A Unified Framework for Shot Type Classification Based on Subject Centric Lens

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

Abstract. *Shots* are key narrative elements of various videos, *e.g.* movies, TV series, and user-generated videos that are thriving over the Internet. The types of shots greatly influence how the underlying ideas, emotions, and messages are expressed. The technique to analyze shot types is important to the understanding of videos, which has seen increasing demand in real-world applications in this era. Classifying shot type is challenging due to the additional information required beyond the video content, such as the spatial composition of a frame and camera movement. To address these issues, we propose a learning framework Subject Guidance Network (SGNet) for shot type recognition. SGNet separates the subject and background of a shot into two streams, serving as separate guidance maps for scale and movement type classification respectively. To facilitate shot type analysis and model evaluations, we build a large-scale dataset *MovieShots*, which contains $46K$ shots from $7K$ movie trailers with annotations of their scale and movement types. Experiments show that our framework is able to recognize these two attributes of shot accurately, outperforming all the previous methods. [1]

1 Introduction

In 1900, film pioneer George Albert Smith firstly introduced shot type transitions into videos, which revolutionized traditional narrative thought and this technique remains widely used in today's video editing [48]. *Shot*, a series of visual continuous frames, plays an important role in presenting the story. It can be recognized from multiple attributes, such as *scale* and *movement*. As illustrated in Fig. 1, five scale types and four movement types of shots are widely adopted in video editing, serving for different scenes and emotional expressions.

We are in the era of web 2.0 where user-generated videos proliferate, the techniques to analyze shot types have seen increasing demand in real-world applications: 1) With the capability of recognizing shot types, the videos shared online can be automatically classified or organized not only by their content,

[1] The dataset and related codes are released here in compliance with regulations.

Fig. 1. Demonstrations of five *scale* types and four *movement* types of video shots sampled from *MovieShots* dataset. It is noticed that shot scales can reveal information of a story from different aspects. For example, *long shots* usually indicate the location information, while *close-up shots* are widely used for emphasizing the identities of the characters. *Medium shots* and *full shots* are good at depicting an event, while *extreme close-up shots* are used for symbolic expressions or intensifying the story emotion. For movement types, we notice that *static shots* are mainly used for narrative purposes and *motion shots* try to track moving objects. *Push shots* aim to emphasize the content information of the main subject while pull shots shrink the figure of the main subject and gradually reveal its surrounding environment

e.g. object categories, but also by shot types. Thus the system will be able to respond to queries like long shots over a city etc. 2) By analyzing the sequence of shots in movies, we may provide a data-driven view of how professional movies are constructed. Such insight can help ordinary users to make videos that look more professional – one can even build softwares to guide video production for amateurs.

Despite the potential value of shot type analysis, it is true that most previous works in computer vision primarily focus on objective content understanding. For example in video analysis, we focus on classifying and localizing the actions [44,17,33], while shot type analysis has been rarely investigated and lacks appropriate benchmark. Existing datasets on shot type classification are either too small or not publicly available. However, we believe that the analysis of cinematic techniques (e.g. shot types) are also equally important.

To facilitate researches along this direction, we construct a new dataset *MovieShots*, which composes of over $46K$ shots collected from public movie trailers, with annotations on five scale types and four movement types. We select out scale and movement from many other shot attributes, as they are the two most common and distinguishable attributes that can uniquely characterize a shot in video, where the *scale* type is decided by the amount of subjects within the frame, and the *movement* type is determined by the camera motion [19].

We further propose a novel subject centric framework, namely Subject Guidance Network (SGNet), to classify the scale and movement type of a shot. The key point here is to find out the dominant subject in a given shot, then we can decide its scale according to the portion it takes, and differentiate between the camera movement and subject movement to determine the movement type. SGNet successfully separates the subject and background in a shot and takes them to guide the full images to predict the labels for scale and movement type.

The contributions of this work are as follows: 1) We construct *MovieShots*, a large-scale $46K$ shot dataset with professionally annotated scale and movement attributes for each shot. 2) SGNet is proposed to classify scale and movement type simultaneously based on the subject centric lens. Our experiments show that this framework greatly improves the classification performance comparing to traditional methods and conventional deep networks TSN [44] and I3D [7].

2 Related Work

Shot Type Classification Datasets. Traditional shot type classifications mainly focus on sports videos [16,14,27]. Sports video is a special kind of video that contains many clips such as video replays or comments, which is hard to transfer to general video scenarios. Previous movie shot type researches [43,6] are limited on their evaluation benchmarks. They collect no more than twenty films with about one-thousand shots. There is no public available dataset to test the functionality of these methods. It is noticed that these datasets annotate either *scale* or *movement* attribute only, lacking a comprehensive description for a shot. In order to solve these limitations, we collect a $10\times \sim 100\times$ larger dataset, with $46K$ video shots annotated with both *scale* and *movement* attributes from more than $7K$ public movie trailers.

Shot Type Classification Methods. Conventional methods for shot *scale* classification use SVM with dominate color region [29], low-level texture features [55,1], or optical flow [27]. Decision tree method [50] sets up fixed rules to classify the scale type of a shot. Scene depth [3] is applied to infer the scale but is limited to the depth approximation accuracy and lacks generalization ability. For *movement* type classification, traditional approaches rely on the manual design of a motion descriptor. *e.g.* [22,34] design *motion vectors* CAMHID and 2DMH to capture the camera movement. [43] leverage optical flow to find an alternative of the motion vectors. However, all these methods heavily depend on hand-crafted features that are not applicable to general cases. Our SGNet separates the subject from image and considers both the spatial and temporal configurations of a given shot, achieving much improved performance with better generalization ability.

