Recognition of Camera Angle and Camera Level in Movies from Single Frames

MATTIA SAVARDI, Department of Medical and Surgical Specialties, Radiological Sciences, and Public Health; University of Brescia, Italy
ANDRÁS BÁLINT KOVÁCS, Film Department; ELTE University, Hungary
ALBERTO SIGNORONI, Department of Medical and Surgical Specialties, Radiological Sciences, and Public Health; University of Brescia, Italy
SERGIO BENINI, Department of Information Engineering; University of Brescia, Italy

The position and orientation of the camera in relation to the subject(s) in a movie scene, namely *camera "level"* and *camera "angle"*, are essential features in the film-making process due to their influence on the viewer's perception of the scene. In this paper, we propose the use of Convolutional Neural Networks (CNNs) for the automatic recognition of camera angles (categorized into five classes: Overhead, High, Neutral, Low, and Dutch) and camera levels (categorized into Aerial, Eye, Shoulder, Hip, Knee, and Ground) in movie frames. Our approach demonstrates remarkable effectiveness even when frames do not prominently feature the human figure. The training, validation, and test datasets are composed of frames sampled from an unprecedented variety of movie shots, freely available images, and labeled frames from cinematographic websites, for a total of over 24,000 images. Classification results for both camera angle and level achieve a weighted average precision and recall above 95%. To foster further research in domains such as movie stylistic analysis, video recommendation, and media psychology, we provide the developed models, annotation tool, and frame data through our project page at https://cinescale.github.io/.

CCS Concepts: • **Computing methodologies** → *Computer vision*; • **Applied computing** → *Media arts*.

ACM Reference Format:
Mattia Savardi, András Bálint Kovács, Alberto Signoroni, and Sergio Benini. 2023. Recognition of Camera Angle and Camera Level in Movies from Single Frames. 1, 1 (May 2023), 11 pages. https://doi.org/10.1145/nnnnnn.nnnnnnn

1 INTRODUCTION

Camera angle and level are fundamental aspects to consider during movie production, alongside other critical elements such as shot scale, camera movement, and shot editing, in order to enhance the storytelling experience for viewers [23, 3].

The importance of *camera angle* lies in its ability to establish a power dynamic between characters, as prior research has demonstrated that low-angle shots (where the viewer is forced to look up at the characters) can convey dominance, strength, and aggression, while high-angle shots (where the viewer looks down at the characters) can imply weakness and vulnerability [9]. Camera angle can

Authors' addresses: Mattia Savardi, Department of Medical and Surgical Specialties, Radiological Sciences, and Public Health; University of Brescia, Brescia, Italy, mattia.savardi@unibs.it; András Bálint Kovács, Film Department; ELTE University, Budapest, Hungary; Alberto Signoroni, Department of Medical and Surgical Specialties, Radiological Sciences, and Public Health; University of Brescia, Brescia, Italy; Sergio Benini, Department of Information Engineering; University of Brescia, Brescia, Italy.

XXXX-XXXX/2023/5-ART $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

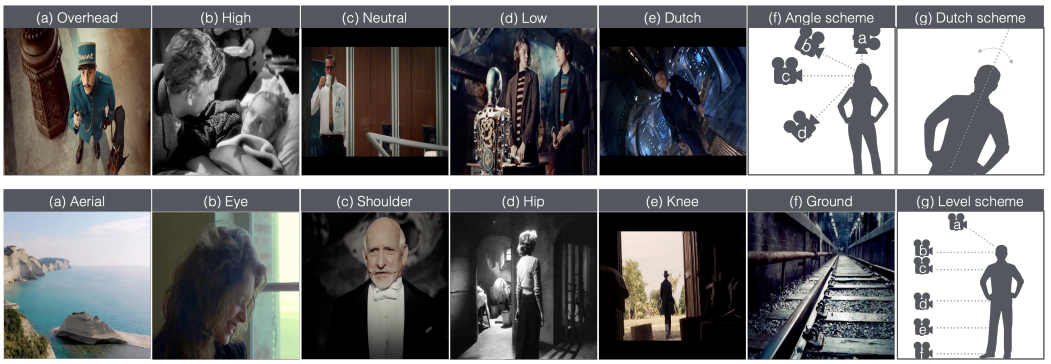

Fig. 1. First row: examples of different camera angle classes, and a reference scheme. Second row: examples of different camera level classes, and a reference scheme.

also influence empathic engagement by affecting the audience's attitudes towards and evaluations of characters [10], products [14] and the credibility of a speaker in promotion videos [24, 13]. Figure 1 (first row) provides a reference scheme and examples of different camera angles.

