Proceedings of the 11th World Congress on Electrical Engineering and Computer Systems and Sciences (EECSS'25)
Paris, France - August, 2025
Paper No. MVML 131
DOI: 10.11159/mvml25.131

Shot Scale beyond Boundaries: A Regression-Based Approach

Sebastian Maisel [1], Samuel Wunderlich[1], Anna Kleinhans [1], Andreas Koch [1]
[1] Institute for Applied Artificial Intelligence, Stuttgart Media University
Nobelstrasse 10, 70569 Stuttgart, Germany
maisel@hdm-stuttgart.de; wunderlich@hdm-stuttgart.de; kleinhans@hdm-stuttgart.de; kocha@hdm-stuttgart.de

Abstract **-** Video or movie analysis often involves examining various audio-visual elements. One significant visual aspect that has attracted research attention is shot scale, which refers to the distance between the camera and the subject in a shot. The impact of shot scale on narrative and viewer emotion has been well researched. Additionally, methods for automatic shot classification have been developed. However, previous studies primarily focused on categorizing shot scales into discrete classes, which can lead to misclassifications. In this work, we propose a novel regression-based approach for shot scale prediction. We show that our method outperforms prior shot classification methods in cross-dataset generalization.

Keywords**:** regression, shot-size, shot-scale, camera, film, video, video analysis

1. Introduction

In recent years, the classification of visual parameters in films and videos has emerged as an important research area for both film studies and multimedia applications. Among these parameters, the shot scale, which refers to the extent of a subject's visibility in a shot, has proven to be a significant factor in movie analysis. Prior studies have, for instance, highlighted the influence of shot scale on viewers' mental state attribution [1] and its impact on narrative engagement ratings [2] which has led to the emergence of the task of automatic shot scale classification. The automation of shot scale classification not only provides a quantitative basis for understanding cinematographic style [3] [4] but can also be utilized in tasks such as film recommendations, automatic attribution of movie's authorship [5], or stylistic analysis. The characterization of individual shots according to their spatial framing - ranging from close-ups to wide shots - offers insights into narrative structure and visual rhetoric, which are central to both the creative, stylistic [4] and analytical disciplines of film and video production.

By employing advanced automatic classification methods for shot scale prediction, efficient, large-scale analysis of vast movie corpora [6] is facilitated, thereby enabling its applications in domains such as video editing [7].

For the purpose of this paper, shot scales are categorized as follows:

Extreme Close-Up (ECU) Shows extreme detail (e.g., an eye, a finger, or part of the face).

Close-Up (CU) Frames a face or key object, isolating it from the background.

Medium Close-Up (MCU) Chest or shoulders up; balances face and some background.

Medium Shot (M) Waist up; subject clearly visible with moderate surroundings.

Medium Wide Shot (MW) Knees up; more environment while keeping focus on the subject.

Wide Shot (W) Full body visible; subject in relation to environment.

Extreme Wide Shot (EW) Subject small in frame, emphasizing setting and surroundings.

In this paper, we propose a novel regression-based approach to shot scale prediction. Our work aims to assess whether the regressive prediction of shot scales, especially when combined with the fusion of face and person bounding boxes, can achieve superior cross-dataset generalization. Such improvements are expected to benefit both theoretical investigations into film language and practical applications in media technology. Building on previous work of Argaw et al. [7] and Savardi et al. [6], we implement a regression model for shot scale prediction. After filtering out noisy labels from the datasets, we assess

the cross-dataset performance of our approach and compare it against the models proposed by Argaw et al. [7] and Savardi et al. [6].

Our main contributions can be summarized as follows:

- Introduction of a novel regression-based approach to shot scale prediction.
- Experiments on the filtered AVE and CineScale datasets demonstrate the superior cross-dataset generalization capability of our model.
- We enable more fine-grained and adjustable analysis of shot scales.

2. Related Work

Although research on shot scale prediction remains relatively limited, several notable approaches have emerged, some of which have informed and inspired the present work.

