applied sciences

MDPI

Article

Comparative Study of Movie Shot Classification Based on Semantic Segmentation

Hui-Yong Bak [1] and Seung-Bo Park [2,*]

[1] Department of Mechatronics Engineering, Inha University, 100 Inha-ro, Michuhol-gu, Incheon 22212, Korea; dkdlenrh@naver.com

[2] Department of Software Convergence Engineering, Inha University, 100 Inha-ro, Michuhol-gu, Incheon 22212, Korea

* Correspondence: molaal@inha.ac.kr; Tel.: +82-32-860-8831

Received: 22 April 2020; Accepted: 12 May 2020; Published: 14 May 2020

Abstract: The shot-type decision is a very important pre-task in movie analysis due to the vast information, such as the emotion, psychology of the characters, and space information, from the shot type chosen. In order to analyze a variety of movies, a technique that automatically classifies shot types is required. Previous shot type classification studies have classified shot types by the proportion of the face on-screen or using a convolutional neural network (CNN). Studies that have classified shot types by the proportion of the face on-screen have not classified the shot if a person is not on the screen. A CNN classifies shot types even in the absence of a person on the screen, but there are certain shots that cannot be classified because instead of semantically analyzing the image, the method classifies them only by the characteristics and patterns of the image. Therefore, additional information is needed to access the image semantically, which can be done through semantic segmentation. Consequently, in the present study, the performance of shot type classification was improved by preprocessing the semantic segmentation of the frame extracted from the movie. Semantic segmentation approaches the images semantically and distinguishes the boundary relationships among objects. The representative technologies of semantic segmentation include Mask R-CNN and Yolact. A study was conducted to compare and evaluate performance using these as pretreatments for shot type classification. As a result, the average accuracy of shot type classification using a frame preprocessed with semantic segmentation increased by 1.9%, from 93% to 94.9%, when compared with shot type classification using the frame without such preprocessing. In particular, when using ResNet-50 and Yolact, the classification of shot type showed a 3% performance improvement (to 96% accuracy from 93%).

Keywords: shot type classification; semantic segmentation; CNN; Mask R-CNN; Yolact

1. Introduction

In films, movie shot types are classified based on the distance between the camera and the subject, and the general types of shots are the close-up shot, the medium shot, and the long shot [1,2]. Among them, close-up shots are used for expressing the emotions and psychology of the characters, with the subject occupying most of the screen. As shown in Figure 1a, emotion or psychology is expressed with the character's eyes, mouth, and facial muscles by making the character's face occupy most of the screen [3]. In medium shots, a portion of the character's body below the waist or elbow is located at the bottom of the screen. Medium shots are also used to express a character's gaze direction and movement, since the character's body above the waist appears on the screen [3]. Figure 1b is an example of the medium shot, which shows the character's gaze direction, motion, and conversation partner. In long shots, the subject occupies about one-sixth of the screen, giving the audience information about the place (inside or outside, in an apartment, a shop, a forest, etc.) and time (day, night, season) [3].

Additionally, the director uses the close-up, medium shot, and the long shot alternately to appropriately place the flow of emotion in the film and the importance of the scene [4]. As such, the shots contain a lot of information, such as the arrangement and flow of emotion, the psychology of the characters, and the important scenes, so classifying the shot type is a very important pre-task in movie analysis.

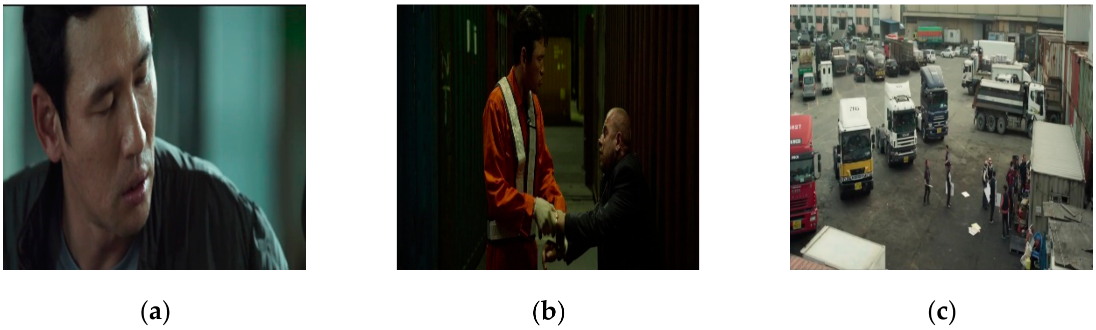


Figure 1. Types of shots in the movie *Veteran*: (**a**) Close-up shot; (**b**) Medium shot; (**c**) Long shot.