Camera level, on the other hand, is a tool for controlling storytelling by determining the viewer's perspective on the scene. Eye-level shots are considered neutral and are often used to show natural conversations between characters, while knee- and ground-level takes can be used to feature characters walking without revealing their face, inducing viewers to imagine what is happening at higher levels. Aerial-level shots, in turn, provide viewers with a reference in space, time, or reality. Using different camera levels can affect the viewer's empathy, with eye-level shots promoting perceived similarity to the subject of the camera [9]. Figure 1 (second row) provides a reference scheme and examples of different camera levels. Ultimately, camera angles and levels can be combined in different ways in movie frames to create different effects.

In this paper, we address the problem of automatic recognition of camera level and angle from single frames with a data driven approach using Convolutional Neural Networks (CNN). Our main contributions can be summarized as follows:

- We collect and make available on the project page a novel dataset of unprecedented dimension: 24,665 frames sampled from a wide range of movies, freely available images, and shots from cinematographic websites;
- We annotate such dataset with the corresponding ground-truth labels for camera angles (9,037 frames) and level (15,628 frames) using an ad-hoc tool named `AniXtract` [7] which we developed for this purpose;
- For the task of automatic recognition of camera angle and level, we introduce a CNN whose architecture incorporates two independent classification heads, each responsible for categorizing a different class, and we train it in an alternating fashion.

Differently from other solutions based on human pose estimation (*e.g.,* [27]), the proposed approach demonstrates remarkable effectiveness even when frames do not feature the human figure.

2 RELATED WORK

Existing works on camera features have primarily addressed the automatic extraction of shot scale and camera motion, while no prior research exclusively focused on recognizing camera angle and level from movie frames.

Regarding shot scale, which refers to the distance between the camera and the subjects, the works in [25, 20, 4, 6] have explored this aspect. Meanwhile, [22] and [28] have delved into over-the-shoulder shots, a specific type of shot framing. Other studies have tackled the automatic estimation of camera motion types in sequences of frames, another fundamental aspect in film-making. Early work, such as [17], employed linear combinations of optical flow models to classify roll, pan, tilt, and zoom shots. This approach was later expanded by [29], which included camera rotation. The work in [26] proposed a Markov random field based motion segmentation algorithm to classify pan, tilt, zoom, tracking, and establishing shots, while [5] used linear Support Vector Machines to classify the same types of shots, employing homography parameters as indicators of camera motion.

More recently, the trend has shifted towards end-to-end deep learning models, which simplify the learning workflows and yield better results. Although traditional schemes were employed for example in art movies [4], [20] was the first to propose measuring shot scale with Convolutional Neural Networks (CNNs) in live-action films, which was later found useful for movie style classification in [23]. In [2], frames were preprocessed with Mask R-CNN and Yolact to obtain a semantic segmentation, which improved the recognition of shot scale. In the context of shot type classification, the paper [18] presents a learning framework called Subject Guidance Network, which separates the subject and background of a shot into two streams, serving as separate guidance maps for scale and movement type classification, respectively. Furthermore, the authors introduce a large-scale dataset called MovieShots, comprising 46,000 shots from 7,000 movie trailers, annotated with their scale and movement types.

Very recently human pose estimation (HPE) methods have been used to identify camera features such as position, scale, and movement in order to enable higher artistic and storytelling interpretation, as in [27]. Other approaches for recognizing elementary camera features using CNNs include those in parliamentary debates [12], music concerts [11], and sporting events [15].

3 DATASET

3.1 Dataset Annotation with AniXtract tool

In order to develop computational models that accurately capture camera features such as angle and level, it is necessary to have a formal representation of the involved data. However, annotating this type of data on a large scale can be complex and time-consuming. To facilitate annotation of frames with respect to camera features, we developed `AniXtract` [7], a graphical application for labeling camera angle, camera level, and shot scale. A screenshot taken from `AniXtract` UI is presented in Figure 2. In automatic mode, `AniXtract` can conveniently recall models of suitably trained neural networks to automatically extract the camera features in movies. Annotated frames can be afterwards manually reviewed and corrected by a human operator. Corrected annotations can then be helpful to enlarge the training set and learn better models in an iterative fashion.