Video Analysis and Understanding in One Shot. Most previous single shot video understanding tasks [17,33] are about action recognition [18,44,46] and temporal action localization [8,53,63]. Video object detection [12,21], video ob-

Table 1. Comparisons with other datasets

	#Shot	#Video	Scale	Move.
Lie 2014 [4]	327	327		✓
Unified 2005 [14]	430	1	✓	
Sports 2007 [59]	1,364	8	✓	
Soccer 2009 [30]	1,838	1	✓	
Cinema 2013 [6]	3,000	12	✓	
Context 2011 [55]	3,206	4	✓	
Taxon 2009 [43]	5,054	7		✓
MovieShots	46,857	7,858	✓	✓

Table 2. Statistics of *MovieShots*

	Train	Val	Test	Total
Number of Movies	4,843	1,062	1,953	7,858
Number of Shots	32,720	4,610	9,527	46,857
Avg. Dur. of Shot (s)	3.84	5.31	3.78	3.95

ject segmentation [61,54], video person recognition [58,49], video-text retrieval [51,60] and some low-level vision tasks, *e.g.* video inpainting [56] video super-resolution [62,31] are also applied in single shot videos. However, research on video shot type is rarely explored, despite of its huge potential for video understanding. We set up a benchmark with our *MovieShots* dataset and conduct a detailed study on it.

3 *MovieShots* Dataset

To facilitate the shot type analysis in videos, we collect *MovieShots*, a large-scale shot type annotation set that contains $46K$ shots from 7858 movies. The details of this dataset are specified as follows.

3.1 Shot Categories

Following previous definition on shot type [19,43,55,28,40,26], shot *scale* is defined by the amount of subject figure that is included within the frame, while shot *movement* is determined by the camera movement or the lens change.

Shot *scale* has five categories: 1) *long shot* (LS) is taken from a long distance, sometimes as far as a quarter of a mile away; 2) *full shot* (FS) barely includes the human body in full; 3) *medium shot* (MS) contains a figure from the knees or waist up; 4) *close-up shot* (CS) concentrates on a relatively small object, showing the face or the hand of a person; (5) *extreme close-up shot* (ECS) shows even smaller parts such as the image of an eye or a mouth.

Shot *movement* has four categories: 1) in *static shot*, the camera is fixed but the subject is flexible to move; 2) for *motion shot*, the camera moves or rotates; 3) the camera zooms in for *push shot*, and 4) zooms out for *pull shot*. While all the four movement types are widely used in movies, the use of *push* and *pull* shots only takes a very small portion. The usage of different shots usually depends on the movie genres and the preferences of the filmmakers.

3.2 Dataset Statistics

MovieShots consists of $46,857$ shots from $7,858$ movie trailers, covering a wide variety of movie genres to ensure the inclusion of all scale and movement types

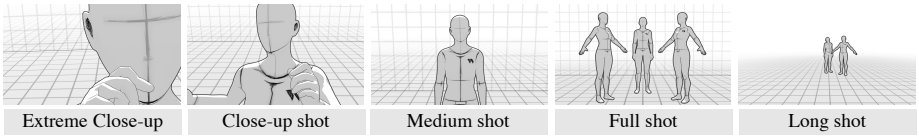

Fig. 2. Prototypes of annotation corresponding to extreme close-up shot, close-up shot, medium shot, full shot and long shot

of shot. Table 1 compares *MovieShots* with existing private shot type datasets, noting that none of them are publicly available. *MovieShots* is significantly larger than others in terms of the shot number and the video coverage, with a more comprehensive annotation covering both the *scale* type and the *movement* type for each shot.

3.3 Annotation Procedure

Building a large-scale shot dataset is challenging in two aspects: appropriate data collections and accurate data annotations. We firstly crawled more than $10K$ movie trailers online. Because movie trailers usually contain advertisements and big subtitles displaying the actor/director information, we firstly cleaned them with auto advertisement/big text detection and went through a second round of manual check. Noting that shot detection is a well solved problem and we used an off-the-shelf approach [42,36] to cut shots in these trailers and filtered out failure cases with manual check. All our annotators are cinematic professionals in film industries or cinematic arts majors, who provide high quality labels. We also set up annotation prototypes for these well defined criterion of shot types, as illustrated in Fig. 2. Additionally, three rounds annotation procedures have been done to ensure the high consistency. We finally achieve 95% annotation consistency, with those inconsistent shots being filtered out in our experiments.

4 SGNet: Subject Guidance Network

In this section, we introduce our Subject Guidance Network (SGNet) for *scale* and *movement* type classification. The overall framework is shown in Fig. 3.

We firstly divide a shot into N clips to capture the contextual variations along the temporal dimension. Each clip passes through a two-branch classification network and outputs a feature vector. The feature vectors coming from the N clips are pooled together and pass through a fully-connected layer to get the final prediction.

It is noticed that, the separation of subject and background information is critical for both two tasks. While the *scale* type depends on the portion of the subject in the shot, the *movement* type relies on the background motion rather than the subject motion, as the changes of the background information are closely related to the camera motion. To reduce the burden of the whole