Two representative studies of shot type classification include classifying shot types based on the proportion of the face occupying the screen and classifying shot types using a convolutional neural network (CNN). In one study that classified shot types from the proportion of the face on-screen, the accuracy of shot type classification was high for close-ups and medium shots (with faces in the frame), but the accuracy in shot type classification for long shots (without faces) was low. Also, shot type classification cannot be performed if there is no face on-screen. Shot type classification using the CNN is able to classify the shot type even if no one is on-screen, but a certain portion of the shots cannot be classified, owing to the absence of a semantic approach that people can grasp (that is, the image cannot be semantically analyzed).

Therefore, in the present study, the performance of shot type classification was improved by pre-processing semantic segmentation of the frames extracted from the movie. Semantic segmentation approaches the images in order to distinguish boundary relationships among objects. The representative technologies of semantic segmentation include the Mask R-CNN and Yolact. We conducted a study for comparative evaluation of the performance attained from using these as preprocesses for shot type classification.

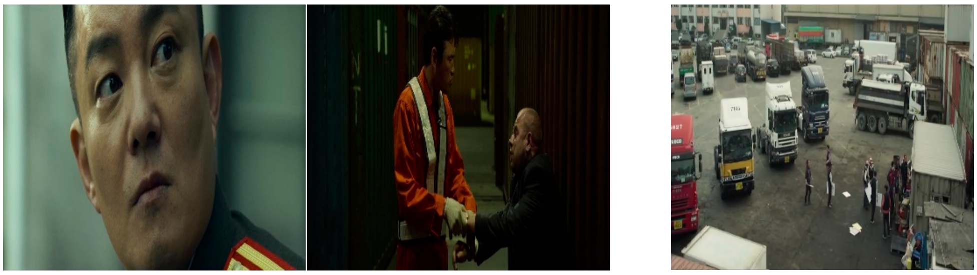
The proposed approach automatically classifies close-up shots, medium shots, and long shots using a CNN and semantic segmentation, and the accuracy of shot type classification is enhanced by applying the boundary relationships of objects to the frames extracted from films using Mask R-CNN and Yolact. For the data, a total of 11,066 frames extracted from 89 films, such as *Inception*, were used.

Section 2 of this article explains the works related to the background knowledge of shot type classification technology. Section 3 explains the structure suggested for shot type classification, and contemplates the experiment conducted, and its results. Lastly, the results are summarized in Section 4.

2. Related Works

2.1. Shot Type Classification

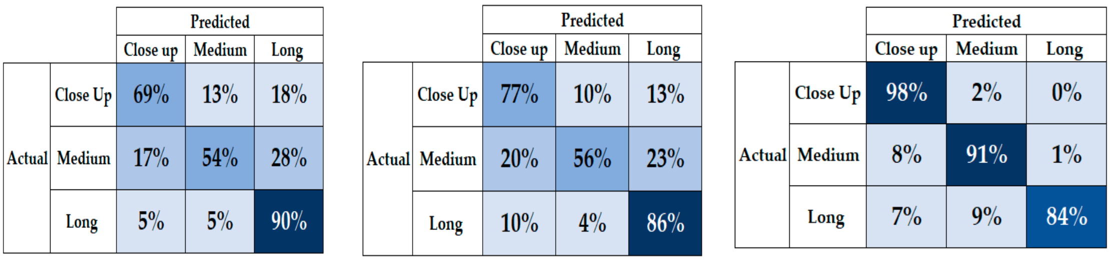
The various studies that have classified shot types can be largely categorized into two methods: Methods that use the face, and methods that use the CNN. The first method classifies shot types based on the proportion of the face that occupies the screen. In this type of study, although the accuracy of movie shot classification is high when there is a face in the frame, as shown in the two frames of Figure 2a, there is a disadvantage in that the accuracy of shot type classification is low when there is no face in the frame, as shown in the frame for Figure 2b. Also, if no person is on-screen, the shot type cannot be classified [5–8].