Shot scale classification Several prior works use CNN-based approaches for shot scale classification. Svanera et al. [5] train and evaluate multiple CNNs for shot scale classification for the automatic attribution of movie authorship. Savardi et al. [6] propose a CNN-based method for classifying shot scales in films for the three shot scales Close-Up, Medium and Wide, using a fine-tuned VGG-16. Vacchetti and Cerquitelli [8] address shot classification across eight shot scales using an ensemble approach that integrates predictions from three fine-tuned VGG-16 models. Their results reveal that misclassifications frequently occur due to samples displaying characteristics of similar classes, possibly indicating the need for a more fine-grained, continuous prediction of shot scales. In comparison, Bak and Park [9] propose an approach to shot scale classification that adds semantic information of the image by preprocessing with semantic segmentation and ResNet-50, achieving an average accuracy of 94.9%. Tsingalis et al. (2012) [10] propose an SVM-based method for classifying movie shots into seven shot scale categories (Extreme Close-Up to Extreme-Wide Shots) using two geometric features derived from facial bounding boxes: the height and width ratios of the actor's face to the video frame. Their approach achieves a high shot-level classification accuracy on a manually annotated dataset. Although effective, the method requires visible faces and their manual boundary annotations, limiting its use to face-present shots. Based on the work of [10], our work relies on automatically generated face- and person-bounding-boxes based on the work of YOLO (v8) by Redmon et al. [11]. Newlin and Arivazhagan [12] compare a quantitive method by which shot scale classification is performed based on the face percentage in the shot with CNN based approaches for video surveillance applications. Other approaches focus on the detection of shot types, which are defined by what is depicted in the shot, as opposed to shot scales [13].

While some of the studies concerned with shot scale classification show promising results, none have adopted a more fine-grained regression-based approach. We aim to close this gap with our work.

Datasets Savardi et al. [14] released the CineScale dataset which includes shot scale annotations for 124 movies. The shot scales are divided into nine categories (Extreme Close-Up, Close-Up, Medium Close-Up, Medium Shot, Medium Long Shot, Long Shot, Extreme Long Shot, Foreground Shot and Insert Shots). The CineScale dataset is used for training and evaluating our regression-based approach. Argaw et al. [7] present the AVE dataset, a benchmark for AI-assisted video editing. It includes around 196,000 annotated shots, with shot scale predicted via a multi-task audio-visual model. AVE supports detailed analysis of cinematographic patterns, therefore parts of the dataset are used for benchmarking different models in this work. To evaluate reliable boundaries between shot scales, we adopt the classification schemes proposed by [15] and [16].

3. Method

The primary focus of this work lies in the comparative analysis of our regressive model against classification-based approaches exemplary employed by AVE and CineScale. In contrast to the fixed-class prediction of these models, our method aims to predict shot scales on a continuous scale, enabling a more flexible downstream use.

In order to enable our model to produce reliable shot scale predictions, we analyze and filter both the CineScale and AVE data set using YOLO models for face and person detection.

3.1. Dataset Statistics

Based on the definition of shots scales in section 1 and the largest face bounding box height of subjects in each frame, we observe significant class overlap and some label noise in the CineScale dataset. In contrast, the AVE dataset offers clearer class boundaries, though it treats all medium shot scales as a single category and exhibits a low number of samples in extreme classes (see Table 1). Figure 1 illustrates the disparities in shot scale label to subject size between the two datasets.

As shown in Table 1, there exists high class imbalance in both datasets, particularly in the AVE dataset, where both extremes contain very few samples. Additionally, label noise is present across all classes, due to the labeling process of both datasets. Since labeling was performed at the shot level in the case of AVE, individual frames within these shots may not consistently reflect the assigned labels. While the whole shot may fall into one of the shot scale classes, individual frames can reflect different shot scales especially if there is a lot of movement of subjects through the scene or camera movement. Training a network on these labels also ignores that a model which only takes one frame as input does not receive enough context information to reliably discern the subject of a shot.

Table 1: Data distribution across classes in the two data sets. Classes are imbalanced in both sets, with medium shots forming the vast majority of data. Extreme shots are rare in comparison.