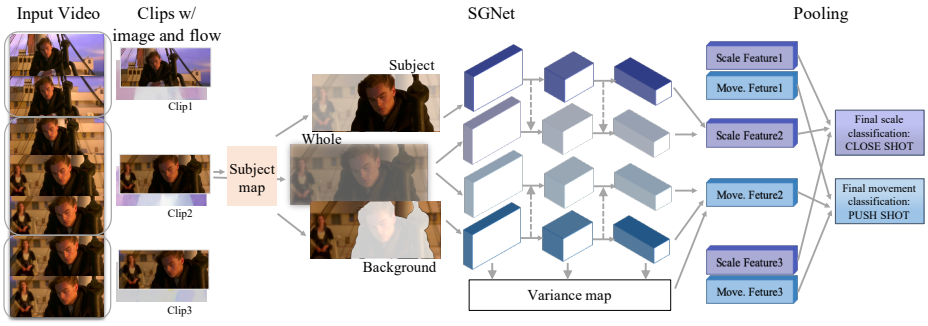

Fig. 3. Pipeline of Subject Guidance Network (SGNet). A subject map is generated from each clip's image. With this subject map, we take subject and background to guide scale and movement prediction respectively

pipeline, a light-weight *subject map generator* is designed to separate the subject and background[2] in both image and flow in an effective way. The *subject map* here is a saliency map sharing the same width and height as the original image with values in range of zero to one. We use *subject map* to guide the whole image to predict the *scale*, and use the *background map* to guide the whole image to predict the *movement* with the help of an obtained *variance map*.

In the following two subsections, we will first introduce the subject map guidance for shot type classification in Section 4.1, and elaborate on our subject map generation in Section 4.2.

4.1 Subject Map Guidance Classification

As discussed before, the separation of subject and background is crucial to shot type classification. In this section, we firstly explain how the obtained subject map guides scale and movement classification respectively.

For *scale* type classification, the subject map $[1 \times W \times H]$ element-wise multiplies with the whole image $[3 \times W \times H]$ to get an subject image $[3 \times W \times H]$. Then we take the whole image and the obtained subject image as two input pathways. Each of them is sent into a ResNet50 [23] network. To apply subject guidance, we fues the subject feature map from different stages of feature representations into the whole image pathway. Specifically, these fusion connections are applied right after pool_1, res_2, res_3, and res_4 in the ResNet50 [23] backbone. The fusion is conducted by lateral connections [11].

For *movement* type classification, the background guidance is applied in a similar way to the scale type prediction. Additionally, a *variance map* module is further introduced, inspired by the fact that the changes of appearance along time is a cue for movement classification. For example, in a *static* shot, the

[2] Background image is equal to the whole image minus the subject part.

appearances of background among different clips are almost the same, while the background appearances changes significantly as time changes in *motion* shots, *push* shots and *pull* shots. We calculate one variance map $\mathbf{V}_m \in \mathcal{R}^{N\times N}$ for each shot among its different clips ($N = 8$ clips in our experiments) at different stages of the backbone ResNet50. Specifically, these stages include those after pool_1, res_2, res_3, and res_4, the same as those used in the previous fusions. We apply inner product between the two normalized feature maps $\mathbf{F}_{m,i}, \mathbf{F}_{m,j} \in \mathcal{R}^{h_m\times w_m\times c_m}$ from two clips i, j at stage m to get $\mathbf{V}_{m,(i,j)}$. The inner product here is equivalent to calculating the cosine similarity, which captures the similarities between different clips. $\mathbf{V}_m$ is achieved by concatenating all possible clips pair $\mathbf{V}_{m,(i,j)}$. Finally, all $\mathbf{V} = \{\mathbf{V}_m\}$ among M stages are concatenated along the channel-wise dimension and are fed into a two-layer FC for classification. The classification results using variance maps are fused with the image classification results for the final prediction.

4.2 Subject Map Generation

Now we elaborate on how we separate the subject from the background with our light-weight *subject map generator*.

Conventional saliency/attention map methods employ hand-crafted visual features or heuristic priors [9,64], which are incapable of capturing high-level semantic knowledge, making the predicted map unsatisfactory. Pre-trained state-of-the-art deep networks [32,13] are usually very large with more than $50 \sim 100$ layers and are not easy to be taken as a submodule in the designed networks to fine tune, considering the high computational costs. From another perspective, to train a randomly initialized subject network from scratch with only shot type label is impractical, since the supervision signal is too weak and the network is unable to converge, considering the subject map is a pixel-wise prediction but the annotation is a video-level label.

To strike a balance between the performance and the computational efficiency, we resort to knowledge distillation (KD) [24,52], considering that it is easy and flexible to learn, and achieves state-of-the-art performance on classification problems [10]. A light-weight *student generator* (a 6-layer CNN) learns from its *teacher network* (a MSRA10K [25] pre-trained R^3Net [13]). Note that a naive KD using $\mathcal{L}_2$ loss is usually suboptimal because it is difficult to learn the true data distribution from the teacher and may result in missing generation details. Therefore, an additional adversarial loss with the help of Generative Adversarial Networks (GAN) [45,47,41,2,20] is adopted. In all, the student generator is trained by minimizing the following three-term loss,

$$\mathcal{L} = \alpha\mathcal{L}_2 + \beta\mathcal{L}_{\text{adv}} + \mathcal{L}_{\text{cross}}.$$

The first loss term L_2 is the least square error between the generated subject map and its corresponding pseudo subject map, which aims to mimic the output of teacher network. L_2 loss alone is not able to teach the student network to generate fine grained details since it does not consider the constraints from

the whole data distribution. The second term $\mathcal{L}_{\text{adv}}$ is given by a learned discriminator, which is trained to compete with the student generator to learn the true data distribution. The discriminator takes the subject map from the teacher network as *real* and the output from student generator as *fake*. Finally, we take cross-entropy loss L_{cross} as our classification loss and be jointly trained with the whole pipeline to encourage right predictions.

5 Experiments

5.1 Experiments Setup

Data. All the experiments are conducted on *MovieShots*. The whole dataset is split into *Train*, *Val*, and *Test* sets with a ratio 7:1:2, as shown in Table 2.