(**a**) (**b**)

Figure 2. (**a**) Shots with faces; (**b**) a shot without a visible face.

The second type of study uses a CNN to classify close-up shots, medium shots, and long shots [1]. In one study using 400,000 frames extracted from 120 movies as data, learning was done using CNNs with AlexNet, GoogLeNet, and VGG-16 structures to check and compare accuracy [9–11]. Figure 3 is a confusion matrix of the experiment results in which the shot types were classified at 74% accuracy with AlexNet, 75% accuracy with GoogLeNet, and 94% with VGG-16. Unlike the first study, this study can classify shot types even if there is no person on-screen. VGG-16, with the highest accuracy, is a structure using 16 network layers among VGGNets. The VGGNet has a disadvantage, however, in that its performance decreases as the network layers increase, and ResNet solves this disadvantage [12].



		Predicted		
		Close up	Medium	Long
Actual	Close Up	69%	13%	18%
	Medium	17%	54%	28%
	Long	5%	5%	90%

(**a**)

		Predicted		
		Close up	Medium	Long
Actual	Close Up	77%	10%	13%
	Medium	20%	56%	23%
	Long	10%	4%	86%

(**b**)

		Predicted		
		Close up	Medium	Long
Actual	Close Up	98%	2%	0%
	Medium	8%	91%	1%
	Long	7%	9%	84%

(**c**)

Figure 3. Comparison of shot scale classification [1]: (**a**) AlexNet; (**b**) GoogLeNet; (**c**) VGG-16.

Shot type classification using a CNN does not semantically analyze and classify images. Instead, it classifies shot types based on image characteristics and patterns. Therefore, additional information is needed in order to access the images semantically, and this can be done through semantic segmentation, which distinguishes the boundary relationships among objects.

2.2. CNN Technology Used for Shot Type Classification

Recently, shot type classification studies using the CNN has been performed, and the most representative structures of the CNN are VGGNet and ResNet [11,12]. When classifying shot types only using a CNN, some types of shots cannot be classified because the method classifies the shot types based on image characteristics and patterns. Some approaches have been studied for shot type classification based on CNN [13,14], which have accomplished some advances. However, their approaches can be applied to specific domains such as sport movies and music concerts. On the other hand, the purpose of our approach was to classify shots of general movies. Thus, we will compare to VGG-16 and ResNet-50 when applied to movies.

In our study, semantic segmentation was applied to improve on the existing studies using the Mask R-CNN and Yolact. To understand these technologies, we discuss VGGNet, ResNet, and semantic segmentation in detail.

2.2.1. VGGNet

VGGNet was a study conducted to understand the influence of network layers on CNN performance [11,15]. To solely determine the effect of network layer, the experiment was conducted using the smallest filter size, 3 × 3, and 11, 13, 16, and 19 network layers were tested. As a result, it was confirmed that the error rate was similar or worse when the performances of 16 network layers and 19 network layers were compared. Based on such, researchers of VGGNet stopped experimenting with larger number of network layers. The structure of CNN that improved this disadvantage is ResNet, which is explained in Section 2.2.2. Figure 4 is the structure of VGG-16 using 16 network layers among VGGNets. As illustrated in Figure 4, VGGNet is a simple structure that consists only of convolution, max pooling, and full connection, so it is widely used based on its simplicity for understanding the structure and convenience in modifying and testing the structure.

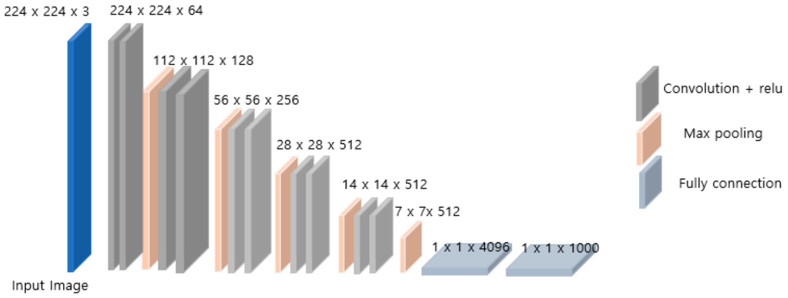


Figure 4. Structure of VGG-16 [16].

The previous shot type classification study using VGG-16 had 94% accuracy, but accuracy may change when applied to a different test group [1].

