action recognition are two-stream CNNs with 2D convolutional kernels. In [16] Simonyan et al. showed that the combination of the appearance and motion has the ability to improve action recognition accuracy. The new era of two-stream CNNs is marked by 3D convolutional kernels, which have outperformed 2D CNNs through the use of large-scale video datasets [1]. Currently, a two-stream model, with spatio-temporal convolutions over appearance and motion, joined at the fully connected layer, shows the most promise [1,8]. There is evidence that combining two-stream CNNs with a third representation can improve the accuracy results [2].

Two-stream architectures, despite being powerful tools for video classification, are considered black boxes due to the little information that they provide about the learning process, aggravated by the amount of parameters to be trained [22]. At the same time that CNNs evolve, several authors have questioned the strength of the models while others try to explain what is happening inside those black boxes. The work presented in [9] introduces an approach that deciphers which portions of the face are more concerned with model predictions. [12] demonstrates that adversaries can easily craft adversarial examples without any internal knowledge of the target network. In [14], the authors found that the question "Husky or Wolf?" does not depend on the animal but the snow in the background. Several works have focused in models unboxing via decision trees [21] and gradient-based visualization techniques [20]. In [4,5] are presented several clues about what are learning patio-temporal CNNs by visualizing the internal excitation of the models but they do not provide any lead about where the excitation occurs: within the background or over the actors.

Until recently, it was believed that 2D CNNs were able to recognize objects through the learning of non-handcrafted features [10]. In [7] is shown that CNNs trained over ImageNet are strongly biased towards recognizing textures rather than shapes. This is the main reason why these models are very susceptible to adversarial attacks [6]. The fact that these models are biased towards texture also makes difficult the training process over small datasets and far from favoring the learning, it causes significant harm to the performance of the model [15].

Given the success of deep learning architectures, there is recent interest to have a deeper understanding of how and what they learn. Until now, some works have been directed to explain or criticize deep architectures used in image classification, but it does not happen in the same way for video processing (3D CNNs).

3 Experiments and Results

To investigate what are learning 3D CNNs we focused on the appearance channel from the I3D architecture. For that, we implement a training procedure for the model [1] published on github[1]. Given that all the models used in our experiments

[1] https://github.com/deepmind/kinetics-i3d.

were trained using our code we conducted the first experiment (Sect. 3.1) to validate the quality of our training procedure. The second experiment (Sect. 3.2) is mainly focused on demonstrating that the appearance channel is learning the largest part of the scene: the area covered by the action or the background, depending on the camera target distance. In the third experiment (Sect. 3.3) are presented several considerations that inform us the appearance channel is more prone to learn the background than to learn the activity and we show that the model is considerably affected by the suppression of the background.

To visualize the outcomes we present histograms for each experiment, in which each bar relates to one class of the 101 classes in the UCF101 dataset. For data representation, we adopted the following convention over the bar chart visualization technique: cyan (light) bars represent how many videos were tested for each class while magenta (dark) bars represent the number of missclassified videos. The 101 classes in the dataset were sorted alphabetically in order to maintain the correspondence between the visualization of different experiments.

3.1 Towards the Reproduction of the Results Presented in [1]

In Table 1.a are presented the results achieved by our implementation after combining both channels, appearance and motion. The variation presented in the reported results is given by the introduction of random parameters in the model, data augmentation and random crop during training. Despite this, our implementation achieves similar results than those reported by the authors and hence can be used as a baseline for further experimentations.