For the aim of this work, we annotate camera angle on five different classes: *Overhead, High, Neutral, Low, Dutch*, as shown in Figure 1 (first row). The angle class describes camera rotation along both lateral (High, Neutral, and Low) and longitudinal (Dutch) axes. In particular, an Overhead-angle indicates a take looking down on a subject from an almost perpendicular direction. On the other hand, we categorize camera level (i.e., the height of the camera in the scene in relation to the subject being framed) into six different classes: *Aerial, Eye, Shoulder, Hip, Knee, Ground*, as shown in Figure 1 (second row). The particular class of Aerial-level is used for shots taken from a considerable height, such as from a plane or a drone, showing a large portion of the surroundings. All annotations were performed by a team of expert film scholars: two independent coders and a third person who made decisions in cases of disagreement.

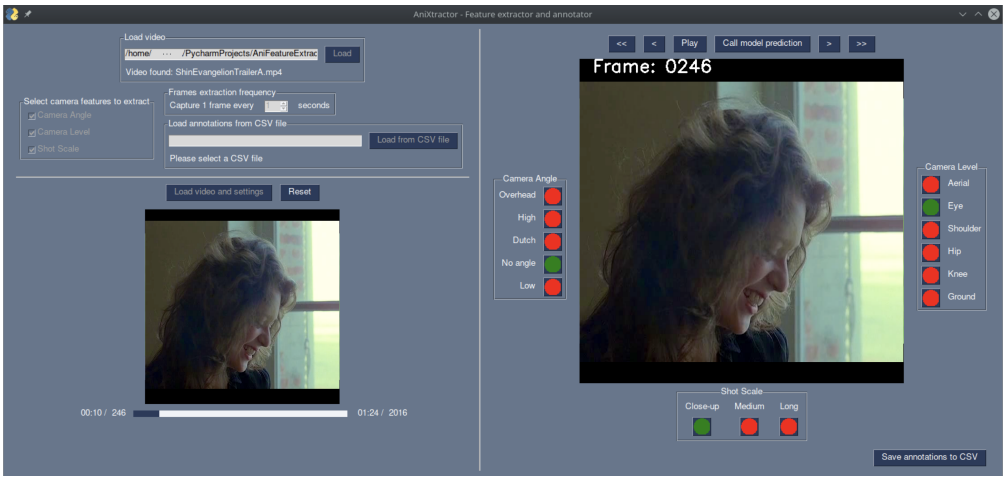

Fig. 2. A screenshot of the AniXtract tool [7] for annotating camera features (angle, level, shot scale). On the left panel it is possible to either load a video (from file or URL, and set a rate for frame extraction) or a previously computed annotation from a *.csv* file. On the right panel it is possible to launch automatic annotation on the video, or to check and/or correct previously computed annotations.

3.2 Dataset Composition

The dataset used for the task of automatic classification of camera angles and levels in movie frames has been collected from various sources:

- All frames in classes Eye-, Shoulder-, and Hip-level were automatically sampled from films by various authors (e.g., Scorsese, Bergman, etc.). The full list of employed movies, both in colors and black&white can be found on the project website;
- Classes that rarely appear in movies, such as Ground- or Knee-level shots were retrieved through Google's image search, then automatically downloaded, and finally manually filtered;
- Most Aerial-level images were extracted (at 10 second intervals) from videos taken by drones over various cities and landscapes from freely available clips on the web;
- Images from other classes were scraped from shot examples taken from Film School Rejects' online database.

The dataset contains a total of over 24,665 images (9,037 for angle, and 15,628 for level) and is made available on the project website. For our subsequent analyses, it is split in 70/10/20 partitions for training, validating and testing. These percentages are chosen to maximize the amount of samples to be used for training the networks, while still having representative testing/validation datasets. The exact number of frames in split datasets for classes of camera angle and level is illustrated in Table 1.