2.2.2. ResNet

ResNet is a study conducted to investigate how the performance of CNN improves as the layer number of the network increases. To find this, a comparison test was conducted for 20 layers and 56 layers. Figure 5 is a table of the experimental results. From the results of the experiment, as shown in Figure 5, the experiment results with 56 network layers had a higher error rate than 20 network layers. In other words, the higher the number of network layers, the higher the error rate [12,17]. This is because gradient vanishing occurs as the number of network layers increases [18]. In general learning, a back-propagation method that finds the weight and bias that minimizes the loss function value is performed by executing forward propagation, which obtains the loss function value through the neural network, then calculates the inverse of forward propagation. Gradient vanishing is a phenomenon in which the slope that becomes the back propagation gradually disappears and results in insufficient learning.

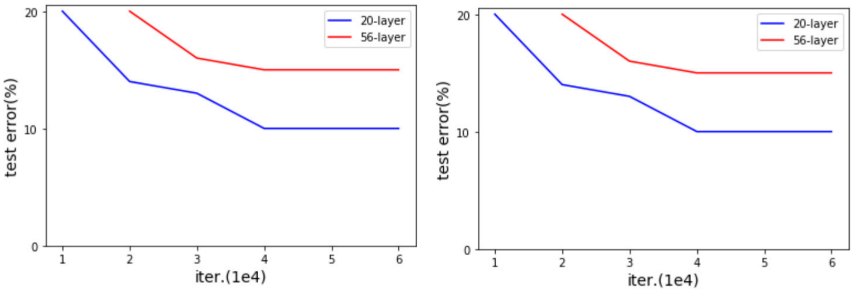


Figure 5. Results of a performance experiment based on network layer change [12,17].

To solve this, the residual block shown in Figure 6b was applied. In general, a CNN uses the plain block shown in Figure 6a, and learns to find the $H(x)$ that processes the input value to output y. As shown in Figure 6b, ResNet applies a skip connection that adds input x to $H(x)$. This performs learning while minimizing $f(x) + x$. Here, since x is a value that does not change, the learning proceeds in order to make $f(x)$ return 0. Because there is a value for x that does not change during the progress of learning through back propagation, at least 1 always exists after performing differentiation. This solved the problem of gradient vanishing by creating a minimum slope for learning [12,17].

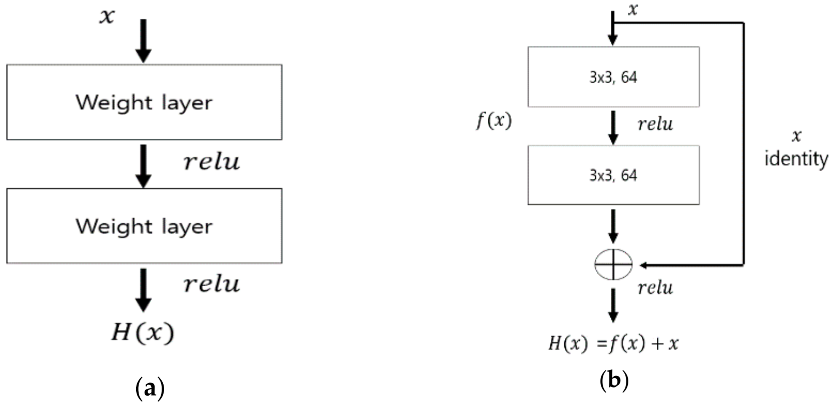


Figure 6. Structures of (**a**) the plain block; (**b**) the residual block [12,17].

2.2.3. Semantic Segmentation

Semantic segmentation classifies which class the images of pixels belong to [19]. As shown in Figure 7, since the boundary of the object can be classified by classifying the class at a pixel level using semantic segmentation, the image can be accessed semantically. In this study, Mask R-CNN and Yolact were used for applying semantic segmentation.

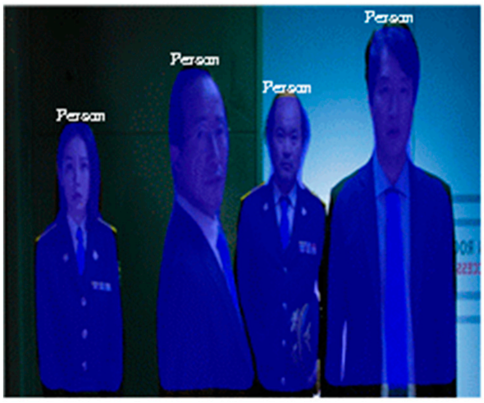


Figure 7. Example of semantic segmentation.