Since we are focused on exposing what part of the scene the appearance channel is learning, we report the results of the RGB model for each partition of the dataset in Table 1.b. In Fig. 1 is shown that the model has, in general, a decent performance over the appearance channel. Classes with higher accuracy are those in which the activity is performed in a specific scenario and the objects involved in the activity are presented in both phases train and test (e.g. $BoxingPunchingBag_{18}$, $PlayingGuitar_{63}$, etc.). On the other hand, classes with

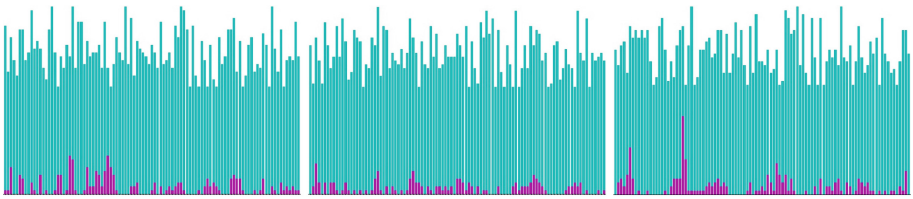
Table 1. In these tables, we report the accuracy achieved by our implementation over the three partitions of the UCF101 dataset; compared with the original publication of Carreira et al. [1]. **a)** presents the accuracy obtained by combining the appearance and motion channels and **b)** presents the accuracy of the appearance and motion channels.

UCF101	Carreira et al. [1]	Ours
Split-1	-	97.60
Split-2	-	97.99
Split-3	-	97.72
Average	98.0	97.77

a)

UCF101	RGB channel	Motion channel
Split-1	94.55	96.40
Split-2	95.66	96.97
Split-3	94.34	96.42
Average	94.85	96.59

b)



a) b) c)

Fig. 1. These pictures depict the performance of the appearance channel from the I3D. **a)**, **b)** and **c)** reflect the results achieved by the model for each partitions on the UCF101 dataset, respectively. The cyan color represents how many instances were evaluated per class and the magenta color denotes the amount of misclassified instances. Horizontal axis refers to the index of each class ordered alphabetically while the vertical axis represents the amount of instances per class (Best seen in color). (Color figure online)

a worse performance are those in which the activities are performed in different scenarios (e.g $CricketBowling_{23}$, $Hammering_{36}$, $CricketShot_{24}$, etc.). The subindex refers to the index of the corresponding bar in the visualization.

Until now, we only have clues that point to the fact that the appearance channel is learning from the background texture. For this reason, we put this hypothesis to a quantitative test where the quality of the appearance channel is evaluated.

3.2 Background Inclusion and Exclusion During the Evaluation Phase

This experiment is focused on demonstrating that the appearance channel is learning the texture of the largest region of the image, say the activity (very close to the camera) or the background (when the activity is performed away from the camera). To accomplish the aim of this experiment we employ a mask

Table 2. These tables report the accuracy achieved by the model trained eliminating the background or the activity. **a)** presents the accuracy of the model when the background are replaced by a black mask and **b)** presents the opposite case, when the activities are replaced by a black mask.

UCF101	RGB channel
Split-1	19.16
Split-2	18.31
Split-3	23.53
Average	20.33

a)

UCF101	RGB channel
Split-1	58.92
Split-2	54.28
Split-3	56.54
Average	56.58

b)