Implementation Details. We take cross-entropy loss for the classification result. The shot is evenly split into 3 clips in training and 25 clips in testing. The fusing function from the subject/background map to whole image is implemented by concatenating the output from the two branches. Image and flow are set up as two inputs and their classification score are fused to get the final results. We train these models for 60 epochs with mini-batch SGD, where the batch size is set to 128 and the momentum is set to 0.9. The initial learning rate is 0.001 and the learning rate will be divided by 10 at the $20th$ and $40th$ epoch.

Evaluation Metrics. We take the commonly used Top-1 accuracy as the evaluation metric. Specifically, in our experiment, we denote Acc_S for scale classification performance and Acc_M for movement classification performance.

5.2 Overall Results

We reproduce DCR [29], CAMHID [22] and 2DMH [34] according to their papers. DCR [29] clusters dominant color sets and predicts shot type based on the ratio of different color sets. CAMHID [22], 2DMH [34] are based on motion vectors. CAMHID [22] takes SVD to get the dominant components. 2DMH [34] disentangles the magnitude and orientation of motion vectors. TSN [44] and I3D [7] are experimented using authors' code repositories. SGNet adopts ResNet50 [23] as the backbone. All the network weights are initialized with pre-trained models from ImageNet [39] unless specially stated.

Overall Results Analysis. 1) *Traditional Methods.* The overall results are shown in Table 3. The performances of DCR [29], CAMHID [22] and 2DMH [34] are restricted by their poor representations.

2) *3D Networks.* For movement classification, I3D-ResNet50 achieves better result than TSN-ResNet50 (img + flow) since it captures more temporal relationships. With Kinetics400 [7] pre-trained, I3D-ResNet50 gets 4.8 boost on Acc_M. But in scale classification, I3D-ResNet50 performs worse than TSN-ResNet50

Table 3. The overall results on shot scale and movement type classification

Models	Acc_S (↑)	Acc_M (↑)
DCR, Li *et al* [29]	51.53	33.20
CAMHID, Wang *et al* [22]	52.37	40.19
2DMH, Prasertsakul *et al* [34]	52.35	40.34
I3D-ResNet50 [7]	76.79	78.45
I3D-ResNet50-Kinetics [7]	77.11	83.25
TSN-ResNet50 (img) [44]	84.08	70.46
TSN-ResNet50-Kinetics (img) [44]	84.18	71.61
TSN-ResNet50 (img + flow) [44]	84.10	77.13
TSN-ResNet152 (img + flow) [44]	84.95	78.02
SGNet (img)	87.21	71.30
SGNet (img + flow)	87.50	80.65
SGNet w/ Var (img)	87.42	80.57
SGNet w/ Var (img + flow)	**87.57**	**81.86**
SGNet w/ Var-Kinetics (img + flow)	**87.77**	**83.72**

(img). The reason might be that I3D-ResNet50 is not good at capturing the spatial configuration of frames in predicting the shot scale. The performance of I3D-ResNet101 is similar to I3D-ResNet50 since deeper 3D networks needs more data to improve the performance.

From another perspective, 3D CNNs are much more computational expensive and need dense samples from videos, which causes the low speed for training and inference. We choose 2D TSN-ResNet50 as our backbone. The results prove that this 2D network can achieve better results than 3D networks with our careful designs. Deep 2D network TSN [44] using image (TSN img) achieves $\sim 30\%$ raise on Acc_S and Acc_M than traditional methods, as it captures high-level semantic information such as the subject contours and the temporal relationship in a shot.

3) *Deeper Backbones.* To show that the improvement does not come from the increase of model parameters, we compare SGNet w/ Var (img + flow) (use ResNet50 backbone) with TSN-ResNet152 (img + flow). SGNet w/ Var (img + flow) outperforms TSN-ResNet152 by a margin of 2.62 on Acc_S and 3.84 on Acc_M, with 15% fewer parameters and 19% fewer GFLOPs.

4) *2D Networks and Kinetics Pre-training.* Our full model SGNet w/ Var (img + flow) which includes subject map guidance, motion information flow, and variance map, improves 3.49 (relatively 4.12%) on Acc_S and 11.40 (relatively 16.11%) on Acc_M compared to TSN (img), and 3.47 (relatively 4.15%) on Acc_S and 4.73 (relatively 6.13%) on Acc_M compared to TSN (img + flow). The full model get further improvements by 0.2 on Acc_S and 1.8 Acc_M with Kinetics [7] pre-trained. This result shows that action recognition dataset can bring more help to shot movement predictions.

Table 4. Comparison of different subject or/and background map guidance.

#	Settings	Acc_S (↑)	Acc_M (↑)
1	Base (TSN w/ Var img+flow)	84.15	77.25
2	Subject only	79.97	74.65
3	Back only	79.60	75.80
4	Base+Subj	**87.57**	80.86
5	Base+Back	87.10	**81.86**
6	Base+Subj+Back	87.31	81.54

Analysis of Our Framework. Based on TSN (img), SGNet (img) takes the advantage of subject map guidance and improves the scale and movement results by 3.13 and 0.96 respectively, which shows the usefulness of subject guidance especially for scale type prediction. With the help of variance map, SGNet w/ Var (img) raise the movement classification performance from 70.46 to 80.57 (relatively 14.35%). Similarly, flow (SGNet img+ flow) helps the model to improve the movement results from 70.46 to 80.65 (relatively 14.46%). These results show that variance map and flow both capture the movement information and contribute to the great performance on movement type classification. As for scale type classification, variance map and flow bring slight improvements ($0.2 \sim 0.3$), which shows that the movement information captured by variance map and flow provide a weak assistance to the scale type classification. Finally, combining variance map and flow, SGNet w/ Var (img + flow) further gains improvement on scale (87.57) and on movement (81.86) classification and achieves the best performance among all (without Kinetics pre-trained).