Mask R-CNN combines the function to detect objects using a Faster R-CNN and the function to perform semantic segmentation using a fully convolutional network (FCN) [20–24]. Yolact predicts the mask coefficients for each instance, creates a prototype mask for the entire image, and then combines the two linearly to identify the object and to set the object boundary [25–27]. Unlike Mask R-CNN, Yolact uses a full range of image space without compressing the image, resulting in better semantic segmentation performance than Mask R-CNN [25].

Instance segmentation identifies different objects for each segmentation in more detail than semantic segmentation, but our study used semantic segmentation only, since the study only required determining the object boundary relationships.

3. Comparison of Shot Type Classification Methods Based on Semantic Segmentation

3.1. Shot Type Based on Semantic Segmentation

This article proposed shot type classification using semantic segmentation and ResNet-50. As shown in Figure 8, semantic segmentation was applied to the frames extracted from the films in order to classify the boundary relationships among objects. ResNet-50 alone cannot semantically approach images. Additionally, semantic segmentation was preprocessed on the frames extracted from the movies in anticipation that the shot type and the surface of the objects were closely related. To apply semantic segmentation, Mask R-CNN and Yolact were used, as shown in Figure 9. The we used the two was to investigate in detail how the preprocessing from semantic segmentation affects the performance of shot type classification. Unlike Mask R-CNN, Yolact uses a full range of image space without compressing the image, resulting in better semantic segmentation performance than Mask R-CNN [25]. We can see that Figure 9c, which had semantic segmentation applied via Yolact, shows semantic segmentation performance superior to that of Figure 9b, for which semantic segmentation was applied via Mask R-CNN. After changing the preprocessed frames to 224 × 224, close-up shots, medium shots, and long shots were classified using ResNet-50 for CNN-based Classifier in Figure 8. The previous shot type classification studies used VGG-16 to classify shot types. In our study, ResNet, a deeper network than VGGNet, was used. In order to use the pretrained model in Keras, ResNet-50 was used from among the ResNets having various network layers. Keras is a deep learning library implemented with Python. Mask R-CNN and Yolact for Semantic Segmentation used the pretrained model. The PC used for the experiment comprised a I7-7700K CPU, 32GB memory, and a GTX 1080Ti graphics card with 11GB memory.

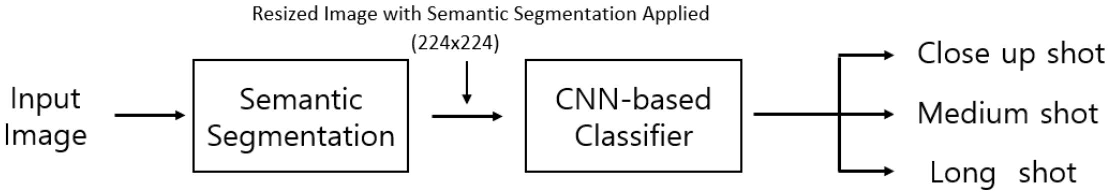


Figure 8. Structure of the suggested semantic segmentation.

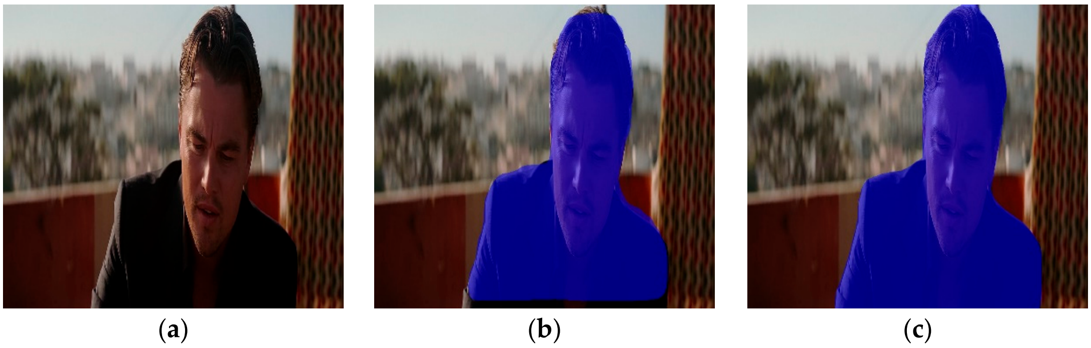


Figure 9. Regular frame examples of semantic segmentation applications: (**a**) Non-mask; (**b**) Mask R-CNN; (**c**) Yolact.