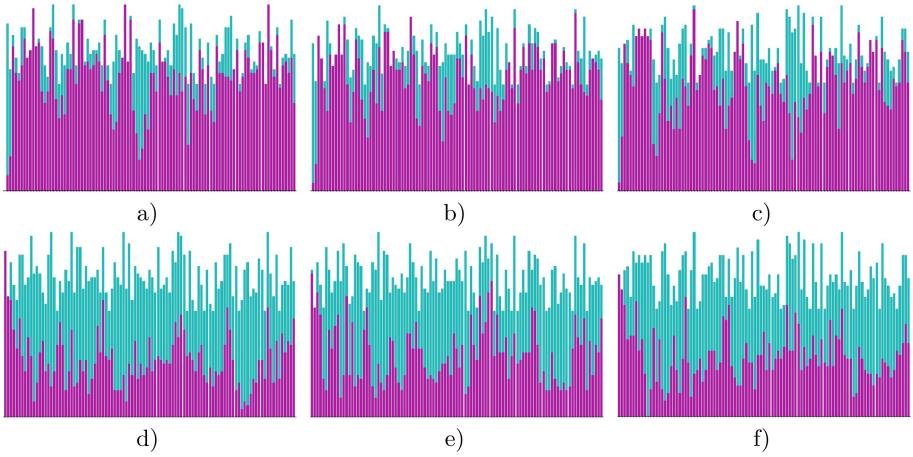


Fig. 2. These images show the performance of the appearance channel from I3D. Figures **a)**, **b)** and **c)** visualize the results achieved by the model when the background is replaced by a black mask (only the activity is taking into consideration). Figures **d)**, **e)** and **f)** visualize the results achieved by the model when the activity is replaced by a black mask (only the background). The cyan color represents how many instances were evaluated per class and the magenta color denotes the amount of misclassified instances. Notice how the images have a complementary behavior (when the results in the first are good in the second are bad). Horizontal axis refers to the index of each class ordered alphabetically while the vertical axis represents the amount of instances per class (Best seen in color). (Color figure online)

r-CNN[2] [18] for human body segmentation. Since the quality of the extracted human masks is not always the best, a morphological operation of dilation is used to reduce this problem and ensure that people involved in the activity are always included in the analysis. However, we manually check the 13320 videos within the dataset for corroborating the quality of such masks. For those cases where the person performing the activity was undetected by the mask, we used the original RGB video.

To assess how sensitive the model can be concerning the area occupied by the activity, the training generated in Sect. 3.1 is selected and the test is performed over the three partitions of the dataset. The first test is carried out taking into account only the area occupied by the activity and the background is replaced by a black mask. Quantitative results are presented in Table 2.a. For the second test, the activity is replaced by the black mask and the results are presented in Table 2.b.

Contrary to the results presented in Sect. 3.1, the results achieved when the background is removed are catastrophic (See Table 2.a). The classes that report better accuracy are those that are performed near to the camera with static background (e.g $ApplyEyeMakeup_1$, $ApplyLipstick_2$, $JumpRope_{47}$,

[2] https://github.com/facebookresearch/detectron2.

$JumpingJack_{48}$ etc.). Note how the performance decreases while the area covered by the activity decreases or the background changes for instances of the same class. This means that a reduction in texture strongly affects the final decision of the appearance channel (e.g $Basketball_8$, $Drumming_{27}$, $SkyDiving_{83}$, etc.). See Fig. 2.

After analyzing Table 2.b , we can arrive at the conclusion that the obtained results are not as good as those in Table 1.b but it do show a better performance than those achieved when the background is removed (See Table 2.a). This experiment provides us evidence about how important is the background for the model. For this particular dataset, the background results more important than the activity itself as the results are almost 3 times better than those reported in Table 2.a (See Fig. 2).

From this experiment, we can conclude the premise that the appearance channel is favoring the largest regions in the videos. When this area corresponds to the execution of the activity, the model ends up learning what is expected (spatio-temporal features of the activity). Unlike this, in cases where the background predominates, the model ends up learning the invariability of the background texture. The reason why the activities carried out near the camera now provide the worst results is because the background lacked texture and that is why the model is learning the activity (See Fig. 3).

3.3 Model Performance Evaluation When the Activity and Background Are Isolated

This experiment tries to answer two fundamental questions of the learning process: (i) how much does the overall performance of the model is affected if the background is the same for all classes during training and testing? and (ii) which is more discriminating for the model the appearance of the activity or the background? To answer the first question, we replace the background of every video in the dataset with a black mask in both phases, training and testing. The results achieved in this experiment are presented in Table 3.a. To answer the second question, the opposite strategy is followed. This time, the action is replaced by a black mask during both phases, training, and testing. Additionally, videos were resized to 224×224 without preserving the aspect ratio to guarantee that action

Table 3. These tables report the accuracy achieved by the model when the activity and background are isolated. **a)** presents the accuracy of the model when background is replaced by a black mask and **b)** presents the opposite case, when the activities are replaced by a black mask. The mask is applied in both phases, training and testing.