	ECU	CU	MCU	M	MW	W	EW
CineScale	3,769	89,640	263,841	80,747	95,461	54,097	8,076
AVE	354	24,667	136,271			19,355	938

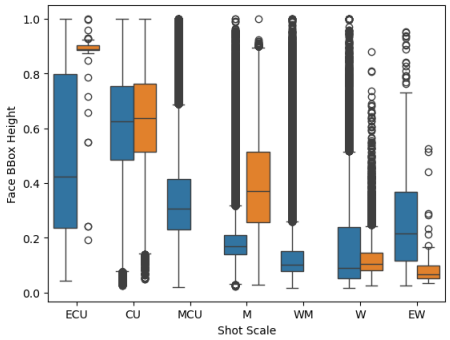

Figure 1: Face bounding box height relative to height of frames for each shot scale in CineScale (blue) and AVE dataset (orange). Both datasets have overlaps in their class labels and differing definitions of shot scales.

3.2. Data Preparation

Based on the distribution of face bounding box heights, we manually examine overlaps between adjacent shot scale classes. Using shot scale definitions, we derive upper and lower bounds for each class based on sampled data, thereby reducing class overlap and minimizing mislabeled instances. To avoid discarding valid samples without visible people, this filtering is applied only to frames in which one or more faces and persons are detected. This process assumes that faces in cinematic shots are typically vertically aligned, which generally holds true in the data. Only a small number of non-standard cases are affected by this assumption.

No filtering is performed on the extreme close-up class, as these shots often feature only partial facial regions (e.g., eyes), making bounding box sizes highly unreliable. For close-up filtering, we further consider the number of persons present in the frame, as this helps distinguish between tightly framed shots and wider compositions featuring multiple subjects. To ensure consistency with the shot scale definitions, we limit our analysis to classes that directly correspond to established shot scales. As such, we exclude non-standard categories such as ’other’ in the AVE dataset and ’foreground’, ’insert’, or ’unavailable’ classes in the CineScale dataset.

We filter 220k samples from the CineScale dataset, leaving us with a total of 527,380 samples. On these we create a new 70/30 split, assigning the provided movies either to test or training set. For the AVE data, we only remove 13k samples, for a total of 170,051 samples. The distribution of face bounding box heights before and after filtering is shown in Figure 2.

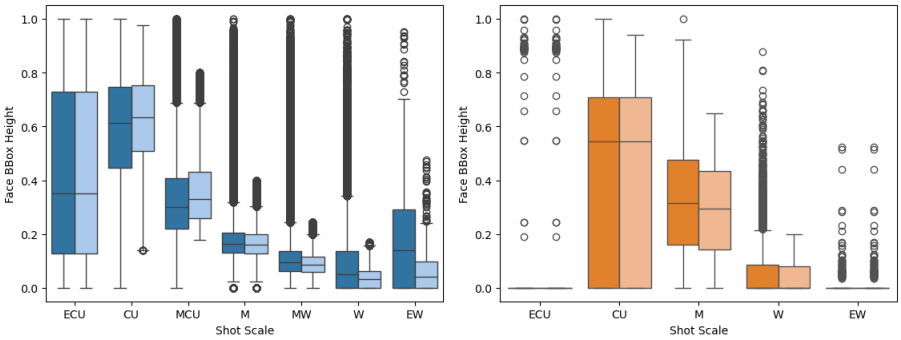

Figure 2: Bounding box heights in CineScale data (left) and AVE (right); before (blue, orange) and after (light blue, light orange) filtering out shots not belonging in the respective shot scale classes. Class overlap is reduced after filtering.

3.3. Model Details

We modify and fine-tune a VGG-16 convolutional network [17] on the CineScale dataset, using images of size 224x224px as input, freezing the weights of all filter layers except the last two convolutional blocks. Additionally, the network receives the bounding boxes of persons and faces as binary image of size 112x112px to an auxiliary convolutional head, with flattened output size 32. Both outputs are concatenated and passed into a modified VGG-16 classifier, which is trained from scratch. In contrast to previous works, we formulate the training objective as regression problem and round the networks output to the nearest integer for classification. The output layer consists of a single neuron instead of a number of neurons corresponding to the number of classes. The output is limited by a sigmoid function, scaled by the total number of classes in the dataset to produce a continuous shot scale estimate $x \in (0,6)$, corresponding to the shot scales in the CineScale dataset. In total our model consists of about 70M trainable parameters.