3.2. Experimental Results

In the experimental data, 11,066 frames were extracted randomly from 89 movies, such as *Inception*, and which were classified by shot type through the ground truth operation. Of the 89 movies, 75 were used for training and validation, and the remaining 14 were used for testing. As shown in Table 1, 8679 frames were used for training and 1787 frames were used for validation, with 600 frames used to measure the accuracy of the shot type classification. The ratios of training data and validation data were set to 82% and 15% for shot, and 85% and 15% for medium shot for supporting enough training data.

Table 1. Number of Frames.

Shot Type	Training	Validation	Testing
Close-up	3586	782	200
Medium	2392	413	200
Long	2701	592	200
Total	8679	1787	600

In order to evaluate the accuracy of the shot type classifications, a general frame and frame to which semantic segmentation was applied using Mask R-CNN and Yolact were classified using VGG-16 and ResNet-50 Afterward, the accuracies were evaluated and compared.

As a result, we confirmed that the average accuracy of the shot type classification preprocessed with semantic segmentation was 1.9% higher than shot type classifications using a general frame, as seen in the results presented in Table 2. Also, as shown in Table 3, shot type classification using ResNet-50 was more accurate than shot type classification using VGG-16.

Table 2. Average Accuracy of Shot Type Classification.

Without Semantic Segmentation	Semantic Segmentation–Based Classification Method
93%	94.9%

Table 3. Accuracy of Shot Type Classification.

Classification Type	Without Semantic Segmentation	With Semantic Segmentation	
		Mask R-CNN Added	Yolact Added
VGG-16	92.5%	94.8%	93.8%
ResNet-50	93.5%	95.3%	96%

A conventional shot type classification study using a CNN classified shot types with 94% accuracy for close-ups, medium shots, and long shots using VGG-16 [1]. In the present study, when classifying close-ups, medium shots, and long shots using VGG-16 on general frames, the shot types were classified with 92.5% accuracy, as seen in Table 3. This was expected to occur due to the differences in the data. In this study, the method to classify the shot types with ResNet-50 after preprocessing for semantic segmentation using Yolact is illustrated as a confusion matrix in Figure 10 and showed the highest accuracy among the experimental results (96%). This classified shots with performance better than classifying shot types with only VGG-16 and ResNet-50, as done in the previous study.

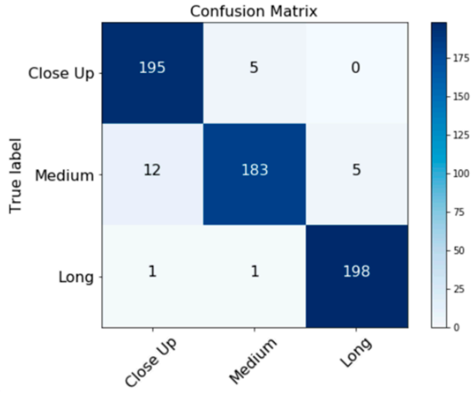


Figure 10. Confusion Matrix of ResNet-50 with Yolact.

The training time of ResNet-50 and Yolact with the highest accuracy was 1654 s, and this model took 66s to assort 600 test frames. So, it took approximately 0.11s per frame. Therefore, this model is difficult to apply in real time to process 30 frames.