4 METHODS

4.1 Model Architecture

We propose a novel CNN architecture for the classification of camera angle and camera level, based on the ResNet [8] family as the backend. The architecture is specifically designed to handle two distinct sets of classes, while the backend is chosen for its superior performance on ImageNet [19] compared to older models like VGG16, as well as its relatively lower number of parameters. The

Table 1. Number of frames for each class in the dataset.

ANGLE	Overhead	High	Neutral	Low	Dutch
Train.	353	3,523	556	1,546	346
Valid.	43	506	75	221	59
Test	109	1,004	164	442	90
Total	505	5,033	795	2,209	495

LEVEL	Aerial	Eye	Shoulder	Hip	Knee	Ground
Train.	5,083	2,466	2,116	947	107	218
Valid.	688	361	303	168	18	26
Test	1,491	696	604	239	29	68
Total	7,262	3,523	3,023	1,354	154	312

intended use of this architecture is on large movie corpora, and therefore computational constraints are also taken into account.

The proposed architecture includes two independent classification heads, each responsible for categorizing a different non-exclusive set of classes. To effectively train the multi-head network, we introduce a custom loss function that considers the outputs of both classification heads. This loss is a weighted sum of two categorical cross entropy applied to each network head.

To ensure effective training, the data from each set of classes is presented to the network in a round-robin fashion. Specifically, the training iterations alternate between the two datasets, with each batch of data used to train the corresponding classification head. This approach allows both classification heads to be trained evenly and effectively and enables the network to learn the most salient features of each set of classes. This mutual relationship between tasks fosters mutual enhancement, leveraging their interdependence to maximize performance and synergy. Additionally, it allows for the production of both sets of classes jointly in the prediction phase, without wasting computational resources.

To ensure that the images in the dataset are harmonized, the input size is resized to 256×256 and then centre-cropped to 224×224 pixels. The feature maps of the last convolutional layer of the backend are flattened and passed to a fully connected (FC) layer with 512 hidden units and ReLU as the activation function. Given the small number of samples, dropout regularization and batch normalization strategies are applied.

The final layer is an FC layer with a softmax activation function and a number of neurons equal to the number of classes to be recognized for each classifier (five and six for the two different heads, respectively).

4.2 Data Augmentation

To address the low number of images in the dataset, we utilize on-the-fly augmentation by applying both geometric and chromatic transformations. Specifically, we employ `TrivialAugment`, a parameter-free augmentation technique [16] that applies a single augmentation to each image. This method has been shown to outperform previous state-of-the-art automatic augmentation techniques.

4.3 Hyperparameters and Cross-validation

The training process employs the Adam optimizer [1] in all experiments. A learning rate of 0.08 and a batch size of 384 are used for 50 epochs, with the best model selected based on validation

Table 2. Performance on camera angle.

ANGLE	Precision	Recall	F1-score	Support
dutch	0.94	0.87	0.90	109
high	0.96	0.98	0.97	1004
low	0.96	0.93	0.94	442
neutral	0.96	0.96	0.96	164
overhead	0.98	0.90	0.94	90
accuracy			0.96	1809
macro avg	0.96	0.93	0.94	1809
weighted avg	0.96	0.96	0.96	1809

Table 3. Performance on camera level.

LEVEL	Precision	Recall	F1-score	Support
aerial	0.99	1.00	1.00	1491
eye	0.91	0.94	0.92	696
ground	0.97	0.84	0.90	68
hip	0.90	0.90	0.90	239
knee	0.92	0.76	0.83	29
shoulder	0.91	0.87	0.89	604
accuracy			0.95	3127
macro avg	0.93	0.89	0.91	3127
weighted avg	0.95	0.95	0.95	3127

performance. Due to the multi-label nature of the problem and the imbalanced samples, the `F1 score` is used as the metric for monitoring network performance. Based on experimental results, ResNet18 is chosen as the best-performing backend and is utilized in the final model. Training is conducted on a workstation equipped with NVIDIA GPUs, and all code is implemented in PyTorch.

5 RESULTS

In Table 2 and Table 3 we show the results obtained on camera angle and level, respectively. Considering the vastness and heterogeneity of the data domain, the limited availability and variety of usable data for some classes, and the presence of errors in the ground-truth in case of ambiguous frames, the obtained scores (precision, recall, and accuracy around 95%) are highly satisfactory.