We employ a modified mean squared error (MSE) loss, where the indicator function $\mathbf{1}_y$ excludes samples correctly assigned to their class from contributing to the loss (see Equation 1). Due to the high class imbalance in the data, losses are weighted by w_y according to their respective class y. Optimization is performed using AdamW, with a learning rate of 5e−4 and a learning rate scheduler that reduces the learning rate upon plateauing (no improvement in validation loss for 50 epochs). The model is trained for 500 epochs, using the same data augmentation strategy as proposed in [6].

$$L(x,y) = w_{yn}(x_n - y_n)^2 \mathbf{1}_{yn}(\lfloor x_n \rceil) \quad (1)$$

4. Experiments

Our experiments aim to evaluate the generalization capability of the regression-based shot scale estimation model in comparison to classification-based models of previous works. We train our model on the CineScale dataset, as the AVE dataset lacks sufficient representation of extreme shot scale classes. To assess generalization, we test all models both on their source datasets and cross-dataset, applying consistent class mappings where necessary.

We further evaluate the official CineScale model trained on CineScale, and a re-implementation of the AVE model based on the original paper on the AVE data. While reproducing CineScale's results, we observe a lower accuracy on our test split than reported on the test split of the original work. However, the high accuracy of the model on our reconstructed training split (close to 90%) suggests that our split preserves a relatively consistent separation between training and test data.

We test whether the regression-based model informed by face and person bounding boxes can generalize better across both datasets and to unseen data, compared to models that rely on discrete class predictions.

4.1. Shot Scale Mappings

As the CineScale and AVE datasets define shot scale classes with differing granularity, mappings are necessary when evaluating models on datasets they were not trained on. These mappings are derived based on face and person bounding boxes distribution for each class (e.g. Figure 1) and follow the shot scale definitions outlined in Section 1.

Table 2: shot scale class mapping based on CineScale definitions for different numbers of output classes

	ECU	**CU**	**MCU**	**M**	**MW**	**W**	**EW**
7 classes	ECU	CU	MCU	M	MW	W	EW
5 classes	ECU	CU		M	W		EW
3 classes	CU			M		W	

4.2. Model Evaluation

Figure 3 shows the normalized confusion matrix for our regression model evaluated on the CineScale dataset. The model predicts the most frequent classes with higher accuracy, while performance drops for extreme classes such as Extreme Close-Up (ECU) and Extreme Wide (EW). This likely stems from a combination of low sample counts and high variance in face bounding box sizes, especially among Extreme Close-Ups that were not fully cleaned by our data filtering. Overall, misclassifications predominantly occur between semantically adjacent classes, indicating that the model captures the ordinal structure of shot scales. Figure 4 supports this, as the model is capable of outputting continuous values between class labels. Mapped onto three classes, the macro-averaged F1 score is 0.67 (Table 3), slightly below that of the original CineScale model but still indicative of good performance given the continuous nature of the task and the inherent class imbalance.

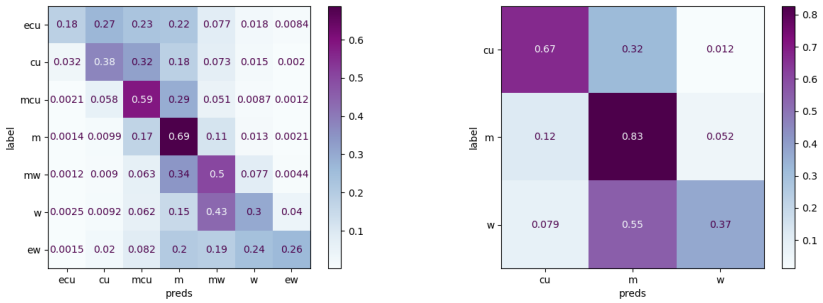

Figure 3: confusion matrix of our model on CineScale data, left: on 7 classes, right: mapped to 3 classes.