5.3 Ablation Studies

We conduct ablation studies on the following designs to verify their effectiveness: 1) subject map guidance, 2) subject map generation, and 3) joint training.

The Effects of Different Subject and Background Map Guidances. In the first block of Table 4, we take TSN model using image and flow with variance map (TSN w/ Var img+flow) as our baseline to test the effects of different subject map guidances. It takes a single-branch ResNet50 as backbone and two models for image and flow respectively, and fuses their scores at the end. We observe that using only subject or background information is inferior to the performance of using the whole image and flow, with $\sim 5/\sim 2$ drop on Acc_S/Acc_M.

Setting 4,5 in Table 4 are two branches setting with either subject or background guidance. In these experiments, we take a two-branch ResNet50 as backbone for image and flow model, one branch for subject/background and the other one for the whole image/flow. The output obtained from the first branch is concatenated with the output of second branch (+Subj and +Back) as guidance, and send to following networks. Generally, subject guidance achieves $0.4 \sim 0.8$

Table 5. Comparison of different subject map generation modules.

#	Settings	Acc_S (↑)	Acc_M (↑)
1	Base (TSN w/ Var img+flow)	84.15	77.25
2	SBS-ResNet50 [37]	83.82	76.36
3	R^3Net-ResNet18-fixed [13]	84.55	78.14
4	R^3Net-ResNet50-fixed [13]	85.10	79.56
5	R^3Net-ResNet18-finetuned [13]	86.15	81.24
6	R^3Net-ResNet50-finetuned [13]	**88.10**	**82.58**
7	Student generator w/ $\mathcal{L}_2 + \mathcal{L}_{cross}$	85.34	79.11
8	Student generator w/ $\mathcal{L}_2 + \mathcal{L}_{\text{adv}} + \mathcal{L}_{cross}$	**87.08**	**81.13**

performance gain on scale classification and background guidance outperforms subject guidance on movement prediction with $0.8 \sim 1.0$ better results.

Setting 6 (Base+Subj+Back) in Table 4 use a four-branch ResNet50. Two branches are for subject guidance, and the rest two are for background guidance. With more information, the performance drops a little since the subject and background information might be mutually exclusive to each other.

The Influence of Different Subject Map Generations. As discussed above, a subject map generation module is needed to guide the network prediction. This module has many alternatives. Table 5 shows the comparisons between our approaches and self-attention SBS (Saliency-Based Sampling Layer) [37], and fine-tuned/fixed models R^3Net-ResNet18/50 [13]. We take TSN model using image and flow with variance map (TSN w/ Var img+flow) as our baseline to test the influence of different subject map generations.

Self-attention generation method SBS (Saliency-Based Sampling Layer) [37] does not bring improvement compared with the baseline. The reason might be that self-attention is hard to learn from these weak labels, *i.e.* shot types. The pre-trained fixed networks (settings 3,4) bring gains to the performance, and the performance increases as the network becomes deeper. Moreover, when we fine tune these networks on our tasks (settings 5,6), the performance improves further with ~ 2 gains on both Acc_S and Acc_M.

Our light-weight subject map generation module is driven by two losses besides the classification cross entropy loss. The performance of using $\mathcal{L}_2$ loss (setting 7) is worse than fine-tuned R^3Net-ResNet50. With the help of both $\mathcal{L}_2$ loss and adversarial loss $\mathcal{L}_{adv}$ (setting 8), student generator is on par with R^3Net-ResNet50. However, compared with R^3Net-ResNet50, our light-weight subject map student generator has 99.8% fewer parameters and 89.4% fewer GFLOPs (shown in Table 6), which largely speeds up the training and inference processes.

Two-task Joint Training. To investigate the relationship between the scale and movement classification, we conduct the joint training experiments on these two tasks, as shown in Table 7. We take our full model SGNet w/ Var (img + flow)

Table 6. Parameters and computational complexity of different networks

Network Architecture	Params(M)	GFLOPs
Student generator	**0.04**	**2.38**
R^3Net-ResNet18 [13]	23.66	19.95
R^3Net-ResNet50 [13]	37.53	22.54

Table 7. Comparison of the performance of joint training sharing different modules.

Settings	Acc_S (↑)	Acc_M (↑)
Separate	87.57	81.86
Joint-training (Share SMG)	**88.12**	**82.19**
Joint-training (Share till res_1)	87.24	81.10
Joint-training (Share till res_4)	86.17	80.29

as the baseline. We testify the coupling of scale and movement type by sharing the same modules from bottom to top gradually. As lined out in the second row, sharing the subject map generation (SMG) module is helpful to the performance, where Acc_S and Acc_M raises 0.55 and 0.33 respectively. However, when we further share these two tasks classification till ResNet50's res_1 and res_4 modules, we observe that joint training is harmful to the performance when they share more branches. These prove that both scale and movement benefit from the subject guidance. The spatial layout learnt from scale and the camera motion learnt from the movement contribute complementaily to the subject map generation. While the subject guidance is shared by both tasks, the distinct goals of the two tasks still require task-specific designs in the later part to learn better representations.

5.4 Qualitative Results

In this section, we show the qualitative results of subject map generation and the variance map computation.

Subject Map. Fig. 4(a) compares our generated subject map with those generated by R^3Net-ResNet50 in fixed setting. Our generated subject map achieves much better generation result that are consistent with our human judgment. The first row in the figure is an over-the-shoulder static close-up shot, where both methods successfully predict the subject woman rather than the man with back head. But our method outputs much less noise. The second row is a full shot. Our method successfully detects both two people and does not include the background stone into the subject map. In the third and fourth row cases, R^3Net-ResNet50-fixed outputs blurred area around the contours of two people while our method obtains a sharp shape of the subject.