UCF101	RGB channel
Split-1	78.27
Split-2	78.68
Split-3	77.13
Average	78.02

a)

UCF101	RGB channel
Split-1	90.69
Split-2	93.10
Split-3	92.90
Average	92.23

b)

performers were fully included in the video. Results of this experiment can be seen in Table 3.b.

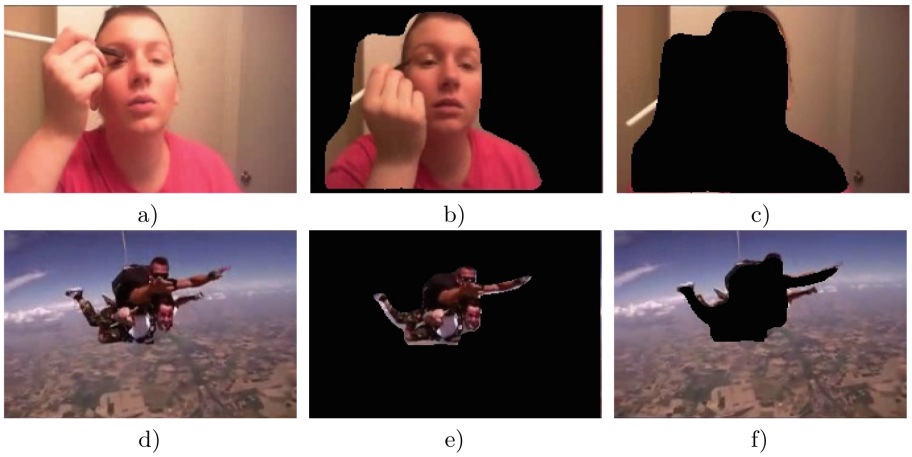


Fig. 3. The first row (subimages a), b) and c)) displays an instance where the model learned the activity and not the background ($ApplyEyeMakeup_1$). In the second row (subimages d), e) and f)), an instance in which the model learned the background instead of the activity is presented ($SkyDiving_{83}$).

Note how the performance of the model is improved when the background is included during the training process. This is due to models pre-trained on ImageNet are biased to learn texture instead of shapes [7]. When the background is removed from training videos, the texture is diminished drastically and hence the performance of the model also decreases. This arises by the fact the I3D model was initialized with weights learned from ImageNet and a lack of texture during the fine tuning process significantly harm the global performance of the model. Finally, it can be concluded that the appearance channel from the I3D model learns in most of the classes the background. This is because most of the activities in this dataset are tied to the context and the largest area in the scenes correspond to the background texture.

4 Discussion and General Considerations

In Fig. 3 are depicted, in three different ways, two instances from two different classes: (i) a) and d) are the original RGB frames, (ii) b) and e) show the human mask over a black background and finally, (iii) c) and f) show the background without the activity. For images within the first row, the background lacks texture and for this reason, the model learns the texture of the activity. For cases, where the background is very similar for different activities, training data produces the desired effect over the model; the appearance of the activity is

learned. The second row of Fig. 3 shows the opposite case. Here, the background is highly textured and is very similar within the same class (videos are recorder from aerial views). For this specific class, the model learns the background.

Another important factor to consider during training is the random cropping of the videos. There is nothing that guarantees the people involved in the activity remain within the selected region (partially or totally) each time the video is included in a batch. Note how, in Table 3.b the performance is similar to produced by the original model in Table 1.b. Finally, we propose two foremost considerations that have to be taken into account when selecting datasets and training the appearance channel from the I3D architecture for human activity recognition. The first one establishes that the strength of the model will be given by the diversity of backgrounds in each activity class. The second aspect to take into consideration is that using data augmentation far from improving the performance of the model creates an irreparable situation since it introduces more of the same texture and this biases the model.