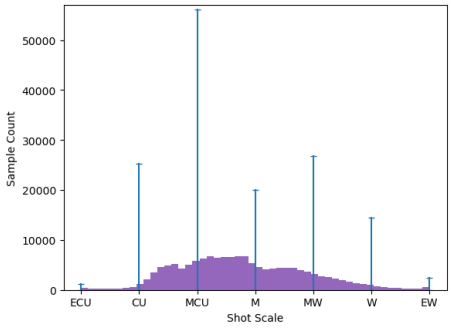

Figure 4: Distribution of raw regression model outputs on our CineScale test set before assigning them to a class. The model can predict continuous values between classes.

4.3. Transfer Results

To evaluate cross-dataset generalization, we evaluate our model - trained on the CineScale dataset - on the AVE dataset using the same shot scale mapping applied in Figure 5. The confusion matrix in Figure 6 reveals similar prediction behavior to that observed on the CineScale dataset, with most misidentifications occurring between adjacent classes. As baseline comparison of the AVE model on their data, refer to Figure 7. Notably, many Extreme-Wide Shots are predicted to belong to the wide category, due to CineScale having many samples with similar face size in the Extreme-Wide category. As with the CineScale evaluation, Extreme Close-up predictions remain less reliable due to uncleaned data in this class. However, as shown in Table 3, our model outperforms both the CineScale and the AVE models in terms of F1-score, achieving improvements of 6% and 32% respectively. Due to dataset imbalances, accuracy is considered a less informative metric in this context.

Table 3: comparison of models in terms of accuracy and F1-score (weighted avg) on the AVE and CineScale datasets

	AVE		**CineScale**	
	Accuracy	**F1-Score**	**Accuracy**	**F1-Score**
AVE (ResNet101)	0.73	0.62	0.74	0.59
CineScale (DenseNet)	0.32	0.33	**0.76**	**0.70**
Ours (VGG16)	**0.77**	**0.65**	0.74	0.67

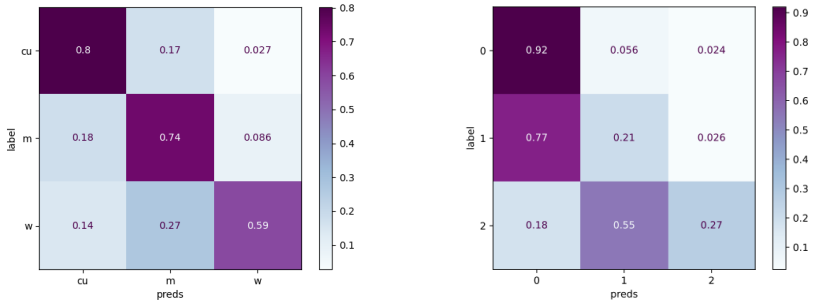

Figure 5: Confusion matrix of CineScale model, left: on CineScale dataset, right: on AVE dataset mapped to 3 classes.

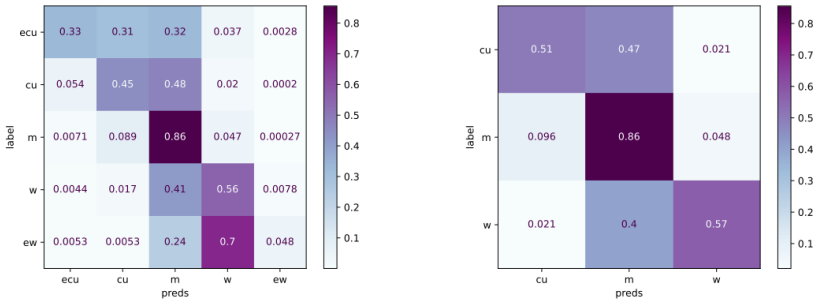

Figure 6: Confusion matrix of our regression model on AVE dataset (trained on CineScale)