More insights on performance can be inspected in Figures 3 and 4, where we show the confusion matrices obtained on the testing sets of camera angle and level, respectively.

5.1 Error analysis

After analyzing the primary reasons for misclassifications of camera angles compared to the ground-truth, the following observations are reported:

- A fraction of Low-angle frames are wrongly classified as High-angle frames. This is often due to images that are challenging to classify even for human annotators (as in Figure 5(b)), or present misleading perspective visual cues that cause the network to respond strongly to the wrong class (as in Figure 5(c)).

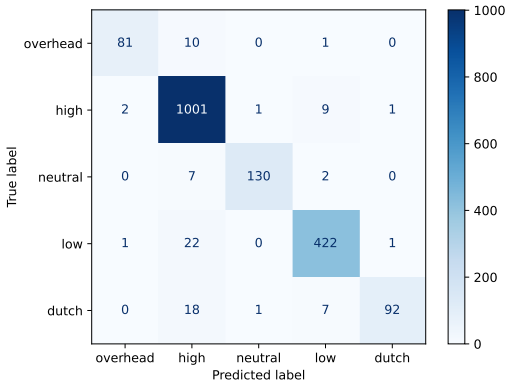

Fig. 3. Performance on camera angle recognition task.

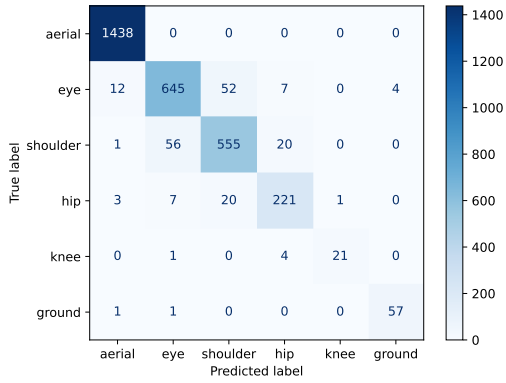

Fig. 4. Performance on camera level recognition task.

- Some Dutch-angle frames are incorrectly labeled as either High- or Low-angles. This is because the Dutch scheme is classified as a possible angle, but in this case, the rotation occurs in the longitudinal plane. As a result, a Dutch shot can also be high, low, or neutral, as these rotations occur on different planes (see e.g., Figure 5(d)).
- Other errors occur because the network struggles to understand the context of various scenes. Many manual annotations rely on knowledge of what is happening in the scene, whereas the network must determine these responses solely based on visual stimuli. For example, in the close-up in Figure 5(a), it is difficult to distinguish whether the subject is standing (Low-angle) or lying (Overhead-angle) without the context given e.g., by adjacent frames.
- Minor misclassifications can occur due to heavily cluttered frames or scenes with unclear geometry, as in Figures 5(e) and (f).

These are the different cases that contribute to errors in camera level prediction:

- Contiguous classes: the majority of errors occur between adjacent camera levels, for example, Eye-level and Shoulder-level or Shoulder-level and Hip-level. Errors that are far away from the diagonal of the confusion matrix are infrequent.
- Multiple subjects: when there are multiple individuals in a frame, the network may have difficulty identifying the correct camera reference because it relies on certain anatomical parts such as eyes, shoulders, or hips to determine the camera level. Frames with multiple subjects can lead to incorrect classification, as shown in Figures 5(i) and (j).
- Other causes of minor misclassifications are due to i) Low contrast: shots taken at night can be challenging to classify due to low contrast, as in Figure 5(l); ii) Unclear content: Frames with unclear content, such as Figure 5(g), where the presence of the sun and the absence of clear contours delineating human faces can result in incorrect classification; (iii) Overlapping labels: Images with overlapping labels, as in Figure 5(k), can be difficult to classify even for human annotators; (iv) Little visual context: Frames with little visual context information, as in Figure 5(h), can be challenging to classify.

In Figure 6 we show the class activation maps (obtained using Grad-CAM [21]) for a couple of erroneously classified frames ("unclear content" in Figure 5(g) and "overlapping labels" in Figure 5(k)). By observing the highlighted regions which are relevant for the obtained prediction, it is possible to formulate hypothesis about the main causes of wrongly predicted frames.