Variance Map. The variance map is important for predicting shot movement. We divide a shot video into 8 clips, plot the variance map for each movement type in the test set and average these variance maps in Fig. 4(b). The variance map is of size 8×8, and these gray scale blocks show the similarity among clips in the variance map. As noted from the plot, the variance map of the static shot is nearly an all-one matrix, meaning that there is no significant change between the eight clips. The near identity matrix shape of motion shots reveal that it has the least similarities between consecutive clips.

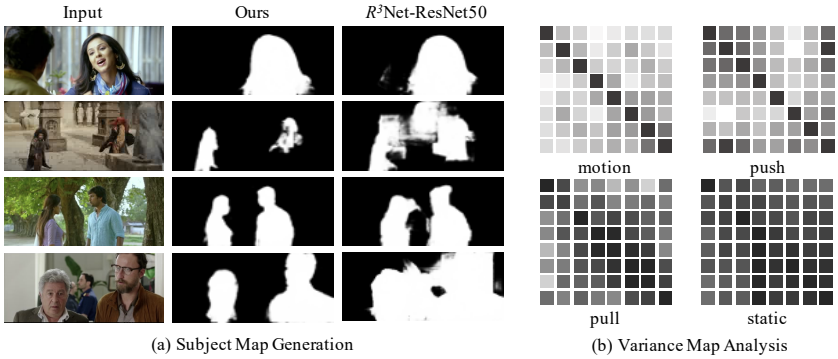

Fig. 4. (a) Comparison of our generated subject map and R^3Net-ResNet50-fixed generated map. (b) Variance map visualization of different movement types using gray-scale colors to indicate the similarity. The lighter the color, the lower the similarity score

6 Application

Shot type analysis has a wide range of potential applications. In this section, we illustrate one such application of realizing automatic video editing with the help of shot type classification.

Automatic Video Editing: Shot Type Changing. Video editors usually try out different shot types to convey emotions and stories, which consumes a lot of time and resources. In many cases, the change of shot type changes the semantic of a movie and effects audience's emotions, revealing the intent of the directors.

The model we propose in this paper could classify the shot type of any given video shot. Fig. 5 shows a shot clip[3] from the famous film *Titanic*, demonstrating how our model can be applied to changing shot scales to achieve a desired artistic expression. The original shot is a *medium shot*. Suppose we want to emphasize the role of speaker in this *dialogue scene*, we may want to use some *close-up shot* to emphasize the speaker. Firstly, we propose many cropping regions randomly depend on the position of the speaker and generate the corresponding candidate shots. Secondly, we use our shot classification model to classify these candidate shots and assign them with confidence scores. Note that the original shot is a single shot. We divide it into four shots depending on the active-speaker [38,15,35,57]. In Shot 1, Rose and Jack walk on the deck of the ship; Shot 2, Rose talks; Shot 3, they stop and Rose looks at Jack; Shot 4, Jack talks, as illustrated in Fig. 5. We change the style of the original shot by selecting parts from the divided four shots and replace them with the candidates with high scores and the desired scale types, as shown in Fig. 6. After these changes, the emotion of this clip turns to be more intense and the speaking cast

[3] [42] is adopted here to cut shots from the film.

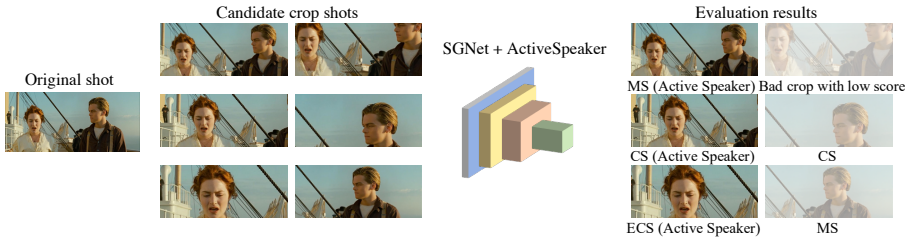

Fig. 5. A sample shot from a dialogue scene in *Titanic*, showing how we use our proposed shot type classification framework to aid the shot type changing

Fig. 6. Editing results on a medium shot clip from *Titanic* to emphasize the speaker and on a full shot clip *Parkour* from the DAVIS dataset [5] to emphasize the action

is being emphasized after changing from a middle shot to a close shot. One more result on DAVIS dataset [5] is also shown in Fig. 6. These results demonstrate the importance of shot type in videos, especially for their emotion and aesthetic analysis.[4]

7 Conclusion

In this work, we construct a large-scale dataset *MovieShots* for shot analysis, which containing $46K$ shots from $7K$ movie trailers with professionally annotated scale and movement attributes. We propose a Subject Guidance Network (SGNet) to capture the contextual information and the spatial and temporal configuration of a shot for our shot type classification task. Experiments show that this network is very effective and achieves better results than existing methods. All the studies in this paper together show that shot type analysis is a promising direction for edited video analysis which deserves further research efforts.

Acknowledgement: This work is partially supported by the SenseTime Collaborative Grant on Large-scale Multi-modality Analysis (CUHK Agreement No. TS1610626 & No. TS1712093), the General Research Fund (GRF) of Hong Kong (No. 14203518 & No. 14205719), and Innovation and Technology Support Program (ITSP) Tier 2, ITS/431/18F.

[4] More results and their corresponding videos are shown in the supplementary videos.