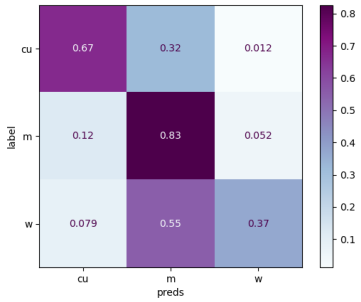

Figure 7: confusion matrix of AVE model on AVE dataset (3 classes)

4.4. Discussion

Our results demonstrate that the regression model generalizes well across datasets. The majority of misclassifications occur between semantically adjacent shot scales, which is consistent with the model's implicit encoding of ordinal relationships. reliably predicting the extreme classes correctly remains a challenge, as sample counts for these classes are typically very low in films and they still possess some label noise. The samples in Figure 8 also demonstrate that our model is capable of producing visually plausible labels, even when the actual labels are not even of an adjacent shot scale. Compared to prior work, our model achieves competitive F1 scores, and benefits from a regression-based output space, allowing post hoc remapping to arbitrary class definitions without retraining. Classification models inherently lack this flexibility due to their reliance on a fixed number of discrete output classes defined at training time. This is particularly valuable in practical applications, where shot scale interpretation may vary by context or users may want to differentiate between specific scales. This also enables more fine-grained comparisons of shot sizes.

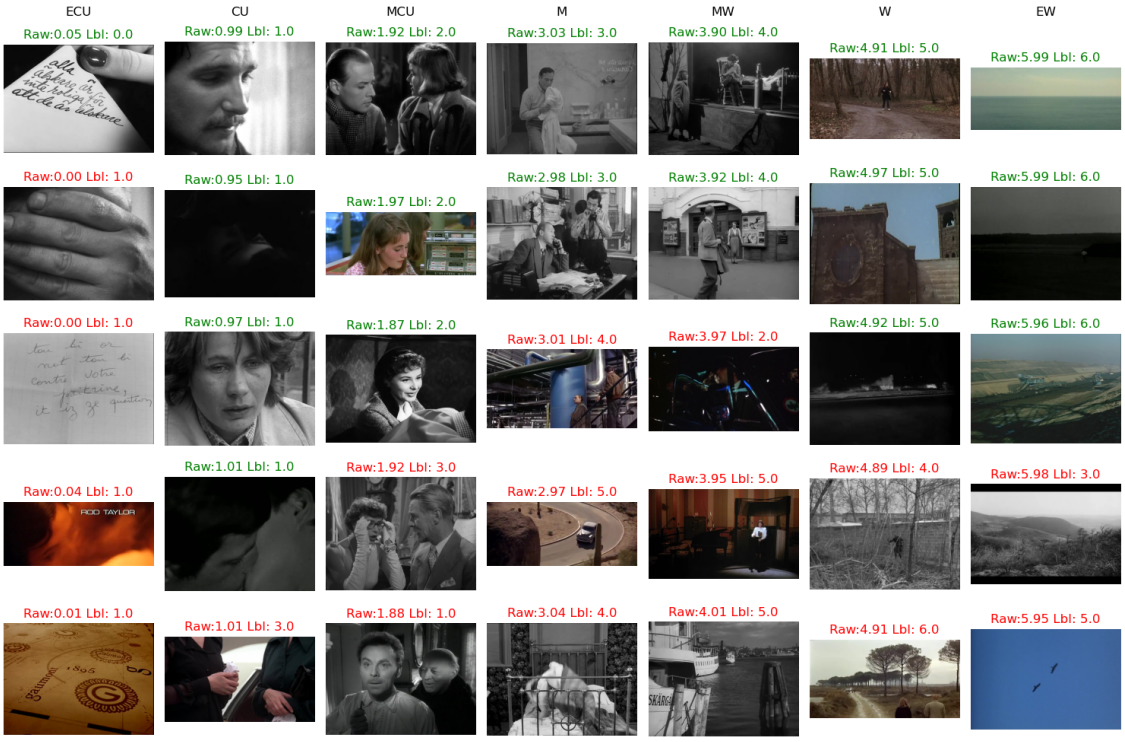

Figure 8: Outputs of the regression model on CineScale test data. We sample random images with raw outputs close to class labels. Green represents the prediction matching the label, red represents a mismatch. We observe a significant portion of predictions being correct despite not matching the originally assigned label

5. Conclusion

We present a regression-based deep learning model for shot scale prediction that is capable of expressing the ordinal and continuous nature of cinematic shot scales. By evaluating the model across two datasets, we demonstrate strong generalization performance and competitive F1 scores in the presence of class imbalance and label noise. Additionally, the continuous output enables flexible post-processing and adaptation to different class schemes without the need for retraining. These properties make the model particularly suitable for real-world applications in video analysis and film studies. Future work may explore improved predictions of extreme categories, and integration into larger pipelines for automated cinematic analysis.