5 Conclusions

In this paper, we provided evidence that clarifies some points that remain unclear when using two-stream CNNs, especially the appearance channel. From the experimentation, we can conclude that the appearance channel from the I3D architecture is learning the texture of the largest region of the video. In other words, if the activity is closer to the camera, the model learns the human texture and becomes in a stronger classifier invariant to scale and background changes. In the opposite case, when the area covered by the background is larger than the area of the performed activity, the model learns the background texture, resulting in a weak classifier, easy to deceive and susceptible to changes in the scale as well as in the background. Besides, this particular model does not work well when different activities are performed over the same background. For these cases, we should pay special attention when using this model in real scenarios where the background can change such as surveillance, autonomous navigation, etc. So we can conclude with the following statement: similar to 2D CNNs, it seems that 3D CNNs are learning the texture of the background and its steadiness. A possible solution is to mask the videos in the training phase keeping only the humans performing the activities; and base the recognition more in the appearance of the activity than in the appearance of the scene.

References

1. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: CVPR, pp. 6299–6308 (2017)
2. Choutas, V., Weinzaepfel, P., Revaud, J., Schmid, C.: Potion: pose motion representation for action recognition. In: CVPR, pp. 7024–7033 (2018)
3. Diba, A., Pazandeh, A.M., Van Gool, L.: Efficient two-stream motion and appearance 3D CNNs for video classification. arXiv preprint arXiv:1608.08851 (2016)

4. Feichtenhofer, C., Pinz, A., Wildes, R.P., Zisserman, A.: What have we learned from deep representations for action recognition? In: CVPR, pp. 7844–7853 (2018)
5. Feichtenhofer, C., Pinz, A., Wildes, R.P., Zisserman, A.: Deep insights into convolutional networks for video recognition. In: IJCV, pp. 1–18 (2019)
6. Ford, N., Gilmer, J., Carlini, N., Cubuk, D.: Adversarial examples are a natural consequence of test error in noise. arXiv preprint arXiv:1901.10513 (2019)
7. Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F.A., Brendel, W.: Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. arXiv preprint arXiv:1811.12231 (2018)
8. Hara, K., Kataoka, H., Satoh, Y.: Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and ImageNet? In: CVPR, pp. 6546–6555 (2018)
9. Khorrami, P., Paine, T., Huang, T.: Do deep neural networks learn facial action units when doing expression recognition? In: CVPR, pp. 19–27 (2015)
10. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)
11. Kumar, A.D., Chebrolu, K.N.R., Soman, K.P., et al.: A brief survey on autonomous vehicle possible attacks, exploits and vulnerabilities. arXiv preprint arXiv:1810.04144 (2018)
12. Narodytska, N., Kasiviswanathan, S.: Simple black-box adversarial attacks on deep neural networks. In: CVPRW, pp. 1310–1318. IEEE (2017)
13. Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Berkay Celik, Z., Swami A.: Practical black-box attacks against machine learning. In: Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security, pp. 506–519. ACM (2017)
14. Ribeiro, M.T., Singh, S., Guestrin, C.: Why should i trust you?: explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD, pp. 1135–1144. ACM (2016)
15. Ringer, S., Williams, W., Ash, T., Francis, R., MacLeod, D.: Texture bias of CNNs limits few-shot classification performance. arXiv preprint arXiv:1910.08519 (2019)
16. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. In: Advances in Neural Information Processing Systems, pp. 568–576 (2014)
17. Soomro, K., Zamir, A.R., Shah, M.: Ucf101: a dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402 (2012)
18. Wu, Y., Kirillov, A., Massa, F., Lo, W.-Y., Girshick, R.: Detectron2. https://github.com/facebookresearch/detectron2 (2019)
19. Zach, C., Pock, T., Bischof, H.: A duality based approach for realtime TV-L^1 optical flow. In: Hamprecht, F.A., Schnörr, C., Jähne, B. (eds.) DAGM 2007. LNCS, vol. 4713, pp. 214–223. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-74936-3_22
20. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8689, pp. 818–833. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10590-1_53
21. Zhang, Q., Yang, Y., Ma, H., Wu, Y.N.: Interpreting CNNs via decision trees. In: CVPR, pp. 6261–6270 (2019)
22. Zhou, H., Alvarez, J.M., Porikli, F.: Less is more: towards compact CNNs. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9908, pp. 662–677. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46493-0_40