References

1. Bagheri-Khaligh, A., Raziperchikolaei, R., Moghaddam, M.E.: A new method for shot classification in soccer sports video based on svm classifier. In: 2012 IEEE Southwest Symposium on Image Analysis and Interpretation. pp. 109–112. IEEE (2012) 3
2. Belagiannis, V., Farshad, A., Galasso, F.: Adversarial network compression. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 0–0 (2018) 7
3. Benini, S., Canini, L., Leonardi, R.: Estimating cinematographic scene depth in movie shots. In: 2010 IEEE International Conference on Multimedia and Expo. pp. 855–860. IEEE (2010) 3
4. Bhattacharya, S., Mehran, R., Sukthankar, R., Shah, M.: Classification of cinematographic shots using lie algebra and its application to complex event recognition. IEEE Transactions on Multimedia **16**(3), 686–696 (April 2014) 4
5. Caelles, S., Pont-Tuset, J., Perazzi, F., Montes, A., Maninis, K.K., Van Gool, L.: The 2019 davis challenge on vos: Unsupervised multi-object segmentation. arXiv:1905.00737 (2019) 14
6. Canini, L., Benini, S., Leonardi, R.: Classifying cinematographic shot types. Multimedia tools and applications **62**(1), 51–73 (2013) 3, 4
7. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6299–6308 (2017) 3, 8, 9
8. Chao, Y.W., Vijayanarasimhan, S., Seybold, B., Ross, D.A., Deng, J., Sukthankar, R.: Rethinking the faster r-cnn architecture for temporal action localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1130–1139 (2018) 3
9. Cheng, M.M., Mitra, N.J., Huang, X., Torr, P.H., Hu, S.M.: Global contrast based salient region detection. IEEE Transactions on Pattern Analysis and Machine Intelligence **37**(3), 569–582 (2014) 7
10. Cheng, Y., Wang, D., Zhou, P., Zhang, T.: A survey of model compression and acceleration for deep neural networks. arXiv preprint arXiv:1710.09282 (2017) 7
11. Christoph, R., Pinz, F.A.: Spatiotemporal residual networks for video action recognition. Advances in Neural Information Processing Systems pp. 3468–3476 (2016) 6
12. Deng, J., Pan, Y., Yao, T., Zhou, W., Li, H., Mei, T.: Relation distillation networks for video object detection. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 7023–7032 (2019) 3
13. Deng, Z., Hu, X., Zhu, L., Xu, X., Qin, J., Han, G., Heng, P.A.: R3net: Recurrent residual refinement network for saliency detection. In: Proceedings of the 27th International Joint Conference on Artificial Intelligence. pp. 684–690. AAAI Press (2018) 7, 11, 12
14. Duan, L.Y., Xu, M., Tian, Q., Xu, C.S., Jin, J.S.: A unified framework for semantic shot classification in sports video. IEEE Transactions on multimedia **7**(6), 1066–1083 (2005) 3, 4
15. Ebrahimi, J., Rao, A., Lowd, D., Dou, D.: Hotflip: White-box adversarial examples for text classification. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). pp. 31–36 (2018) 13

16. Ekin, A., Tekalp, A.M.: Shot type classification by dominant color for sports video segmentation and summarization. In: 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). vol. 3, pp. III–173. IEEE (2003) 3
17. Fabian Caba Heilbron, Victor Escorcia, B.G., Niebles, J.C.: Activitynet: A large-scale video benchmark for human activity understanding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 961–970 (2015) 2, 3
18. Feichtenhofer, C., Fan, H., Malik, J., He, K.: Slowfast networks for video recognition. arXiv preprint arXiv:1812.03982 (2018) 3
19. Giannetti, L.D., Leach, J.: Understanding movies, vol. 1. Prentice Hall Upper Saddle River, New Jersey (1999) 2, 4
20. Goldblum, M., Fowl, L., Feizi, S., Goldstein, T.: Adversarially robust distillation. Thirty-Fourth AAAI Conference on Artificial Intelligenc (2020) 7
21. Guo, C., Fan, B., Gu, J., Zhang, Q., Xiang, S., Prinet, V., Pan, C.: Progressive sparse local attention for video object detection. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3909–3918 (2019) 3
22. Hasan, M.A., Xu, M., He, X., Xu, C.: 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 (2014) 3, 8, 9
23. 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) 6, 8
24. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. In: NIPS Deep Learning and Representation Learning Workshop (2015) 7
25. Hou, Q., Cheng, M.M., Hu, X., Borji, A., Tu, Z., Torr, P.: Deeply supervised salient object detection with short connections. IEEE TPAMI **41**(4), 815–828 (2019) 7
26. Huang, Q., Xiong, Y., Rao, A., Wang, J., Lin, D.: Movienet: A holistic dataset for movie understanding. In: The European Conference on Computer Vision (ECCV) (2020) 4
27. Jiang, H., Zhang, M.: Tennis video shot classification based on support vector machine. In: 2011 IEEE International Conference on Computer Science and Automation Engineering. vol. 2, pp. 757–761. IEEE (2011) 3
28. Kowdle, A., Chen, T.: Learning to segment a video to clips based on scene and camera motion. In: European Conference on Computer Vision. pp. 272–286. Springer (2012) 4
29. Li, L., Zhang, X., Hu, W., Li, W., Zhu, P.: Soccer video shot classification based on color characterization using dominant sets clustering. In: Pacific-Rim Conference on Multimedia. pp. 923–929. Springer (2009) 3, 8, 9
30. Li, L., Zhang, X., Hu, W., Li, W., Zhu, P.: Soccer video shot classification based on color characterization using dominant sets clustering. In: Muneesawang, P., Wu, F., Kumazawa, I., Roeksabutr, A., Liao, M., Tang, X. (eds.) Advances in Multimedia Information Processing - PCM 2009. pp. 923–929 (2009) 4
31. Li, S., He, F., Du, B., Zhang, L., Xu, Y., Tao, D.: Fast spatio-temporal residual network for video super-resolution. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019) 4
32. Li, X., Zhao, L., Wei, L., Yang, M.H., Wu, F., Zhuang, Y., Ling, H., Wang, J.: Deepsaliency: Multi-task deep neural network model for salient object detection. IEEE Transactions on Image Processing **25**(8), 3919–3930 (2016) 7

