A Unified Framework for Shot Type Classification Based on Subject Centric Lens (Supplementary Material)

Anyi Rao[1], Jiaze Wang[1], Linning Xu[1], Xuekun Jiang[2],
Qingqiu Huang[1], Bolei Zhou[1], and Dahua Lin[1]

[1] CUHK - SenseTime Joint Lab, The Chinese University of Hong Kong
[2] Communication University of China
{anyirao, hq016, bzhou, dhlin}@ie.cuhk.edu.hk, xkjiang@cuc.edu.cn
jzwang.cuhk@gmail.com, linningxu@link.cuhk.edu.cn

1 Implementation Details

1.1 Student Generator and Discriminator Settings

The student generator contains 6 convolutional layers. Each layer adopts the convolutional filter that has a 3×3 kernel with stride 1 and padding 1. The input and output of each layer are of dimension 224×224. At the last layer, a sigmoid function is applied to output the subject map with each pixel value ranging from 0 to 1.

The discriminator has 12 convolutional layers. The kernel sizes are set to 3×3. It down samples an input with width 224 × height 224 to a binary output for the real or fake classification.

1.2 Experiments Settings for I3D

Our I3D models are initialized from 2D CNNs pre-trained on ImageNet [4], and we choose ResNet [2] as our backbone. Every input clip has 8 frames, which is sampled from 64 consecutive frames with a stride of 8. The I3D input size is $8 \times 3 \times 224 \times 224$. The batch size is set to 128 and the momentum is set to 0.9. We train the 3D models for 60 epochs with mini-batch SGD, The initial learning rate is 0.01 and the learning rate will be divided by $30th$ at the $40th$ and $50th$ epoch. Location jittering, horizontal flipping, corner cropping, and scale jittering are used for data augmentation.

2 Accuracy Analysis for Individual Class

As shown in Table 1, we list the performance of our full SGNet w/ Var (img+flow) for each individual class. The portion of each class is the same in the training, validation and testing sets, which corresponds to the natural distribution of movie data.

Table 1. Each individual class's accuracy

Classes	ECS	CS	MS	FS	LS	Staic	Motion	Push	Pull
Accuracy (↑)	85.8	84.5	86.9	87.0	95.1	95.0	78.5	28.4	27.4
Data percentage	19%	22%	22%	20%	17%	60%	27%	11%	2%

Table 2. Confusion matrix of shot scale prediction

Classes	ECS	CS	MS	FS	LS
ECS	**0.86**	0.12	0.01	0.01	0.00
CS	0.08	**0.84**	0.08	0.00	0.00
MS	0.01	0.08	**0.87**	0.04	0.00
FS	0.00	0.01	0.06	**0.88**	0.04
LS	0.00	0.00	0.01	0.04	**0.95**

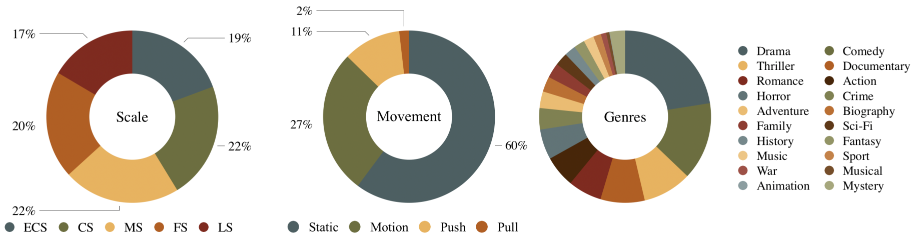
In shot scale prediction, each class's performance is quite similar. The class with the highest accuracy is *long shot* (95.1%) and the class with the lowest accuracy is *close-up shot* (84.5%). Observed from The confusion matrix of scale as shown in Table 2. It is reasonable to see that shot is misclassified to the neighborhood of its ground truth. However, the performance for shot movement prediction varies greatly for different classes due to the imbalanced data distribution. *Static* achieved the highest accuracy 95.0% while the accuracies of *pull* and *push* are around 30%, because these classes take small portions in the whole dataset. Thus, how to solve this unbalanced data problem is very important in the future works.

3 Automatic Video Editing Details

Video editing is the central art of film making, where the shot type plays an important role in conveying the underlying stories and emotions. The usage of different shots usually depends on the movie genres and the preferences of the filmmakers. The model we propose in this paper could classify the shot type of a givens video shot and generate edited shot candidates from random proposals for users, based on their flavors and specified shot types.