3.3. Discussion

The existing study of shot type classification performed classification with 94% accuracy using VGG-16 from among the structures of CNNs. The CNN does not classify shot types based on semantical analysis of frames in the way humans classify shot types because it classifies shot types only based on image characteristics and patterns. To improve this, the present study applied preprocessing with semantic segmentation on the frames extracted from the films. This was to approach the frame after distinguishing the boundaries of the objects in the images. By separating the boundaries of the objects, the frames could be accessed semantically. Mask R-CNN and Yolact were used. The results of the experiment illustrate that shot type classification using ResNet-50 and Yolact had the highest accuracy among the experiments, with 96% accuracy. The average accuracy of shot type classification after preprocessing with semantic segmentation increased by 1.9% (to 94.9%) compared to shot type classification of normal frames. The reason shot type classification using ResNet-50 and Yolact is superior to shot type classification using ResNet-50 and Mask R-CNN is assumed to result from better semantic segmentation performance than Yolact [25]. Additionally, the reason the performance of shot type classification after preprocessing with semantic segmentation is higher than the shot type classification that uses frames without preprocessing from semantic segmentation is that the frames were accessed through semantic segmentation. As a result of analyzing the error rate of 4% in shot type classification by ResNet-50 and Yolact, we confirmed that there are frames in the ground truth task where even humans find it difficult to classify the shot type because of mixed two shot type. This portion accounts for a fraction of the 4% error rate, an example of which is shown in Figure 11. Figure 11 was classified as a close-up in the ground truth work but was classified as a medium shot by ResNet-50 and Yolact.

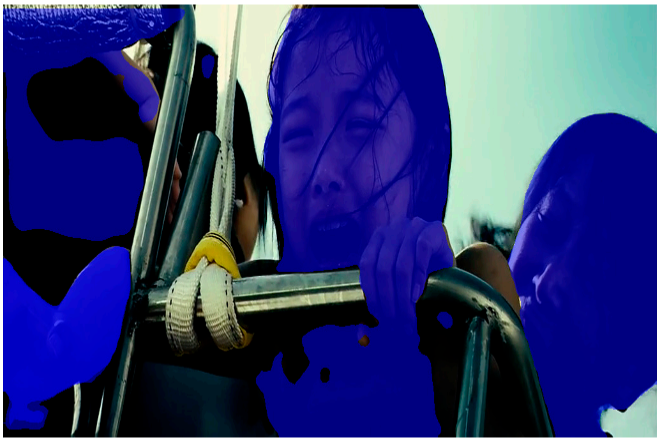


Figure 11. Frame with an Error in Shot Type Classification.

4. Conclusions

The shot type decision is a very important pre-task in movie analysis because each shot contains a lot of information, such as the emotions and the psychology of the characters and the space information. In order to analyze a large number of movies, a technique that automatically classifies shot types is required. Unlike previous studies, our study classified close-up shots, medium shots, and long shots using preprocessing with semantic segmentation and ResNet-50. Throughout this study, shot types were classified with an average accuracy of 94.9%, which is better than the average accuracy of 93% in a previous study without semantic segmentation preprocessing. In particular, when categorized with

ResNet-50 and Yolact, the performance improved by 3% to 96% accuracy, which is superior to the 93% accuracy of the previous study.

As a result of analyzing the 4% error rate of the shot type classification using ResNet-50 and Yolact, in a portion of the 4% error rate, there were frames in the ground truth task where it was difficult even for humans to classify the shot type. To solve this, a further study will be conducted for shot type classification using additional information from each segment based on instance segmentation (which identifies different objects for each segment) instead of semantic segmentation.

Author Contributions: Data curation, H.-Y.B.; Software, H.-Y.B.; Validation, H.-Y.B. and S.-B.P.; Writing—original draft, H.-Y.B.; Writing-review & editing, S.-B.P.; All authors have read and agreed to the published version of the manuscript.