References

[1] K. E. Bálint, J. N. Blessing, and B. Rooney, "Shot scale matters: The effect of close-up frequency on mental state attribution in film viewers," *Poetics*, vol. 83, 2020.

[2] S. Benini, M. Savardi, K. Bálint, A. B. Kovács and A. Signoroni, "On the influence of shot scale on film mood and narrative engagement in film viewers," *IEEE Transactions on Affective Computing*, vol .13, no. 2, pp.592-603, 2022.

[3] S. Benini, M. Svanera, N. Adami, R. Leonardi, and A. Kovács, "Shot scale distribution in art films," *Multimedia Tools and Applications*, vol. 75, no. 23, 2016.

[4] A. B. Kovács, "Shot scale distribution: An authorial fingerprint or a cognitive pattern?," *Projections*, vol. 8, no. 2, pp. 50-70, 2014.

[5] M. Svanera, M. Savardi, A. Signoroni, A. B. Kovács, and S. Benini, "Who is the film's director? Authorship recognition based on shot features," *IEEE MultiMedia*, vol. 26, no. 4, pp. 43–54, 2019.

[6] M. Savardi, A. Signoroni, P. Migliorati, and S. Benini, "Shot scale analysis in movies by convolutional neural networks," in *IEEE International Conference on Image Processing (ICIP)*, Athens, Greece, 2018, pp. 2620–2624.

[7] D. M. Argaw, F. C. Heilbron, J.-Y. Lee, M. Woodson, and I. S. Kweon, "The anatomy of video editing: A dataset and benchmark suite for ai-assisted video editing," in *E Computer Vision – ECCV 2022: 17th European Conference*. Tel Aviv, Israel: Springer, 2022, pp. 201–218.

[8] B. Vacchetti and T. Cerquitelli, "Cinematographic shot classification with deep ensemble learning,"*Electronics*, vol. 11, no. 10, 2022. [Online]. Available: https://www.mdpi.com/2079-9292/11/10/1570

[9] H.-Y. Bak and S.-B. Park, "Comparative study of movie shot classification based on semantic segmentation," *Applied Sciences*, vol. 10, no. 10, 2020. [Online]. Available: https://www.mdpi.com/2076-3417/10/10/3390

[10] N. Vretos, I. Tsingalis, and I. Pitas, "Svm-based shot type classification of movie content," in *IEEE Mediterranean Electrotechnical Conference*, vol. 3, 2012.

[11] J. Redmon, S. K. Divvala, R. B. Girshick and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 779-788.

[12] N. S. R and A. S, "Shot classification for human behavioural analysis in video surveillance applications," *ELCVIA Electronic Letters on Computer Vision and Image Analysis*, vol. 22, no. 2, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:265179030

[13] M. Svanera, S. Benini, N. Adami, R. Leonardi, and A. B. Kovács, "Over-the-shoulder shot detection in art films," in *13th International Workshop on Content-Based Multimedia Indexing (CBMI)*, pp. 1–6, 2015.

[14] M. Savardi, A. B. Kovács, A. Signoroni, and S. Benini, "Cinescale: A dataset of cinematic shot scale in movies," *Data in Brief*, vol. 36, 2021.

[15] D. Arijon, *Grammar of the film language.* Los Angeles: Silman-James Press, 1991.

[16] H. Zettl, *Sight, sound, motion: applied media aesthetics.* Boston, MA: Wadsworth Cengage Learning, 2011.

[17] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *3rd International Conference on Learning Representations (ICLR 2015),* 2015, pp.1-14.