6 CONCLUSION

This paper presents a data-driven approach for automatic classification of camera angle and camera level in movie frames using an original two-head CNN architecture. The proposed model is able to achieve high accuracy in distinguishing among five different camera angle classes and six diverse camera level categories, even when the human figure is not displayed. Such an approach has potential applications in genre recognition, stylistic analysis, and movie recommendation.

Regarding future improvements, we suggest two main directions. The first direction is to augment the training data with other shots to improve the balance of shot populations, especially in the most problematic categories. The second direction is the large-scale application of camera angle and level automatic recognition to the problem of the emotional characterization of movies and their psychological impact on viewers. This is an interesting and potentially valuable application of the presented method, as it could contribute to a quantitative assessment of the emotional content of movies.

Overall, the paper presents a promising approach for automatic annotation of cinematographic elements in movies, with potential applications in film analysis and recommendation, as well as psychological research.

ACKNOWLEDGMENTS

The authors would like to thank MSc students Andrea Ferrari and Gianluca Gualandris for their contributions to the research and to the development of `Anixtract` tool.

REFERENCES

[1] Diederik P. Kingma and Jimmy Ba. 2015. Adam: a method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. Yoshua Bengio and Yann LeCun, (Ed.) http://arxiv.org/abs/1412.6980.

[2] Hui-Yong Bak and Seung-Bo Park. 2020. Comparative study of movie shot classification based on semantic segmentation. *Applied Sciences*, 10, 3390.

[3] S. Benini, M. Savardi, K. Bálint, A. B. Kovács, and A. Signoroni. 2019. On the influence of shot scale on film mood and narrative engagement in film viewers. *IEEE Transactions on Affective Computing*, 1–1. DOI: 10.1109/TAFFC.2019.2939251.