33. Monfort, M., Andonian, A., Zhou, B., Ramakrishnan, K., Bargal, S.A., Yan, Y., Brown, L., Fan, Q., Gutfreund, D., Vondrick, C., et al.: Moments in time dataset: one million videos for event understanding. IEEE transactions on pattern analysis and machine intelligence (2019) 2, 3
34. Prasertsakul, P., Kondo, T., Iida, H.: Video shot classification using 2d motion histogram. In: 2017 14th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON). pp. 202–205. IEEE (2017) 3, 8, 9
35. Rao, A., Lau, F.: Automatic music accompanist. arXiv preprint arXiv:1803.09033 (2018) 13
36. Rao, A., Xu, L., Xiong, Y., Xu, G., Huang, Q., Zhou, B., Lin, D.: A local-to-global approach to multi-modal movie scene segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10146–10155 (2020) 5
37. Recasens, A., Kellnhofer, P., Stent, S., Matusik, W., Torralba, A.: Learning to zoom: a saliency-based sampling layer for neural networks. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 51–66 (2018) 11
38. 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) 13
39. 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) 8
40. Savardi, M., Signoroni, A., Migliorati, P., Benini, S.: Shot scale analysis in movies by convolutional neural networks. In: 2018 25th IEEE International Conference on Image Processing (ICIP). pp. 2620–2624. IEEE (2018) 4
41. Shou, Z., Lin, X., Kalantidis, Y., Sevilla-Lara, L., Rohrbach, M., Chang, S.F., Yan, Z.: Dmc-net: Generating discriminative motion cues for fast compressed video action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1268–1277 (2019) 7
42. 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) 5, 13
43. 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, 4
44. Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., Van Gool, L.: Temporal segment networks: Towards good practices for deep action recognition. In: European conference on computer vision. pp. 20–36. Springer (2016) 2, 3, 8, 9
45. Wang, X., Zhang, R., Sun, Y., Qi, J.: Kdgan: Knowledge distillation with generative adversarial networks. In: Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (eds.) Advances in Neural Information Processing Systems 31, pp. 775–786. Curran Associates, Inc. (2018) 7
46. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018) 3
47. Wang, Y., Xu, C., Xu, C., Tao, D.: Adversarial learning of portable student networks. In: Thirty-Second AAAI Conference on Artificial Intelligence (2018) 7

48. Wikipedia: As seen through a telescope, https://en.wikipedia.org/, accessed: 2020-02-18 1
49. Xia, J., Rao, A., Xu, L., Huang, Q., Wen, J., Lin, D.: Online multi-modal person search in videos. In: The European Conference on Computer Vision (ECCV) (2020) 4
50. Xiao-Feng Tong, Qing-Shan Liu, Han-Qing Lu, Hong-Liang Jin: Shot classification in sports video. In: Proceedings 7th International Conference on Signal Processing, 2004. Proceedings. ICSP '04. 2004. vol. 2, pp. 1364–1367 vol.2 (Aug 2004) 3
51. Xiong, Y., Huang, Q., Guo, L., Zhou, H., Zhou, B., Lin, D.: A graph-based framework to bridge movies and synopses. In: The IEEE International Conference on Computer Vision (ICCV) (2019) 4
52. Xu, G., Liu, Z., Li, X., Loy, C.C.: Knowledge distillation meets self-supervision. In: European Conference on Computer Vision (ECCV) (2020) 7
53. Xu, H., Das, A., Saenko, K.: R-c3d: Region convolutional 3d network for temporal activity detection. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 5783–5792 (2017) 3
54. Xu, K., Wen, L., Li, G., Bo, L., Huang, Q.: Spatiotemporal cnn for video object segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1379–1388 (2019) 4
55. 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, 4
56. Xu, R., Li, X., Zhou, B., Loy, C.C.: Deep flow-guided video inpainting. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019) 4
57. Xu, X., Dai, B., Lin, D.: Recursive visual sound separation using minus-plus net. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 882–891 (2019) 13
58. Yang, J., Zheng, W.S., Yang, Q., Chen, Y.C., Tian, Q.: Spatial-temporal graph convolutional network for video-based person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3289–3299 (2020) 4
59. Yang, Y., Lin, S., Zhang, Y., Tang, S.: Statistical framework for shot segmentation and classification in sports video. In: Computer Vision – ACCV 2007. pp. 106–115. Springer Berlin Heidelberg (2007) 4
60. Yuan, L., Wang, T., Zhang, X., Tay, F.E., Jie, Z., Liu, W., Feng, J.: Central similarity quantization for efficient image and video retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3083–3092 (2020) 4
61. Zeng, X., Liao, R., Gu, L., Xiong, Y., Fidler, S., Urtasun, R.: Dmm-net: Differentiable mask-matching network for video object segmentation. arXiv preprint arXiv:1909.12471 (2019) 4
62. Zhang, H., Liu, D., Xiong, Z.: Two-stream oriented video super-resolution for action recognition. arXiv preprint arXiv:1903.05577 (2019) 4
63. Zhao, Y., Xiong, Y., Wang, L., Wu, Z., Tang, X., Lin, D.: Temporal action detection with structured segment networks. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2914–2923 (2017) 3
64. Zhu, W., Liang, S., Wei, Y., Sun, J.: Saliency optimization from robust background detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2814–2821 (2014) 7