In the *Titanic* editing case, the original shot is obtained from a single shot of the film, where we apply shot detection [5] to obtain the original shot clip. The original resolution of it is 1080P. We first cut it into four segments using active-speaker [3]. Then we transfer the second and fourth segments from medium shot to close shot using the pipeline proposed in the Figure 6 in the paper. After this transfer, the resolution of these two shots are about 500P∼720P. Then they are downscaled to 480P. And the first and third segments are directly downscaled from 1080P to 480P without other changes.

In the *DAVIS* editing cases, each original video is a single shot with 720P resolution. In the first and second examples *parkour* and *boxing*, we select a



(a) Shot Type Statistics (b) Genres of Movie Trailers covered in *MovieShots*

Fig. 1. (a) Proportion of each scale and movement type in *MovieShots* dataset; (b) The movie trailers used in *MovieShots* cover a wide range of genres

segment out from the original and change the shot type style accordingly. In the third example *man-bike*, we change the whole shot from full shot to close shot. The changed segments and unchanged segments are all downscaled to 480P.

4 Dataset

4.1 Annotation Details

The criterion of each shot class annotation is made under the supervision from professional film producers and scholars in the field of cinematic arts. We also reference to previous works [1,6,7]. Note that, all the annotations are done by the professionals.

Scale Annotation. The shot scale is categorized into five classes, long shot, full shot, medium shot, close-up shot and extreme close-up shot.

1. Long shot: The long shot is taken from a great distance, it almost always an exterior shot and shows much of the locale.
2. Full shot: The full shot barely includes the human body in full, with the head near the top of the frame and the feet near the bottom.
3. Medium shot: The medium shot contains a figure from the knee or waist up.
4. Close-up shot: The close-up shot concentrates on a relatively small object such as an actors face.
5. Extreme close-up shot: Extreme close-up shot is a variation of close-up shot, it focus on more detail of the object than close-up shot.

Notion: Although each shot category has clear definition, in film-making, there are still few shots in the fuzzy zone of two categories, especially in full shot and medium shot. In this case, we take the main character facing the camera as the criterion to judge the scale of the shot.

Movement Annotation. The shot movement is classified into four categories, static shot. pans and tilts shot, pull shot, push shot.

1. Static shot: The camera is fixed, but the subject is free to move.
2. Motion shot: The camera moves or rotates.
3. Pull shot: The camera zooms in.
4. Push shot: The camera zooms out.

Notion: If a shot is taken by a camera that is carried by the photographer, it is a hard case, since the photographer has a slightly body shake. In this case, our criterion is that if the annotator is unable to clearly judge the direction of shot movement, this is a static shot.

4.2 Copyright Issue

We use publicly available trailers rather than full movies in our dataset and provide the URLs instead of the original videos, for which copyright is not a problem.

References

1. Giannetti, L.D., Leach, J.: Understanding movies, vol. 1. Prentice Hall Upper Saddle River, New Jersey (1999) 3
2. 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. pp. 770–778 (2016) 1
3. Roth, J., Chaudhuri, S., Klejch, O., Marvin, R., Gallagher, A., Kaver, L., Ramaswamy, S., Stopczynski, A., Schmid, C., Xi, Z., et al.: Ava-activespeaker: An audio-visual dataset for active speaker detection. arXiv preprint arXiv:1901.01342 (2019) 2
4. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. International journal of computer vision **115**(3), 211–252 (2015) 1
5. Sidiropoulos, P., Mezaris, V., Kompatsiaris, I., Meinedo, H., Bugalho, M., Trancoso, I.: Temporal video segmentation to scenes using high-level audiovisual features. IEEE Transactions on Circuits and Systems for Video Technology **21**(8), 1163–1177 (2011) 2
6. Wang, H.L., Cheong, L.F.: Taxonomy of directing semantics for film shot classification. IEEE Transactions on Circuits and Systems for Video Technology **19**(10), 1529–1542 (2009) 3
7. Xu, M., Wang, J., Hasan, M.A., He, X., Xu, C., Lu, H., Jin, J.S.: Using context saliency for movie shot classification. In: 2011 18th IEEE International Conference on Image Processing. pp. 3653–3656. IEEE (2011) 3