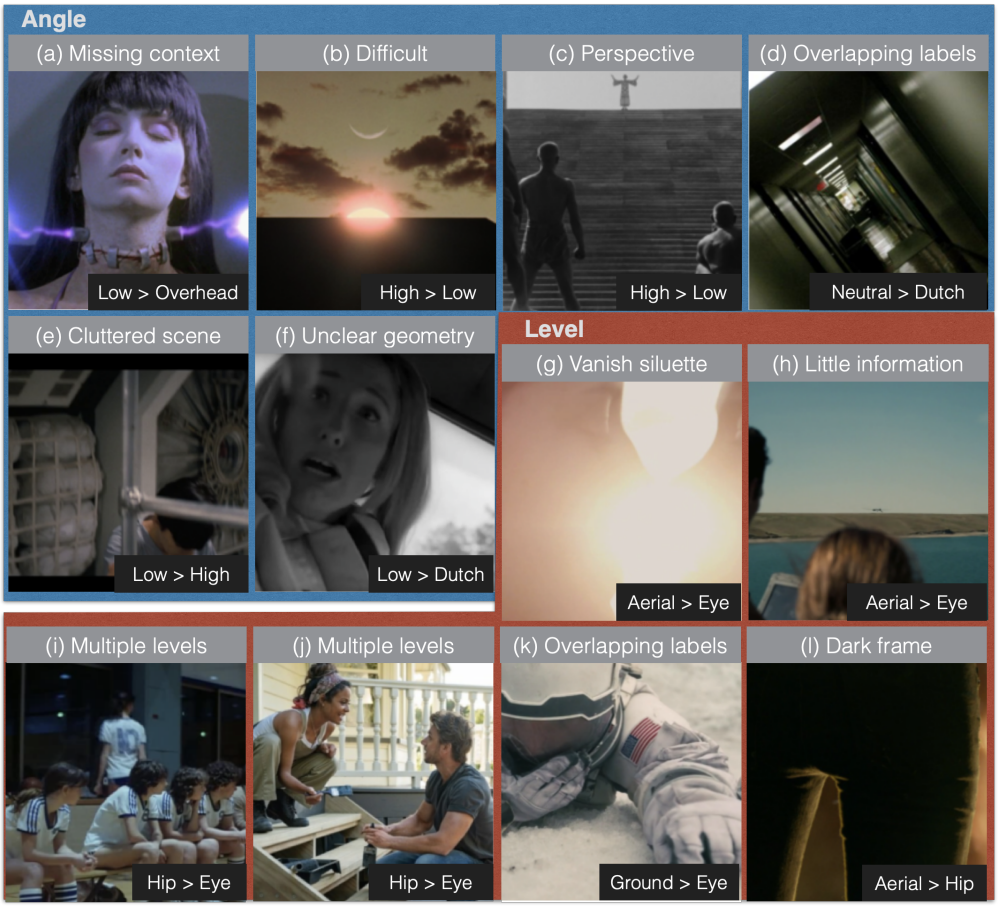

Fig. 5. Error analysis: examples of main causes of misclassification. All frames are tagged with *predicted-label > ground-truth-label*.

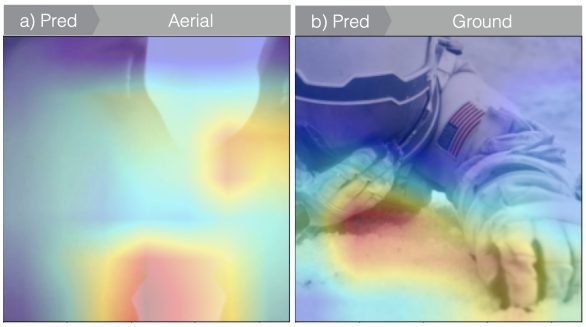

Fig. 6. Class activation maps for some selected challenging frames. The frames are from Fig. 5(g-h), tagged with the target class for a visual explanation.

[4] Sergio Benini, Michele Svanera, Nicola Adami, Riccardo Leonardi, and Andras B. Kovacs. 2016. Shot scale distribution in art films. *Multimedia Tools and Applications*, 75, (Dec. 2016). DOI: 10.1007/s11042-016-3339-9.

[5] S. Bhattacharya, R. Mehran, R. Sukthankar, and M. Shah. 2014. Classification of cinematographic shots using lie algebra and its application to complex event recognition. *IEEE Transactions on Multimedia*, 16, 3, (Apr. 2014), 686–696. DOI: 10.1109/TMM.2014.2300833.

[6] Ines Cherif, Vassilios Solachidis, and Ioannis Pitas. 2007. Shot type identification of movie content. *2007 9th International Symposium on Signal Processing and Its Applications*, 1–4.

[7] Gianluca Gualandris, Mattia Savardi, and Sergio Benini. 2021. AniXtract. https://github.com/Mad0Scientisto/AniXtract. [Online; accessed 13-March-2023]. (2021).

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778.

[9] Wei Huang, Judith S. Olson, and Gary M. Olson. 2002. Camera angle affects dominance in video-mediated communication. In *CHI '02 Extended Abstracts on Human Factors in Computing Systems* (CHI EA '02). Association for Computing Machinery, Minneapolis, Minnesota, USA, 716–717. ISBN: 1581134541. DOI: 10.1145/506443.506562.

[10] Tess Lankhuizen, Katalin E Bálint, Mattia Savardi, Elly A Konijn, Anne Bartsch, and Sergio Benini. 2020. Shaping film: a quantitative formal analysis of contemporary empathy-eliciting hollywood cinema. *Psychology of Aesthetics, Creativity, and the Arts*.

[11] Jen-Chun Lin, Wen-Li Wei, Tyng-Luh Liu, Yi-Hsuan Yang, Hsin-Min Wang, Hsiao-Rong Tyan, and Hong-Yuan Mark Liao. 2018. Coherent deep-net fusion to classify shots in concert videos. *IEEE Transactions on Multimedia*, 20, 11, 3123–3136. DOI: 10.1109/TMM.2018.2820904.

[12] Pedro A. Marín-Reyes, Javier Lorenzo-Navarro, Modesto Castrillón Santana, and Elena Sánchez-Nielsen. 2016. Shot classification and keyframe detection for vision based speakers diarization in parliamentary debates. In *CAEPIA*.

[13] Thomas A. McCain, Joseph Chilberg, and Jacob Wakshlag. 1977. The effect of camera angle on source credibility and attraction. *Journal of Broadcasting*, 21, 1, 35–46. DOI: 10.1080/08838157709363815.