Funding: This article is the result of a basic research project performed under the support of National Research Foundation of Korea funded by the Korean government (Ministry of Education) in 2017 (No. NRF-2017R1D1A1B03036046).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Savardi, M.; Signoroni, A.; Migliorati, P.; Benini, S. Shot scale analysis in movies by convolutional neural networks. In Proceedings of the 25th IEEE International Conference on Image Processing (ICIP), Athens, Greece, 7–10 October 2018; pp. 2620–2624.
2. Katz, S.D.; Katz, S. *Film Directing: Shot by Shot*; Michael Wiese Productions: Studio City, CA, USA, 1991.
3. Thompson, R. *Grammar of the Edit*, 2nd ed.; Focal Press: Waltham, MA, USA, 2009.
4. Canini, L.; Benini, S.; Leonardi, R. Affective analysis on patterns of shot types in movies. In Proceedings of the 7th International Symposium on Image and Signal Processing and Analysis (ISPA), Dubrovnik, Croatia, 4–6 September 2011; pp. 253–258.
5. Cherif, I.; Solachidis, V.; Pitas, I. Shot type identification of movie content. In Proceedings of the 9th International Symposium on Signal Processing and Its Applications, Sharjah, UAE, 12–15 February 2007; pp. 1–4.
6. Tsingalis, I.; Vretos, N.; Nikolaidis, N.; Pitas, I. Svm-based shot type classification of movie content. In Proceedings of the 9th Mediterranean Electro Technical Conference, Istanbul, Turkey, 16–18 October 2012; pp. 104–107.
7. Marín-Reyes, P.A.; Lorenzo-Navarro, J.; Castrillón-Santana, M.; Sánchez-Nielsen, E. Shot classification and keyframe detection for vision based speakers diarization in parliamentary debates. In *Conference of the Spanish Association for Artificial Intelligence*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 48–57.
8. Baek, Y.T.; Park, S.-B. Shot Type Detecting System using Face Detection. *J. Korea Soc. Comput. Inf.* **2012**, *17*, 49–56. [CrossRef]
9. Krizhevsky, A.; Sutskever, I.; Hinton, G. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*; Neural Information Processing Systems: Nevada, MA, USA, 2012.
10. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 1–9.
11. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale ImageRecognition. *arXiv* **2014**, arXiv:1409.1556.
12. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
13. Lin, J.C.; Wei, W.L.; Liu, T.L.; Yang, Y.H.; Wang, H.M.; Tyan, H.R.; Liao, H.Y.M. Coherent Deep-Net Fusion to Classify Shots In Concert Videos. *IEEE Trans. Multimed.* **2018**, *20*, 3123–3136. [CrossRef]
14. Minhas, R.A.; Javed, A.; Irtaza, A.; Mahmood, M.T.; Joo, Y.B. 'Shot classification of field sports videos using alexnet convolutional neural network. *Appl. Sci.* **2019**, *9*, 483. [CrossRef]

15. Jun, H.; Shuai, L.; Jinming, S.; Yue, L.; Jingwei, W.; Peng, J. Facial Expression Recognition Based on VGGNet Convolutional Neural Network. In Proceedings of the 2018 Chinese Automation Congress (CAC), Xi'an, China, 30 November–2 December 2018; pp. 4146–4151.
16. Muhammad, U.; Wang, W.; Chattha, S.P.; Ali, S. Pre-trained VGGNet Architecture for Remote-Sensing Image Scene Classification. In Proceedings of the 2018 24th International Conference on Pattern Recognition (ICPR), Beijing, China, 20–24 August 2018; pp. 1622–1627.
17. Li, B.; He, Y. An Improved ResNet Based on the Adjustable Shortcut Connections. *IEEE Access* **2018**, *6*, 18967–18974. [CrossRef]
18. Huang, G.; Sun, Y.; Liu, Z.; Sedra, D.; Weinberger, K.Q. Deep networks with stochastic depth. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 646–661.
19. Romera-Paredes, B.; Torr, P.H. Recurrent instance segmentation. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 312–329.
20. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
21. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-time Object Detection with Region Proposal Networks. *arXiv* **2015**, arXiv:1506.01497.
22. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2980–2988.
23. Almutairi, A.; Almashan, M. Instance Segmentation of Newspaper Elements Using Mask R-CNN. In Proceedings of the 18th IEEE International Conference On Machine Learning And Applications (ICMLA), Boca Raton, FL, USA, 16–19 December 2019; pp. 1371–1375.
24. Su, H.; Wei, S.; Yan, M.; Wang, C.; Shi, J.; Zhang, X. Object Detection and Instance Segmentation in Remote Sensing Imagery Based on Precise Mask R-CNN. In Proceedings of the IGARSS 2019—2019 IEEE International Geoscience and Remote Sensing Symposium, Yokohama, Japan, 28 July–2 August 2019; pp. 1454–1457.
25. Bolya, D.; Zhou, C.; Xiao, F.; Lee, Y.J. YOLACT: Real-Time Instance Segmentation. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; pp. 9156–9165.
26. Ying, H.; Huang, Z.; Liu, S.; Shao, T.; Zhou, K. Embedmask: Embedding coupling for one-stage instance segmentation. *arXiv* **2019**, arXiv:1912.01954.
27. Benbarka, N.; Zell, A. FourierNet: Compact mask representation for instance segmentation using differentiable shape decoders. *arXiv* **2020**, arXiv:2002.02709.