[14] Joan Meyers-Levy and Laura A. Peracchio. 1992. Getting an angle in advertising: the effect of camera angle on product evaluations. *Journal of Marketing Research*, 29, 4, 454–461. DOI: 10.1177/002224379202900406.

[15] Rabia A. Minhas, Ali Javed, Aun Irtaza, Muhammad Tariq Mahmood, and Young Bok Joo. 2019. Shot classification of field sports videos using alexnet convolutional neural network. *Applied Sciences*, 9, 3. DOI: 10.3390/app9030483.

[16] Samuel G Müller and Frank Hutter. 2021. Trivialaugment: tuning-free yet state-of-the-art data augmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 774–782.

[17] Sang Cheol Park, Hyoung S. Lee, and Seong Whan Lee. 2004. Qualitative estimation of camera motion parameters from the linear composition of optical flow. English. *Pattern Recognition*, 37, 4, (Apr. 2004), 767–779. DOI: 10.1016/j.patcog.2003.07.012.

[18] Anyi Rao, Jiaze Wang, Linning Xu, Xuekun Jiang, Qingqiu Huang, Bolei Zhou, and Dahua Lin. 2020. A unified framework for shot type classification based on subject centric lens. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*. Springer, 17–34.

[19] Olga Russakovsky et al. 2015. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision (IJCV)*, 115, 3, 211–252.

[20] Mattia Savardi, Alberto Signoroni, Pierangelo Migliorati, and Sergio Benini. 2018. Shot scale analysis in movies by convolutional neural networks. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, 2620–2624. DOI: 10.1109/ICIP.2018.8451474.

[21] Ramprasaath R. Selvaraju, Abhishek Das, Ramakrishna Vedantam and Michael Cogswell, Devi Parikh, and Dhruv Batra. 2016. Grad-cam: why did you say that? visual explanations from deep networks via gradient-based localization. *CoRR*, abs/1610.02391. http://arxiv.org/abs/1610.02391.

[22] M. Svanera, S. Benini, N. Adami, R. Leonardi, and A. B. Kovács. 2015. Over-the-shoulder shot detection in art films. In *2015 13th International Workshop on Content-Based Multimedia Indexing (CBMI)*. (June 2015), 1–6. DOI: 10.1109/CBMI.2015.7153627.

[23] M. Svanera, M. Savardi, A. Signoroni, A. B. Kovács, and S. Benini. 2019. Who is the film's director? authorship recognition based on shot features. *IEEE MultiMedia*, 26, 4, (Oct. 2019), 43–54. DOI: 10.1109/MMUL.2019.2940004.

[24] Robert K. Tiemens. 1970. Some relationships of camera angle to communicator credibility. *Journal of Broadcasting*, 14, 4, 483–490. DOI: 10.1080/08838157009363614.

[25] Nicholas Vretos, Ioannis Tsingalis, and Ioannis Pitas. 2012. Svm-based shot type classification of movie content. In (Mar. 2012).

[26] H. L. Wang and L. Cheong. 2009. Taxonomy of directing semantics for film shot classification. *IEEE Transactions on Circuits and Systems for Video Technology*, 19, 10, (Oct. 2009), 1529–1542. DOI: 10.1109/TCSVT.2009.2022705.

[27] Hui-Yin Wu, Luan Nguyen, Yoldoz Tabei, and Lucile Sassatelli. 2022. Evaluation of Deep Pose Detectors for Automatic Analysis of Film Style. In *Workshop on Intelligent Cinematography and Editing*. Rémi Ronfard and Hui-Yin Wu, (Eds.) The Eurographics Association. ISBN: 978-3-03868-173-1. DOI: 10.2312/wiced.20221047.

[28] M. Xu, J. Wang, M. A. Hasan, X. He, C. Xu, H. Lu, and J. S. Jin. 2011. Using context saliency for movie shot classification. In *2011 18th IEEE International Conference on Image Processing*. (Sept. 2011), 3653–3656. DOI: 10.1109/ICIP.2011.6116510.

[29] Xingquan Zhu, Xiangyang Xue, Jianping Fan, and Lide Wu. 2002. Qualitative camera motion classification for content-based video indexing. In (Dec. 2002), 1128–1136. DOI: 10.1007/3-540-36228-2_140.